AN ABSTRACT OF THE THESIS OF

Duy M. Nguyen for the degree of Master of Science in Robotics presented on November 27, 2018.

Title: Initial Designs for Improving Conversations for People Using Speech Synthesizers

Abstract approved: _____

William D. Smart

This thesis aims to determine the impact that Augmentative and Alternative Communication (AAC) devices have on social interactions, and then improves the AAC user experience through a user focus design process. AAC devices enable people who cannot speak to communicate with others. Unfortunately, they are tedious to use and make social interaction a dissatisfying experience. The thesis consists of three main studies. The first study focus on gathering information on behaviors of communicative partners, specifically gaze behaviors, of AAC users and how those behaviors impact the users. The second study takes a form of a focus group with people who are experienced with AAC devices. The main discussions on the focus group are to gather ideas about daily interactions of actual AAC users, and brainstorm design ideas for technologies that can improve their interactions with others. Finally, the last study aims to get feedback on prototypes created from ideas in

the second study.

Initial Designs for Improving Conversations for People Using Speech Synthesizers

by

Duy M. Nguyen

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented November 27, 2018
Commencement June 2019

Master of Science thesis of Duy M. Nguyen presented on November 27, 2018.

APPROVED:

Major Professor, representing Robotics

Head of the School of Mechanical, Industrial and Manufacturing Engineering

---

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Duy M. Nguyen, Author

ACKNOWLEDGEMENTS

I would like to thank ...

Dr. Bill Smart for his guidance and support throughout the process of this thesis. My time in his lab was an invaluable and mind-opening experience as I was given opportunities to work with a variety of state-of-the-art technology.

Dr. Frank Bernerri for teaching me how to design, conduct, and analyze research studies. His thoughtful perspectives and comments made me into a better researcher. I want to thank him for giving me a foundation for my experimental practices and a roadmap to look at whenever I need to get over roadblocks during this whole process.

Dr. Heather Knight for broadening my knowledge on social robotics and creating HRI study. Her class gave a more realistic vision of how robots should interact in the real world.

Dr. Kathleen Bogart for filing the gap in my knowledge of disability research. Her insights and teachings allow me to design a better and more inclusive studies.

Dr. Matthew Johnston for his interests and support for my project. I really appreciate his time and accommodations to be in my committee in such a short notice.

Matt Rueben for being a great friend and role model for the last five years. I want to thank him for taking care of me, making me feel welcome in the lab, and always being available to answer my random questions. I will miss all the time we spent talking about future research ideas.

Wendy Xu, Abhi Agnihotri, and Chandra Char for spending their time talking with me about all the problems I have in this research project and in my personal life at Oregon State University. I want to thank them for being great friends and always there when I needed.

Ramee Kelly and Saritha Suram for all your help in the process of drafting my IRB. They made the long and laborious IRB submission less scary and more enjoyable.

Christopher Eriksen, Christopher Bollinger, Jeffrey Klow, Austin Nicolai, Austin Whitesell and friends in the Personal Robotics Lab for always willing to help me with any technical problems I had throughout this project.

Amber Fultz, Morgan Stosic, Rafael Robles, Meghan Heineman, and friends in the Interpersonal Sensitivity Lab for their feedback in the designs of the studies in this project.

Hannah Stone for helping me running and analyzing the study.

Mary Rebar for being such a great help throughout my last two studies. I want to thank her for all the time and effort that she has spent to organize and recruit people for my studies.

Ralph and Suzi, Nancy and Steve, and the people in the ALS support group at Eugene and Salem for their help and inputs in my project.

My family for always supporting me. I want to thank them for trusting me and always be available for me. I want to thank my brother for all his time editing the thesis. For my parents, *con biết ơn ba mẹ rất nhiều vì đã nuôi dạy và giúp đỡ con trong suốt quá trình thực hiện luận án.*

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

LIST OF APPENDIX FIGURES

LIST OF APPENDIX TABLES

LIST OF APPENDIX TABLES (Continued)

Initial Designs for Improving Conversations for People Using Speech Synthesizers

Part I

Introduction

Amyotrophic Lateral Sclerosis (ALS), also known as Lou Gehrig's disease, is a progressive neurodegenerative disease that currently affects more than 12,000 American, with around 6,000 people diagnosed each year. As the most common motor neuron disease around the world, ALS causes the brain nerve cell (upper motor neuron) or the spinal cord (lower motor neuron) to degenerate and weakens the links that transport signal from the brain to voluntary muscles across the body. As the result, the patients will gradually lose the ability to perform day-to-day activities such as waking, coordinating hand movement, swallowing, speaking, etc. According to reports from the National Institute of Neurological Disorders and Stroke, the life expectancy of ALS patients is around three to five years after diagnosis, and only 10% of them can live more than ten years [1]. At this moment, there is still no definitively known cause for the disease. Additionally, there are no effective cures available either [2].

As the disease progresses, around 80% of the patients will develop dysarthria and eventually lose their ability to communicate. The condition tends to cause a large amount of stress and discomfort to the patients as they cannot have meaningful conversations nor verbally express feelings toward their loved ones during their final moments [3]. To help those patients, researchers around the world have created a speech synthesizer device called the Augmentative and Alternative Communication (AAC) device. These machines allow people with speech impairments to convey their thoughts and intentions to others by inputting them to a computer, which will synthesize the message into a spoken response. There are multiple versions of AAC devices, with the oldest and simplest version of the devices including

only pictures that represented basic needs of the patients, such as needing an assistant, feeling cold, wanting to eat, or expressing happiness etc. [4]. Recent versions of the device incorporate a text to speech function, which can greatly improve the patients′ communicative attempts. The input mechanisms of those devices range from typing with keyboard to using eye gaze detection to select characters from a screen [5]. As a result of the increase in variety and usability of the devices, almost 96% of ALS patients who are recommended an AAC intervention accept the device and use it to the end of their life [6].

Assistive devices allow ALS patients to reconnect with their family members and continue to communicate with others. It is a small step in the right direction, but many improvements are still needed for developing a better and more comfortable social experience for AAC users. Multiple AAC users report that AAC device invokes boredom, loss of interest, and lower perception of trustworthiness from their communicative partners. The root of those uncomforted feeling can be contributed to the mechanical and impersonal tone of computerized voice of the device. [7] [8] [9]. This problem has plagued the AAC since its development. Fortunately, various potential solutions have slowly emerged over the years, such as better speakers with a wider range of pitch and customized voices [10].

Unfortunately, there exists another major drawback of AAC devices that has not yet been remedied. When a patient uses an AAC device, she or he needs to type her or his response out, which is an extremely slow process that can greatly extend the duration of typical pauses in normal conversations [11] [12]. Only a few seconds of those atypical pauses are enough to cause confusion and discomfort to

the communication partner [13] [14].

This prolonged response time impacts other aspects of the conversation as well. Multiple studies in conversation with AAC users showed that their exchanges are dominated by a) closed end, yes-no questions from the AAC users' communication partners, b) a lack of initiation from AAC users, and c) a lack of interpersonal coordination in turn-taking between AAC users and their partner [15] [16] [17]. Many AAC users report being aggravated by the fact that they are unable to properly convey their ideas [15]. As a result, AAC users suffer from low quality of face-to-face conversations, especially with strangers. This, in turn, limits the amount of rapport they can achieve with others.

The ultimate goal of this research program, where the current report is its first step, is to improve the quality of life for ALS patients by creating technology that allowing them to have a more coordinated and natural conversation with others. The thesis consist of three major states: 1) gathering initial data and ideas from the AAC users, 2) designing and validating the prototype generated from those ideas, and 3) summarizing the findings and suggesting future directions.

Part II

Gather Initial Data and Ideas from Users

This chapter starts with a background information on the important of attention in building rapport during conversations and our approach in designing a better systems for the AAC users. The next section covers the first study that is designed to gather behaviors data from interactions between the AAC users and their communication partners. Finally, we ends with a description about a focus group study in which we brainstorm design ideas for our system with actual AAC users.

Chapter 1 Background Information

1.1 The roles of attention in building rapport

It is not uncommon for people to interact with strangers as though they have known each other their whole life. In contrast, there are cases where long-time acquaintances behave as if they are complete strangers. The first case is an example of high rapport interaction when people just click with each other and become a harmony and unified group. On the other hand, the second case would be described as an interaction devoid of rapport; one that feels disconnected and awkward. In social psychology, rapport is a construct that is associated with the quality of the relation or connection between individuals at a group level ([18]). Linda Tickle-Degnen and Robert Rosenthal (1990) have identified three different components to rapport within face-to-face interactions: emotional positivity, coordination, and attention [19].

Emotional positivity represents the good feelings toward people that an individual is interacting with. It is an individual's first impression of his or her communication partners that can set an initial tone for the interaction. For example, people tend to enjoy conversation with attractive individuals more than their counterpart [20].

After the first impression period, the outcome of the interaction relies more heavily on the coordination between the participants. Coordination can be understood in terms of the "chemistry" between people; how well they understand the conversation by regulating the turn-taking exchanges and having smooth flows of verbal and nonverbal behaviors [21].

Last but not least is attention. Attention plays a major role throughout the interaction by acting as the bridge connecting emotional positivity and coordination. At the beginning of the interaction, the attention is focused on identifying the positive cues from the interactants based on their biological appearance. As the conversation begins, the attention slowly shifts to the topic of the conversation and nonverbal cues such as eye gaze, hand moment, body position, etc. to create a more cohesive interaction. Hence, in the conversational setting, attention is the most important role in the development of rapports between people [19].

1.1.1 What is attention?

Attention is a common and important cognitive process in daily social interaction, but although it is an intuitive construct, its precise nature remains elusive. E. Bruce

Goldstein (2011, p.82), a cognitive psychologist at University of Pittsburgh who has published multiple textbooks on the topic, defines it vaguely as, "the ability to focus on specific stimuli or location" [22]. Alternatively, Daniel Kahneman (1973, p.2), a harbinger in studying attention, whose book, Attention and Effort, is cited by thousands of researchers around the world, views attention as "a label for some of the internal mechanisms that determine the significance of stimuli" [23]. In other words, attention is more of a label we ascribe to an inferred causal agent than it is a reference to an objectively describable neural event. It's not always clear, for example, whether scientists operationalize attention by measuring what is interesting to us, or operationalizing what is interesting to us by measuring what we appear to be attending to.

1.1.2 Measuring attention in the conversational setting

For capturing signs of attention, psychologists have applied multiple methods, from measuring brain signals, detecting reaction delay, to coding eye gaze behaviors [18] [24] [25] [26] [27] [28]. Brain signals are analyzed through a recording device implanted into the primary visual cortex, which in turn measures the firing rate of the neuron in order to evaluate the attention level [24] [25]. Due to the intrusive nature of the technique, it is used mostly on animals to gain understandings of the physiological reaction in our brain. The second popular technique to identify attention is the detection of reaction delay. This method is mainly used for task-oriented experiments in which participants perform certain cognitive activities while under

the influences of different stimuli. The times for finishing each activity are recorded and analyzed to show whether participants are distracted by the stimuli or not [28]. Therefore, eye gaze is the most appropriate methods to measure the attention in a setting meant to simulate social conversations [18] [27].

Adam Kendon (1967) used gaze to study attention in typical conversations [11]. He recorded multiple films of people having conversations and annotated their eye movement in each frame. His study showed that people looked at the communicative partners around 41% of the time while talking and 58% of the time while listening. These numbers are similar to a later study conducted by Argyle and Ingham (1972) in which the participants gazed at their partners is 37% of the time while talking and 68% of the time while listening [29]. It means that in a typical conversation, people spend around 50% of all their time looking at their partners. Specifically, they tend to look at their partners less while speaking than while listening to the other person. These important studies demonstrate how an investigation studying the impact of attention people pay to one another during a conversation can be compared to baseline normal conversations.

1.2 Our approach in designing the systems

There have been studies that analyze behaviors of people in natural conversations through time [29] [11], but no one has yet looked at the actions of individuals in interactions between AAC users and their communicative partners. Therefore, we designed our first study to gather behaviors data from conversants in this unique

setting. The study will give us a better understanding of the people's behaviors in this unusual interaction, and valuable data for a future program that we might want to develop.

While the results in the first study is a great starting point, the study used convenience participants whom might not have any experiences with AAC devices. A second study was conducted to gather design inputs from people that had experiences with an AAC device. This new study was inspired by Hee Rin Lee et. al.'s paper [30] on how to use participatory design (PD) method to incorporate technologies into people's daily life. The PD method emphasizes the important of including the users into the designing process to ensure the technology will fit the users' needs.

Chapter 2 First Study: Gathering Behaviors Data

The study consisted of three five-minute interactions between a pair of participants. One of the two participants was assigned to be an AAC user, who had to use a text-to-speech device to communicate, while the other participant could converse normally. In the first interaction, the AAC user typed his or her response using the keyboard. In the second interaction, the same setting was employed except the AAC user worked with an Xbox controller instead of a keyboard. The purpose of always having the Xbox controller interaction after the keyboard interaction was that we wanted the participants to familiarize themselves with the experimental setting and the communication system. In order to increase the external validity of

the study as a simulation of normal AAC usage, we could not put naive participants immediately into a situation where the technological constraints were so novel and challenging that it would overwhelm any other psychological phenomenon we were interested in assessing. Furthermore, as most people in our current generation utilize some sort of communication technology such as texting, Skype, Facebook Messenger, etc., in their daily activities, we expected that the first interaction would come as second nature to them, which would not be far different from their typical conversation. Between each interaction, participants were given a packet of surveys to measure the rapport between them. Finally, a third conversation was held where participants could freely talk to each other. The purpose of this interaction was a mean for participants to fully express themselves without any handicap and was not relevant to the goal of this study.

2.1 Materials

2.1.1 Survey

In this experiment, we employed nine different surveys. However, for the analysis in this thesis, we focused solely on the demographic survey and interaction assessment survey, which measured the rapport between two participants (Appendix A).

2.1.2 Hardware and software

The study involved three cameras: two of them recording each participant's face and the third one recording both participants from the side. We had a set of computer system, which included a NUC computer system, a monitor, a speaker, a keyboard, a mouse, and an Xbox controller. The monitor was the only equipment visible on the table. All other parts of the computer were hidden under the table to prevent any unnecessary distraction. The monitor was pushed a little to the side and turned toward the AAC user to prevent it from turning into an unintentional wall between the pair of participants, which could disrupt their ability to perceive their partner's nonverbal cues (Figure 2.1). Additionally, this setting was more aligned with how most AAC system is set up in the real world.

Figure 2.1: *The setting for an Xbox controller interaction*

For the text-to-speech program, we modified the Festival software from Black Alan [31] and used voices from the CMU Database [32] with the voice named RMS for male participants and SLT for female participants. Another open-source

software named Xboxdrv was utilized to allow us to operate the Xbox controller as a mouse and type by clicking on a virtual keyboard (Figure 2.1). All of the software was on an Ubuntu operating system.

We used Adobe Premiere Pro software to create split-screen videos by combining videos from the two cameras that were recording the participants' face (Figure 2.2). According to a meta-analysis from [27], the assessment of gaze can be improved if it is done on split-screen videos and has a slow-motion option.

Figure 2.2: *Split-screen videos created from 2 cameras recording participants in an interaction*

2.2 Participants

This experiment recruited 160 participants (33 males and 127 females, $M_{age} = 19.7$) from Oregon State University. Participants were from introduction psychological courses open to students in any major. They received extra credits for their contribution to the study.

We took out ten dyads in which there were technical issues happened during

the interaction, two dyads in which the participants had already participated in previous interaction, and nine dyads in which the participants knew each other [1]. The reason that we took out participants that already knew each other was that we wanted to focused on the interaction between two strangers. Table 2.1 shows the decomposition of the gender of the participants that would be used for the analysis.

Table 2.1: *Gender of participants in the study*

Dyads	Frequency
Male (AU) - Male (CP)	2
Male (AU) - Female (CP)	9
Female (AU) - Male (CP)	16
Female (AU) - Female (CP)	41

2.3 Procedure

When the participants arrived at the lab the first participant was always assigned to be the AAC user, called Participant AC, and the second participant was his or her communication partner, called Participant CP. This was done because more time was needed to teach the participant on how to use an AAC device. There were two experimenters for each session, and each of them was assigned to one of the participants throughout the entry study.

Each participant was lead into a different room and was advised not to talk to each other outside of the interaction to prevent any unaccounted influence on

[1] These were participants who rated 4 or higher in the Final Question Survey (Appendix A)

their rating. Participants were given a brief introduction of the study and a consent form. After the participants fully understood and signed the consent form, both participants were given the background information surveys to complete (Appendix A). Additionally, participant AC received a training on the AAC device after he or she finished the surveys. Then, participant CP was lead to the computer room where participant AC was sitting in. The experimenters began the first interaction and asked the participants to remove their hat or glasses because those could obstruct or reflect the light toward the camera.

Next, the experimenters started the calibration process and asked each participant to follow a moving finger to several specific locations with their eyes while keeping their head still in order to calibrate the video images of their gaze to standard fixed locations (e.g., right eye of partner, left eye of partner, chest of partner, middle of the table, monitor, empty area to the right of his or her partner, and empty area to the left of his or her partner). After the calibration, the experimenters went to the room next door and signaled the start of the interaction by turning on the headlight in the computer room. Throughout all interactions, if the AAC user faced any technical difficulty, she or he could raise her or his hand and an experimenter would come to aid her or him. After five minutes, the experiments turned the light off and waited five seconds before coming back into the room. This procedure ensured that each conversation across all experimental sessions was constant in duration.

After each conversation, the experimenter in charge of participant CP led him or her to the room next door while participant AC remained in the same spot. The

experimenters gave them a set of Post Conversation Survey that has the Interaction Assessment Scale (Appendix A) to complete. After the pair was done with their survey and participant AC received her training for the X-box controller condition, participant CP was guided back to the computer room and began the same process as the first interaction: calibrating eye gaze, interacting for five minutes, separating to a different room, and working on a set of Post Conversation Survey that included the Interaction Assessment Scale. The same process was repeated for the third time, but both participants, at that moment, could talk normally. Additionally, instead of the Post Conversation Survey, the participants were given a set of Post Experiment Survey that had a question about whether the participants knew each other (Appendix A). Participant CP was then led back to the computer room one last time and an experimenter would start the debriefing process. After making sure both participants did not have any concerns or questions about the study, the experimenters lead the participants outside and thanked them for their time.

2.4 Coding process

The study implemented two different coding processes: talking code and eye gaze code.

2.4.1 Talking code

For the talking code, there were five different categories: *speaking*, *typing*, *self-simultaneous speech (SSS)*, *hovering*, and *no speaking (NS)*. *Speaking* was coded the moment the participant CP speak or the computer started to synthesize Participant AC′s response; *typing* was coded when participant AC hit the keyboard or clicked on the Xbox controller; *SSS* was coded when participant AC was typing while the computer was speaking at the same time; *hovering* was coded when Participant AC had his or her hand on top of the keyboard or the controller but had yet to actually hit or type anything; and *NS* was coded for the actions that did not belong to any of the four previous categories. The experimenters annotated the talking behaviors for both participants in the first and second interaction.

2.4.2 Eye gaze code

For eye gaze location coding, there were four different locations coded: *face*, *monitor*, *keyboard*, and *other* (e.g., looking around): *face* was coded when participant CP was looking at the upper half of his or her partner′ face (the red region in Figure 2.3); *keyboard* included the body of the participant AC toward the middle of the table (the yellow region in Figure 2.3); *monitor* was coded when participant CP looked at the monitor direction including from the top of the screen to the base on the table (the blue region in Figure 2.3); and *around* was participant CP looking around the room or to places that did not belong to the other three categories (the green region in Figure 2.3). The coders pinpointed the precise frame when the

gaze location change and marked it with the name of the new gaze location.

Figure 2.3: *Coding schema for eye gaze location*

2.4.3 Coder agreement

Several different coders were needed because it took over fifteen hours to measure each interaction. In order to assess coder reliability, all coders coded the same two interactions generating a very large sample of 18,000 (30 frames/sec * 60 sec/min * 5 min/conversation * 2 conversations/pair of participants) measurements per pair of participants in which their agreement could be assessed. We computed the reliability of each of the four gaze categories and each of the five taking categories individually by dummy coding the nine-category nominal scale into nine separate binary (present/absent) scales. For example, a *Face gaze* variable was created where a frame was coded as 1 if the target was looking at his or her partners′ face and coded as 0 if the target was looking at the monitor, the keyboard, or anything else. When two or more coders coded the $36,000^{th}$ frames from two pairs of participants, intercoder reliability can be estimated by a simple correlation coefficient calculated

with the number of measurements made by the two coders [33].

2.4.3.1 Gaze coding

Table 2.2: *Correlations between coders coding on gaze location*

		Coder ID			
		A	D	L	R
Face	D	0.87	–		
	L	0.90	0.79	–	
	R	0.76	0.81	0.72	–
	S	0.95	0.80	0.90	0.64
Around	D	0.84	–		
	L	0.93	0.90	–	
	R	0.73	0.91	0.72	–
	S	0.96	0.95	0.94	0.59
Monitor	D	0.86	–		
	L	0.85	0.67	–	
	R	0.44	0.71	0.39	–
	S	0.86	0.85	0.80	0.41
Keyboard	D	0.49	–		
	L	0.82	0.59	–	
	R	0.60	0.53	0.52	–
	S	0.84	0.54	0.79	0.61

For the gaze categories, we were mainly interested in the coders' agreement of face and around gaze. The reason was that face gaze could be a sign for when attention was directed to the AAC user (Participant AC) and around gaze could be an indicator of distraction. Gazing at the monitor and keyboard was an interesting behavior as it could be interpreted as either inattention or mutual attention. In Table 2.2, the average correlation of coder agreement is greater than .70. Thus, we

believed that it was acceptable to use the data from the coders for our analysis.

2.4.3.2 Taking coding

Table 2.3: *Correlations between coders on coding talking*

			Coder ID		
			H	M	R
Participant AC		M	0.91	–	
	NS	R	0.66	0.65	–
		A	0.82	0.82	0.67
		M	0.94	–	
	Speaking	R	0.92	0.92	–
		A	0.88	0.87	0.87
		M	0.93	–	
	Typing	R	0.86	0.85	–
		A	0.82	0.83	0.80
		M	Null	–	
	Hovering	R	Null	0.56	–
		A	Null	0.24	0.29
		M	Null	–	
	SSS	R	Null	0.56	–
		A	Null	0.24	0.29
Participant CP		M	0.82	–	
	NS	R	0.69	0.73	–
		A	0.78	0.82	0.72
		M	0.82	–	
	Speaking	R	0.69	0.73	–
		A	0.78	0.82	0.72

Note. NS: no speaking; SSS: self-simultaneous speech;
Null: we cannot calculate the result as coders do not code them

In this talking annotation, the three important characteristics of the conversation were non-speaking (NS), speaking, and typing. Self-simultaneous speech (SSS)

or hovering tag were created for certain specific occasions and appeared only for a short amount of time. From Table 2.3, it is shown that the correlations between coders at NS, speaking, and typing were relatively high, as most of them were above .65. For this reason, we believed it was acceptable to use their data for the analysis.

2.5 Analysis and hypotheses

There were three different aspects of the data that were analyzed in this study: the typing and gaze behaviors from the CP participant; the rapport from the participants in the dyad; and the relationship between the gaze behaviors and the rapport in the interaction.

For the typing and gaze behaviors, this analysis focused on two characteristics of the data: 1) the total duration of a certain behavior in a time interval, and 2) the average duration for each time the action appears during that time interval. We applied a log-transformation to those behavior characteristics because their distributions were highly skewed. Employing a logarithm function on a skewed distribution will make it resemble a normal distribution[2]. Finally, due to the logarithm function performs differently between values smaller than one and values greater than one, we added one to our original data before applying the transformation to ensure consistent behaviors across our data ($log(time + 1)$).

[2]A separate analysis using the un-transformed data were conducted in Appendix D

2.5.1 Typing behaviors

This analysis aimed to confirm whether our stimulus worked as we intended such that the participants needed more time to construct their response using an Xbox controller than a keyboard. Our hypotheses were that 1) the total typing of AAC users was greater in the second conversation than the first interaction, 2), on average, the AU participants needed more time to construct each response using the Xbox controller than typing with the keyboard, and 3) the composing time, both total and average, in the first interaction and the second interaction were correlated to each other.

2.5.2 Gaze behaviors

For the gaze behaviors, we focused mainly on two instances: 1) when the CP participants looked at the AU participants′ face and 2) when the CP participants were inattentive (looking around the room). Besides analyzing the relationship of the gaze behaviors between the two interactions, we also looked at the change of those behaviors over the course of each interaction. We divided the 5-minute (300-second) interaction into three 100-second intervals and analyzing the relationship between the first 100 seconds and the last 100 seconds in the interaction.

2.5.2.1 The whole interaction

We expected the second interaction would produce longer waiting time than the first interaction. As a result, our first hypothesis would be that the CP participants would look at the AU participants less, both overall and on average, in the second interaction than the first interaction. Additionally, our second hypothesis was that the communication partner would be more inattentive, both overall and on average, in the second interaction than the first interaction.

Between the two conversations, we believed that the gaze behaviors of the CP participants would be consistent. Our third hypothesis was that there would be correlations in the gaze behaviors between the keyboard and Xbox conditions.

2.5.2.2 Within the interaction

We anticipated that the and awkward long pauses from using a speech synthesizer device would have a bigger impact on the gaze behaviors in the later state of the interaction than the beginning of the interaction. Hence, for both conversations, we first hypothesized that the communication partners look the AAC users less at the end of the interaction than at the beginning of the interaction. Our second hypothesis was that the communication partner would be more inattentive at the end of the interaction than at the beginning of the interaction.

Furthermore, we also expected that the gaze behaviors of the communication partner to be consistent across the interaction. This led to our third hypothesis that there would be correlations in the gaze behaviors between the start and the

end of both conversations.

2.5.3 Rapport

This analysis aimed to learn about how rapport changed between two interactions and between two partners. Our first hypothesis was that the rapport rating in the first interaction would be higher than the second interaction for both the AAC users and their partners because of the longer pauses in the second interaction. Secondly, we hypothesized that the AAC users would rate the rapport lower than their communicative partner in both interactions because they would have to communicate through a text-to-speech device. Next, the third hypothesis was that there would be a positive correlation in rapport rating between first and second conversation for both AU and CP participants. Finally, we hypothesized that there would be a positive correlation in the rapport ratting of the AAC users and their communicative partners in both interactions.

2.5.4 Relationship between gaze behaviors and rapport

It is suggested that rapport is associated with how much a person pay attention toward the other [19]. Our first pair of hypotheses was that there was a positive correlation between the amount of time, total and average, CP participants look at their partner and the rapport of both participants toward the interactions; and a negative correlation between the inattentive duration from CP participants and

how both participants rate the rapport from the interactions. As we anticipated that the stimulus, which is using an speech synthesizer, had not influenced the rating of rapport in both conditions yet. We hypothesized that there would be no correlation between the gaze behaviors of the CP participants and how both participants rated the rapport of each interaction. However, we believed that the gaze behaviors of the CP participants at the end of the interaction would indicate the rapport rating of both participants. Our last pair of hypotheses was that there was a positive correlation between the face gaze from the communication partners at the end of the conversation and the rapport rating for both interactions; and a negative correlation between the inattentive behaviors of the communication partners and how both participants rated their rapport in both interactions.

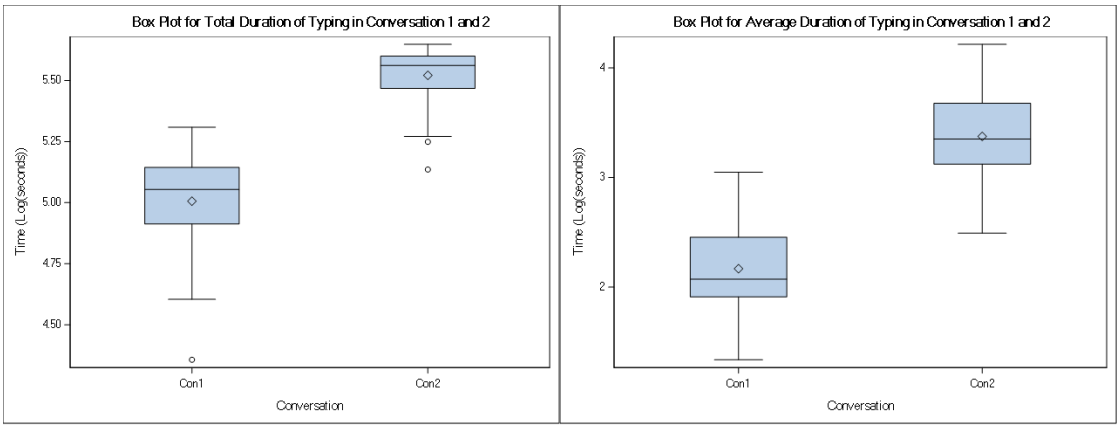
2.6 Results

2.6.1 Typing behaviors

The Figure 2.4 and Table 2.4 show descriptive information about the typing time in the first and second interaction.

Table 2.4: *Descriptive statistics on typing behavior in the whole interaction (log time)*

Interaction	Duration	N	Mean	SD
1	Total Duration	59	5.01	0.19
	Average Duration	59	2.17	0.39
2	Total Duration	59	5.52	0.11
	Average Duration	59	3.38	0.41



(a) Total duration of typing (b) Average time of typing

Figure 2.4: *Log time graph for typing behavior*

From Table 2.5 the AU participants spent significantly longer, in total, conducting the response in the second interaction than in the first interaction ($t(58) = 21.83, p < .001$). The average composing time for each response in the second interaction was significantly greater than the average composing time in the first interaction ($t(58) = 29.19, p < .001$).

Table 2.5: *Paired t-test on typing between the interaction (log time)*

	Mean	SD	DF	t-value	p
Total Duration (Int. 2 - Int. 1)	0.52	0.18	58	21.83	<.001
Average Duration (Int. 2 - Int. 1)	1.21	0.32	58	29.19	<.001

Table 2.6 shows that there was significant and positive correlation between the total composing time in the first interaction and the second interaction ($r(59) = 0.37, p < .01$); and between the average composing time in the first interaction and the second interaction ($r(59) = 0.68, p < .001$).

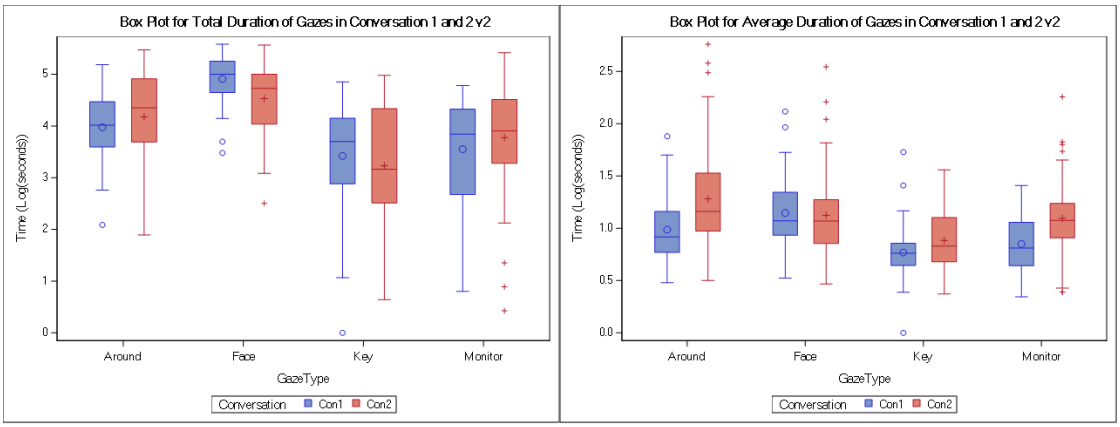
Table 2.6: *Pearson (r) correlation on typing between the interaction (log time)*

	N	r	p
Total Duration (Int. 2 and Int. 1)	59	0.37	.004
Average Duration (Int. 2 and Int. 1)	59	0.68	<.001

2.6.2 Gaze behaviors

2.6.2.1 The whole interaction

The Figure 2.5 shows us the box-plot of the four coded gaze behaviors: around (inattentive), face, keyboard, and monitor. However, we focused solely on the face gaze and inattentive gaze, which had the descriptive statistics displayed in Table 2.7.



(a) Total duration of gaze (b) Average time of gaze

Figure 2.5: *Log time graph for gaze behavior in the whole 5 minutes interaction*

Face Gaze

For the whole interaction, CP participants looked at the AU participants signifi-

Table 2.7: *Descriptive statistics on gaze behavior in the whole interaction (log time)*

Gaze Type	Int.	Duration	N	Mean	SD
Face gaze	1	Total Duration	59	4.91	0.48
		Average Duration	59	1.15	0.33
	2	Total Duration	59	4.53	0.67
		Average Duration	59	1.12	0.40
Around gaze	1	Total Duration	59	3.98	0.62
		Average Duration	59	0.99	0.28
	2	Total Duration	59	4.18	0.88
		Average Duration	59	1.28	0.51

Table 2.8: *Paired t-test on face gaze behaviors between the interaction (log time)*

	Mean	SD	DF	t-value	p
Total Duration (Int. 2 - Int. 1)	-0.38	0.44	58	-6.65	<.001
Average Duration (Int. 2 - Int. 1)	-0.02	0.22	58	-0.76	.449

cantly less in the second interaction than in the first interaction ($t(58) = -6.65, p < .001$). The duration of each time in which the CP participants looked at the AU participants in the second interaction was similar to that of the first interaction. (Table 2.8)

Table 2.9: *Pearson (r) correlation on face gaze behaviors between the interaction (log time)*

	N	r	p
Total Duration (Int. 1 and Int. 2)	59	0.75	<.001
Average Duration (Int. 1 and Int. 2)	59	0.83	<.001

From Table 2.9, there were strong and positive correlation between the duration of time that CP participants focused on the AU participants in the first and second

interaction for both overall time ($r(59) = 0.75, p < .001$) and average time ($r(59) = 0.83, p < .001$).

Inattentive Gaze

Table 2.10: *Paired t-test on inattentive gaze behaviors between the interaction (log time)*

	Mean	SD	DF	t-value	p
Total Duration (Int. 2 - Int. 1)	0.20	0.63	58	2.43	.018
Average Duration (Int. 2 - Int. 1)	0.29	0.34	58	6.62	<.001

Overall, CP participants were moderately more inattentive toward AU participants in the second interaction than in the first interaction ($t(58) = 2.43, p < .05$). The average time for each moment that the CP participants were unfocused in the Xbox controller condition was significantly longer than the average unfocused time in the keyboard condition ($t(58) = 6.62, p < .001$) (Table 2.10).

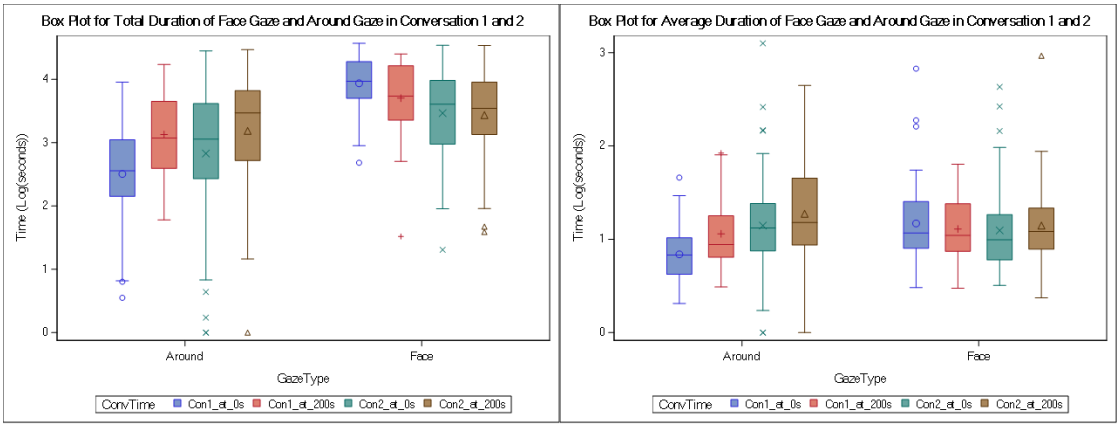
Table 2.11: *Pearson (r) correlation on inattentive gaze behaviors between the interaction (log time)*

	N	r	p
Total Duration (Int. 1 and Int. 2)	59	0.70	<.001
Average Duration (Int. 1 and Int. 2)	59	0.77	<.001

Table 2.11 showed us that there were strong and positive correlation between the duration of time CP participants did not paid attention toward the AU paticipants between the first and second interaction for both overall time ($r(59) = 0.70, p < .001$) and average time ($r(59) = 0.77, p < .001$).

2.6.2.2 Within the interaction

The Figure 2.6, and Table 2.12 shows descriptive information about the typing time in the first and second interaction.



(a) Total Duration of Gaze (b) Average Time of Gaze

Figure 2.6: *Log time graph for gaze behavior at the beginning 100s and at the end 100s of the two interaction*

Face Gaze

For the first interaction, the Table 2.13 shows us that the total time for face gaze behaviors of the communication partners at the end of the conversation was shorter than that of the beginning of the conversation ($t(58) = -5.48, p < .001$). On average, the duration of face gaze from the communication partners at the start of the interaction was similar to that of the end of the interaction.

For the second interaction, there were no differences in gaze behaviors of the CP participants between the beginning and the end of the interaction, for both total duration and average duration (Table 2.13).

There were strong and positive correlations between the total time that the

Table 2.12: *Descriptive statistics on gaze behavior within the first interaction and second interaction (log time)*

Int.	Gaze Type	Moment	Duration	N	Mean	SD
1	Face gaze	First 100s	Total Duration	59	3.93	0.41
			Average Duration	59	1.17	0.41
		Last 100s	Total Duration	59	3.70	0.57
			Average Duration	59	1.11	0.32
	Around gaze	First 100s	Total Duration	59	2.51	0.76
			Average Duration	59	0.84	0.27
		Last 100s	Total Duration	59	3.13	0.60
			Average Duration	59	1.06	0.35
2	Face gaze	First 100s	Total Duration	59	3.47	0.66
			Average Duration	59	1.09	0.45
		Last 100s	Total Duration	59	3.43	0.68
			Average Duration	59	1.15	0.40
	Around gaze	First 100s	Total Duration	59	2.83	1.08
			Average Duration	59	1.15	0.55
		Last 100s	Total Duration	59	3.19	0.93
			Average Duration	59	1.27	0.54

Table 2.13: *Paired t-test on face gaze behaviors within the first interaction and second interaction (log time)*

Int.	Duration	Mean	SD	DF	t-value	p
1	Total (Last 100s - First 100s)	-0.23	0.33	58	-5.48	<.001
	Average (Last 100s - First 100s)	-0.06	0.31	58	-1.49	.142
2	Total (Last 100s - First 100s)	-0.03	0.33	58	-0.72	.476
	Average (Last 100s - First 100s)	0.05	0.26	58	1.55	.126

communicative partners focusing at the AAC user at the start and at the end of first conversation ($r(58) = .82, p < .001$), and second conversation ($r(58) = .88, p < .001$) (Table 2.14).

From Table 2.14, the average time for each moment the conversational partners

Table 2.14: *Pearson (r) correlation on face gaze behaviors within the first interaction and second interaction (log time)*

Int.	Duration	N	r	p
1	Total (Last 100s and First 100s)	59	0.82	<.001
	Average (Last 100s and First 100s)	59	0.68	<.001
2	Total (Last 100s and First 100s)	59	0.88	<.001
	Average (Last 100s and First 100s)	59	0.82	<.001

looking at the AAC user′s face at the start of first and second interaction were strongly and positively correlated with that of the end of first ($r(68) = .68, p < .001$) and second interaction ($r(68) = .82, p < .001$), respectively.

Inattentive Gaze

Table 2.15: *Paired t-test on inattentive gaze behaviors within the first interaction and second interaction (log time)*

Int.	Duration	Mean	SD	DF	t-value	p
1	Total (Last 100s - First 100s)	0.62	0.47	58	10.14	<.001
	Average (Last 100s - First 100s)	0.22	0.26	58	6.47	<.001
2	Total (Last 100s - First 100s)	0.36	0.67	58	4.10	<.001
	Average (Last 100s - First 100s)	0.13	0.40	58	2.44	.018

From Table 2.15, CP participants spent significantly more time, overall, looking around the room at the end of the first conversation comparing to the beginning of the interaction ($t(58) = 10.14, p < .001$). On average, the communicative partners looked around longer at the end of the first interaction than at the beginning of the interaction ($t(58) = 6.47, p < .001$)

For the second interaction, the conversational partners spent more time, in total, wandering their gaze at the end of the conversation comparing to the start

of the conversation ($t(58) = 4.10, p < .001$). On average, the duration of each time the partners looked around the room at the end of the interaction was often longer than the start of the interaction ($t(58) = 2.44, p < .05$) (Table 2.15).

Table 2.16: *Pearson (r) correlation on inattentive gaze behaviors within the first interaction and second interaction (log time)*

Int.	Duration	N	r	p
1	Total (Last 100s and First 100s)	59	0.78	<.001
	Average (Last 100s and First 100s)	59	0.67	<.001
2	Total (Last 100s and First 100s)	59	0.79	<.001
	Average (Last 100s and First 100s)	59	0.73	<.001

The Table 2.16 shows us that there were strong and positive correlations between the total times that the conversational partners were unfocused at the beginning and at the end of the keyboard interaction ($r(59) = .78, p < .001$); and between the total times that the partners were inattentive at the beginning and at the end of Xbox controller interaction ($r(59) = .79, p < .001$).

The average time of each moment the conversational partners of the AAC users looking around at the room at the beginning of first interaction and second interaction were significantly and positively correlated with that of at the end of first interaction ($r(59) = .67, p < .001$) and second interaction ($r(59) = .73, p < .001$), respectively (Table 2.16).

2.6.3 Rapport

The Graph 2.7 and Table 2.17 displayed the descriptive statistics on rapport ratting for both participants in the interaction.

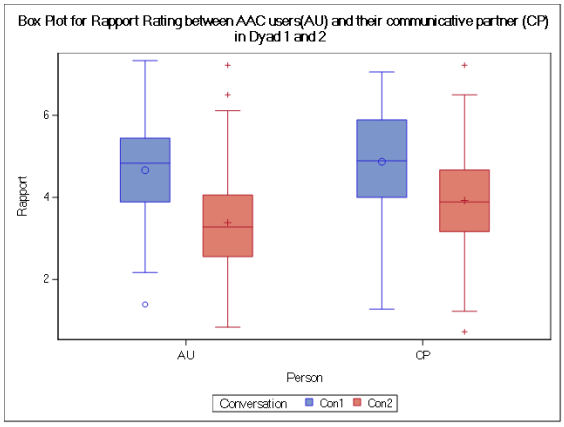


Figure 2.7: *Rapport rating graph*

Table 2.17: *Descriptive statistics on rapport*

Role	Interaction	N	Mean	SD
CP	1	59	4.87	1.15
	2	59	3.92	1.24
AU	1	59	4.66	1.25
	2	59	3.38	1.36

The communicative partners and the AAC users rapport ratings in the second interaction were less than the rating in the first interaction (CP: $t(58) = -7.60, p < .001$; AU $t(58) = -9.34, p < .001$). There were no differences in rapport rating between the communicative partners and the AAC users in the first conversation. However, the t-test suggested that the AAC users felt worse in the

Table 2.18: *Paired t-test on rapport between the interactions and between the participants*

	Mean	SD	DF	t-value	p
Int. 2 CP - Int. 1 CP	-0.94	0.95	58	-7.60	<.001
Int. 2 AU - Int. 1 AU	-1.28	1.05	58	-9.34	<.001
Int. 1 AU - Int. 1 CP	-0.20	1.46	58	-1.08	.286
Int. 2 AU - Int. 2 CP	-0.54	1.70	58	-2.46	.017

second interaction comparing to their partners ($t(58) = -2.46, p < .05$)) (Table 2.18).

Table 2.19: *Pearson (r) correlation on rapport between the interactions and between the participants*

	N	r	p
Int. 1 CP and Int. 2 CP	59	.68	<.001
Int. 1 AU and Int. 2 AU	59	.68	<.001
Int. 1 AU and Int. 1 CP	59	.26	.040
Int. 2 AU and Int. 2 CP	59	.15	.270

From Table 2.19, there were strong and positive correlations in rapport rating between first interaction and second interaction for both the AAC users ($r(59) = .68, p < .001$) and their communicative partners ($r(59) = .68, p < .001$). There was a moderate and positive correlation in rapport rating between the AAC users and their partners in the first conversation ($r(59) = .26, p < .05$). However, the rapport rating between the AAC users and their partners were similar in the second interaction.

2.6.4 Relationship between gaze behaviors and rapport rating

2.6.4.1 Face gaze and rapport

First Interaction

Table 2.20: *Pearson (r) correlation between rapport and face gaze behaviors in first interaction*

			Rapport of CP		Rapport of AU	
		N	r	p	r	p
Whole interaction	Total	59	.31**	.016	.20	.132
	Average	59	.19	.148	.16	.227
Start of interaction	Total	59	.27**	.039	.16	.220
	Average	59	.20	.122	.14	.282
End of interaction	Total	59	.30**	.021	.16	.221
	Average	59	.15	.261	.19	.152
** p <.1; ** p <.05; *** p <.01*						

Table 2.20 shows a moderate and positive correlation between the rapport rating of the CP participants in the first interaction and the total amount of time that they focused at the AU participants in the first interaction ($r(59) = .31, p < .05$). The rapport rating of the CP participants was significantly and positively correlated with the total time that they focused at the AAC user at the beginning ($r(59) = .27, p < .05$) and at the end ($r(59) = .30, p < .05$) of the first conversation. There were no correlations between the average duration of face gaze in the first interaction from the CP participants and their rapport rating.

Overall, the face gaze behaviors of the communicative partners were not correlated with the rapport of the AAC users.

Second Interaction

Table 2.21: *Pearson (r) correlation between rapport and face gaze behaviors in second interaction*

			Rapport of CP		Rapport of AU	
		N	r	p	r	p
Whole interaction	Total	59	.19	.147	.11	.407
	Average	59	-.02	.900	-.05	.726
Start of interaction	Total	59	.19	.154	.03	.798
	Average	59	-.01	.945	-.11	.389
End of interaction	Total	59	.18	.169	.13	.337
	Average	59	-.04	.737	.003	.985
			** p <.1; ** p <.05; *** p <.01*			

The Table 2.21 shows no correlations between face gaze behaviors of the communicative partners in the second interaction and the rapport rating of both participants.

2.6.4.2 Inattentive gaze and rapport

First Interaction

The Table 2.22 shows that the rapport rating of the communicative partner in the first interaction was negatively and significantly correlated with the average time that they were unfocused ($r(59) = -.38, p < .01$). Additionally, there was a weak correlation between the rapport rating of the communicative partners and their overall inattentive behaviors in the first interaction ($r(59) = -.23, p < .1$).

At the beginning of the first interaction, the rapport rating of the CP participants were moderately and negatively correlated with the average duration of

Table 2.22: *Pearson (r) correlation between rapport and inattentive gaze behaviors in first interaction*

			Rapport of CP		Rapport of AU	
		N	r	p	r	p
Whole interaction	Total	59	-.23*	.081	-.02	.886
	Average	59	-.38***	.003	-.19	.160
Start of interaction	Total	59	-.21	.103	-.05	.681
	Average	59	-.29**	.026	-.19	.153
End of interaction	Total	59	-.28*	.035	-.07	.592
	Average	59	-.42***	<.001	-.14	.289
** p <.1; ** p <.05; *** p <.01*						

inattentive behaviors of them ($r(59) = -.29, p < .05$); and were weakly and negatively correlated with the total inattentive time of the communicative partner ($r(59) = -.21, p = .1$).

At the end of the fist interaction, the rapport rating of the communicative partners were strongly and negatively correlated with the average duration of the inattentive gaze from the CP partners ($r(59) = -.42, p < .01$); and were moderately and negatively correlated with the total inattentive time of the CP participants ($r(59) = -.28, p < .05$).

Overall, the inattentive gaze behaviors of the communicative partners were not correlated with the rapport of the AAC users in the first interaction.

Second Interaction

The rapport rating of the communicative partners in the second interaction was negatively and significantly correlated with the average time that they spent looking around in the second interaction ($r(59) = -.36, p < .01$). Additionally, the rap-

Table 2.23: *Pearson (r) correlation between rapport and inattentive gaze behaviors in second interaction*

			Rapport of CP		Rapport of AU	
		N	r	p	r	p
Whole interaction	Total	59	-.18	.168	.04	.768
	Average	59	-.36***	.006	-.12	.368
Start of interaction	Total	59	-.09	.495	.11	.403
	Average	59	-.30**	.023	-.03	.841
End of interaction	Total	59	-.19	.148	.04	.771
	Average	59	-.32**	.013	-.05	.703
** p <.1; ** p <.05; *** p <.01*						

port rating of the communicative partners significantly and negatively correlated with the average time they were inattentive at the beginning ($r(59) = .30, p < .05$) and at the end ($r(59) = .32, p < .05$) of the second conversation. There were no correlations between the total duration of inattentive gaze in the second interaction from the CP participants and their rapport rating (Table 2.23)

Finally, the inattentive gaze behaviors of the communicative partners were not correlated with the rapport of the AAC users in the second interaction.

2.7 Discussion

2.7.1 Typing behavior

Our stimulus was working as we intended, and the AC participants needed to use significantly more time (around 22.26 seconds more for each typing moment [3])

[3]From Table D.2 in Appendix D

to construct their responses to using the Xbox controller instead of the keyboard. Additionally, if the participants struggled with using the keyboard, they were likely to perform worse in the second interaction than other participants.

2.7.2 Gaze behavior

The face gaze and inattentive gaze behaviors of the communication partner were consistent across and between interactions

Face gaze Between the two interactions, CP participants gazed at the AU participants around 38.14 seconds less, in total,[4] in the second interaction than the first interaction. However, the duration of each time the CP participants looked at the AU participants was similar between the two interactions. This suggested a decrease in the frequency of how often the communicative partner looked at the AAC user from interaction 1 to interaction 2. Looking at the first interaction, the data also suggested a decrease in the frequency of how often CP participants look at the AU participants from the start of the interaction to the end of the interaction. Furthermore, this showed that the average gaze duration was consistent and around 2.36 seconds [5], which was a little less than the number reported from Argyle and Ingham [29] (2.95 seconds).

[4]From Table D.5 in Appendix D
[5]From Table D.4 in Appendix D

Inattentive gaze Between the two interactions, CP participants were attentive toward their partner around 24.72 seconds more, in total,[6] in the second interaction than the first interaction. Additionally, CP participants, on average, spent around 1.39 seconds more looking around the room in the second interaction than the first interaction. Within both interactions, the communicative partners were less attentive near the of the end of the conversation than at the beginning of the conversation. Additionally, each distracting moment lasted longer at the end of the interaction than at the beginning of the interaction.

In conclusion, the long pauses in the second interaction seemed to only have an effect on the inattentive behaviors of the communicative partners than on how often they focused on the AAC users. They might not know where to focus when waiting for responses from the AAC users.

2.7.3 Rapport

As we hypothesized, both participants felt worse about the conversation in the Xbox condition than the keyboard condition due to the long pauses in the Xbox condition.

Both participants felt similar to each other in the first interaction, which suggested that using the keyboard to communicate did not have any major impacts to the interactions. However, typing with the Xbox controller influenced the AAC users and made them felt worse about the interaction than their partners.

[6]From Table D.12 in Appendix D

While both participants felling toward the interaction was associated with each other in the keyboard condition, they were completely out of sync with each other in the Xbox controller condition. We expected the complexity of typing with an Xbox controller made it difficult for both participants to connect with each other.

2.7.4 Relationship between rapport and gaze behaviors

In the first interaction, the CP participants looked at their partner less often and gaze around longer if they felt less rapport toward the interaction. In the second interaction, only the inattentive gaze was associated with the CP participants′ feeling toward the interaction. Against our hypothesis, the effect of the speech synthesizer on the communicative partners had already happened in the first 100 seconds of both conversations.

Interestingly, the gaze behaviors of the CP participants had no influences on how the AAC users felt toward the interaction. An explanation for this result could be that the participants were not used to communicate with a speech synthesizer device and had to pay additional attention to composing the response instead of on their partners. Furthermore, as the visual attention of the AU participants were majorly occupied by the typing tasks, the verbal cues from the CP participants might become a larger influence on how the AAC users felt toward the interaction.

2.7.5 Limitations and future work

The analysis suggested that the frequency would be an important characteristic of the gaze behaviors. However, this was not included in this analysis and should be looked at in the future. Additionally, further examination on the gaze behaviors of the communicative partners while waiting for the AAC users′ response would be needed, because the long pauses were the unique aspect of this kind of interactions.

Another major limitation of the study was that the AAC users were not familiar with the technology. The unfamiliar and laborious typing task could potentially impact how participants felt and behaved in our study. A future study with people that had experiences with AAC devices would be needed to get a clearer picture of how speech synthesizer is used in daily conversation.

Chapter 3 Second Study: Gathering Ideas from the Users

The study was a 90-minute focus group and aimed to understand how the behaviors of the communication partners in the first study affect people that had experience with speech synthesizers. Participants of the focus group watched clips from the first study and discussed actions of the communication partners in the clips. The topic of discussion included: what behaviors of the communication partners were good; what behaviors of the communication partners were bad; and how a person could help to signal the communication partners to stop doing those bad behaviors. Afterward, an experimenter gave them a brief introduction about assistive robots and then started a new discussion about how a robot could nudge the

communication partners to behave better.

3.1 Materials

3.1.1 Survey

A demographic survey was used to get the basic information of the participants. Additionally, the survey was designed to have a lot of white space around each answer because we wanted to accommodate people that lost the ability to control their fine muscle and could not precisely circle the answer (Appendix B).

3.1.2 Videos

The study used six videos from the first study. The videos were picked by us under three criteria: 1) the rapport rating of the participants in those video should be diverse because we wanted to capture both bad and good interaction; 2) the actions of the participants in those video should be largely different to each other because the author believed it would give a diverse behaviors of the communication partners; 3) they were solely from the second interaction because they had a lot of long and awkward pauses (Table 3.1).

Table 3.1: *Rapport rating of videos used in the study*

	Rapport	
Video ID	AU	CP
12	3.11	4.44
33	1.17	5.39
36	2.50	3.67
62	2.72	4.00
67	4.17	3.44
79	2.89	5.11

3.2 Participants

For this study, we explicitly recruited participants who had experiences with AAC devices through either using or interacting with it. There were ten participants (4 males and 6 females, $M_{age} = 55.3$) came to two focus groups: nine people in the first focus group and one person in the second group. The groups included four people with ALS, four caretakers, and two people that were neither of them at the moment but had experiences with speech synthesizer devices. Participants were recruited from an ALS support group through the group organizer. They received $20 for their participant in the study.

3.3 Procedure

The first focus group took place in a conference room at Salem Hospital and the second focus group took place at OSU. Participants were seated around a table with a moderator (the author) in the middle . As the moderator was not familiar

with jargon in the ALS community, the support group organizer, who had spent a long time working with ALS patients, had agreed to help as a translator for the discussion. The role would include explaining the ALS jargon, rephrasing the discussion questions from the moderator to make it more related to the participants, and repeating the responses from participants that had trouble verbalizing their responses.

When all of the participants had arrived at the determined location, the experimenter went over the consent form again with the participants and made sure that the participants understand the consent form. Next, the participants completed a demographic survey (Appendix B). Before starting the discussion, the experimenter looked at the survey to make sure that all of the participants were comfortable being recorded [1]. If any of the participants said no in the survey, we would close the lid of all of our camera, except the one that focused on the experimenter. The reason was that we wanted to link between the experimenter behaviors and the showed stimulus (the videos), with the comments from the participants. Finally, the researchers asked the participants to only use their first name in the discussion to maintain the privacy of the participants.

The study officially started with the participants watched interaction clips from the first study. Throughout the clips, the experimenter would ask the participants to identify important moments or behaviors. Key moments or behaviors were defined as the moments or behaviors from the communicative partner in the clips

[1]The consent explicitly mentioned that the whole discussion would be audio recorded. The video record is optional and it only happens if all of the participants consented to its

(a) Fetch Robot (b) Blossom Robot

Figure 3.1: *Robots in the second study*

that were deemed inappropriate or appropriate from the participants. For each inappropriate moments or behaviors, the whole group would discuss why those behaviors were inappropriate and how a person typically fix those situations. Depending on how many ideas were generated, the experimenter might not be able to discuss all videos. The videos discussion would stop after forty-five minutes or an hour. Next, the focus group would start generating ideas about how a robot could be used to mediate those inappropriate behaviors. The experimenter gave a brief presentation about assistive robots and the available robots in our lab, which was the Fetch and Blossom robot, that could be used for this study (Figure 3.1). The purpose of the presentation was to let the participants get some basic ideas about the appearance and capability of the robots. Then the group discuss about the designs of the robot and potentially actions that a robot could do to mediate those inappropriate actions mentioned in the previous section. At the end of

the study, the experimenter would give the participants their compensations and thanked them for sharing their thoughts.

3.4 Coding Process

The purpose of this coding schema was to record how the group think about the behaviors of the communicative partners in the clips and their ideas on how technologies could improve the bad interactions.

The transcripts of both focused groups were made by the author and a lab member. We used a categorizing strategy mentioned in Maxwell's *Qualitative research design* book to develop our coding scheme in the list below [34].

1. **Behavior**: Verbal and nonverbal communication actions of the conversation partners in the clips that the participants in the focus group like and dislike, such as: looks, smile, yawn, ... There are three coding subcategories: Good, Bad, and Neutral.

2. **Attitude**: The participants' comments about the feelings or emotions displayed by the communication partners in the clips, such as: happy, bored, interested, ... There are three coding subcategories: Good, Bad, and Neutral.

3. **Limitation**: Any comments of the participants on the limitation of the stimulus, such as: video quality, characteristics of the people in the clips, ...

4. **Improvement**: Any comments of the participants on how the interaction can be improved either from a person, a robot, or an AAC device function.

There are three coding subcategories: person, robot, and AAC device.

5. **Design/Implementation**: Practical design and implementation ideas of the robot such as sizes and locations of the robot.

6. **Miscellaneous comments**; Any interesting comments that do not belong to the other categories.

3.5 Result and Discussion

In this section, I will focus on the good and bad behaviors of the communications partners. Then, I will talk about the design ideas for the the robot and ideas for additional feature for the speech synthesizers.

3.5.1 Good behaviors

The people in the focus group seemed to prefer communicative partners to actively lead the conversations. They would like the communicative partners to "[take] control of the [interaction]" and "anticipate [the ACC] next questions and answer that as part of [their] response". That would be extremely useful at the beginning of the interaction when the two people were trying to understand each other. As we expected, the participants seemed to appreciate attentive behaviors such as "maintain[ing] eye contact" and leaning toward the AAC user. However, as it took a lot of effort and time to compose a response, the people in the focus group highly valued the genuineness and empathy from the communicative partners in the clips.

They wanted the partners to only "ask what [the partners] really care" to know about the AAC users, be "truly interested in [the AAC users]", and be patient for the response.

3.5.2 Bad behaviors

People in the focus groups seemed to disapprove of the impatience and indifference from the communicative partners. A person in the focus group was frustrated to see the communicative partners "[kept] looking at the monitor screen and [tried] to guess the word" every few seconds. They also did not like it when the communicative partners were looking around and inattentive toward the interactions. Furthermore, the focus group had strong opinions when the communicative partners laughed at the effort to communicate by the AAC users.

3.5.3 Design ideas for robot

The participants preferred the robot to be small, portable, and attachable to their wheelchair. The big robot would "draw a lot of attention" and might interfere with their daily activities. Interestingly, participants mainly fixated on how to cue the communicative partners using "the eyes" of the robots. It was suggested that the "eyeball could flash and then move ... to refocus someone′s attention". The groups were a little skeptical about letting the robot verbally tell the communicative partners to focus on the AAC users because it could easily be perceived as being

rude. A interesting idea coming from the discussion was that the robot should not only prevent bad behaviors but also promote positive behaviors. Furthermore, it was suggested from the discussion that verbal cues would be better for positive feedback and nonverbal cues would be preferred for negative feedback. Lastly, a participant mentioned that:

> "She was just too nervous and then make the whole interaction bad between two people. So, like if somehow we can make her less nervous and less uncomfortable when interacting with the AAC users, then it would be really nice, they will have a better interaction."

Therefore, it would be preferable if the robot could shorten the nervousness and discomfort of the novice communicative partners at the beginning of the interaction.

3.5.4 Design ideas for features of speech synthesizers

The "quickfire" function was highly suggested throughout the discussion. It was the feature that allowed the users to pre-compose common responses and save it to a hot-key in their AAC users. The group also suggested to "forget perfect English and drop it down to the basics" while typing out responses. An algorithm that could predict a whole sentence from a few words would be a potential solution for this problems.

As for the long pauses caused by using the AAC device, there were two ideas coming from the discussion. The first idea was that the communication partner

could see the monitor of the AAC device and help guess what the AAC user was trying to say. However, it was noticed that it was only done with someone the AAC users knew and trusted because there were some communicative partners who made guesses in every second. A feature that could speak out comprehensive phrases or filler words in a response composing by an AAC users could be a solution. The reason was that it still allowed the communicative partners to keep up with the AAC users while lowering the risk of them being disruptive and invasive. Secondly, there could be a waiting screen showed up on the back of the AAC device such as the flashing "dot dot dot" bubble while you were waiting for a text.

Finally, the researcher noticed that the group truly disliked the impatient and rude communication partners. However, they also realized that not everyone could pay full attention to the AAC users in every moment of the interaction. Therefore, if we could develop a feature on either a robot or an AAC device that could get the communicative partners attention back right before the speech synthesizer verbalized the responses, for instance: a “ding before the [AAC user] speaks”.

3.5.5 Limitations of the study

A few people in the group mentioned the quality of the voice from the speech synthesizer used in the video clips. However, we believed that it did not have a lot of impact on the communicative partners in those critical moments in which they were just waiting for the response from the AAC users.

Another drawback of the study was the size of the group. Two people in the

groups, who had trouble speaking and used an AAC device instead of just interacted with, were not able to give a lot of feedback in the discussion. We had tried to eliminate this by conducting the study with people in the same focus group, who knew how to interact around each other. However, we believed our group was a little crowded to get any in-depth feedback from the actual AAC users.

3.6 Summary

For the design ideas of a mediator robot, the group would prefer to have a small and friendly appearance. It was suggested that the robot should give verbal feedback for positive behaviors and nonverbal feedback for negative behaviors. Additionally, it would be nice if the robot could shorten the nervousness of the novice communicative partners at the beginning of the interaction.

For developing features of the AAC devices, the first idea would be an algorithm that could generate a whole sentence based on a few keywords. Next, we should consider another feature that could speak out comprehensive phrases or filler words in a response composing by an AAC user.

Overall, the group wanted the technology to stop bad behaviors and encourage good behaviors from the communicative partners, to make the waiting time shorter and less awkward, and to signal the communicative partners before the speech synthesizer starts to speak.

Part III

Design and Validate the Prototype

This chapter focuses on designing technologies from the suggestions of the AAC users in the previous state. The background research and implement of those ideas are in the first section. The next section describes another focus group study to validate those designs.

Chapter 4 Creating the Prototype

The first section of this chapter covers the background research for choosing the appearances and behaviors of the robots. Then, we explain the implementation of the algorithm used in our feature for the AAC device in the second section.

4.1 Choosing the appearance of robots

In this section, we will first look at the traditional intervention techniques that are used in conversation with a speech synthesizer. The next part will look at some prior work on robots that are used to mediate human-human interaction. Finally, we will describe the two prototypes that we have chosen in the last section.

4.1.1 Traditional intervention techniques

The traditional intervention techniques require the communication partners to participate in a communication program tailored to AAC users [17] [15] [35] [36]. There are two main techniques that are incorporated into the training. The first technique uses the ImPAACt program, which is a communication partner instructional pro-

tocol [35]. The training focuses on four types of prompting techniques to encourage AAC users to take their time for typing out the responses. The program also applies a contingent responding technique to reinforce communicative attempts and supports utterance expansion from the AAC users.

The second technique is the Milieu Teaching techniques [36] used for children with neurodegenerative diseases. The descriptions of all the techniques can be accessed in Ann Kaiser and Courtney Wright's paper. However, there are some relevant techniques that can be implemented into a robot such as respond to child communication using the child's mode and words, imitate child actions and model with the AAC mode, using mirroring and mapping including the ACC: Imitate child actions and model with the AAC mode and spoken words.

Those programs require the caretakers or the communication partners to go through a long training process [35] [36]. Therefore, it would not be applicable in the situation where the AAC users need to talk with a stranger. By using a robot as a mediator, we want to nudge the communicative partners toward the right behaviors without going through the long training process.

4.1.2 Related work on mediator robot in human-human interaction

Robots as a mediator in a human-human interaction is a new research area with multiple potential perspectives to explore. Most of the work on robots in human-human interaction tends to focus on how the robot can fit in a human group [37] [38]. As the robot's goal is to maintain the dynamic of the group, it only carries out

actions that fit our social norm and does not take any initiative toward improving the interaction between human in the group.

KIP1, designed by Hoffman et al. is one of the first social robots that aim to remedy the conflict in human-human conversation. By using implicit cues such as nonverbal behaviors, KIP1 can make people aware of their own behaviors and, hence, nudge the people to fit their behaviors without compromising the natural communication flow between the two people [39].

On the other hand, another robot designed by Ronald Arkin's group uses an explicit approach, such as verbally asking an individual to do certain actions. The robot's job is to intervene in the conversation if it detects conflicts between the patient and the caregiver [40] [41].

There seems to be two type of robotics design that either has a non-humanoid appearance or resemble a person. The actions from non-humanoid robot tend to be implicit and less have minimal impact on the natural flow of the interaction [39]. Adversely, the behaviors and expressions of a humanoid robot are easy for a human to read and understand as it can be used to teach emotions expression for kids with autism [42] [43]. Following the participatory design in the previous section, we create two different robot prototypes: a non-humanoid robot and a humanoid robot, to ask the AAC users about their preference.

4.1.3 The robot prototypes

4.1.3.1 Non-humanoid robot: Blossom robot

The robot is designed and developed in Dr. Guy Hoffman's lab at Cornell University [44]. I choose this robot because of several reasons which are that anyone can build this robot, and it is cheap and easy to customize. Finally, is that the robot is small, which is a characteristic that the participants in the first focus group wanted. Figure 4.1 is a Blossom Robot that has been built in our lab.

Figure 4.1: *Blossom robot*

4.1.3.2 Humanoid robot: Furhat avatar

We choose an interactive avatar for the humanoid robot because the software is available for us and it can be small depending on the monitor [1]. The avatar is developed by Dr. Gabriel Skantze at KTH Royal Institute of Technology in Sweden [45] (Figure 4.2). We pick an avatar display because some robots do use a displayed screen for their face such Baxter and Furhat.

Figure 4.2: *Furhat avatar*

4.2 Potential algorithms for improving the typing speech

To shorten the long and awkward pauses in conversation with AAC users, we propose an idea that the AAC device can automatically vocalize portions of the

[1] This specific avatar robot is suggested by Dr. Olov Engwall at KTH Royal Institute of Technology, Sweden

user's written response after a certain time. Furthermore, the spoken units to be comprehensible and sensible. It is similar to how people usually speak in their daily life, we often just speak out comprehensive phrases as we simultaneously construct the rest of our response. To achieve this, we investigate methods for "chunking" sentences into understandable chunks or phrases. Since the sentences will not be fully constructed when the AAC device needs to vocalize understandable chunks, we specifically investigate methods for chunking incomplete sentences.

There exist many methods for chunking complete sentences and one of the most common methods is noun phrase chunking (NP-chunking) in which the algorithm search for chunks corresponding to individual noun phrases. There are multiple approaches to NP-chunking task such as graphical models or support vector machine(SVM). For graphical models, the majority of the works have been done on two popular models: Hidden Markov Model (HMM)[46] and Conditional Random Field (CRF)[47]. The HMM model made predictions on whether a word belongs to a phrase by looking at preceding words. In contrast, the CRF model looks not only at preceding words but also the proceeding words. While the graphical model gives a generative model approach to our problem, SVM shows a different direction by applying the discriminative model to the chunking task[48]. The algorithm simply looks at the entire input that is the incomplete sentence in this case and categorize the words in that sentence as to whether it is the beginning of a new phrase or a part of an ongoing phrase.[2]

[2]The author collaborated with Christopher Eriksen from Oregon State University and Wuga at the University of Toronto to examine the accuracy of those three algorithms in a class project. Appendix F is a copy of the development and testing the algorithms of the project.

In this project, we chose the CRF algorithm as our method to develop because it has better accuracy than HMM (Appendix F) and because we need a real-time prediction, which is hard to achieve with SVM. We use a software toolkit named CRF++[49] for our CRF chunking because it is one of the best available toolkits for working with CRF algorithm.

Chapter 5 Third Study: Validating the Design

We wanted to create a study to gather feedback about our prototypes that we were working with. The focus group was designed to get ideas about appearance and behavior of two different type of robots (Blossom robot and Furhat avatar); and a feature for the AAC device. We assumed that some of the participants were not in the first focus group, and created a summary video to get them up-to-date. The video consisted of the first thirty or forty-five seconds of the five clips in the prior focus group. The goal of the videos was to show the participants the bad, common, and good behaviors of typical communicative partner while interacting with an AAC user. Afterward, the moderator would show the first robot design and ran a demo to show the capabilities of the robot. If it was the Blossom robot, the demo would consist of the robot as a "movie partner" in which the robot would "watch" the movie with you and react to the scenery and emotions of the characters in the movie. If it was the Furhat avatar, the demo would be a guessing game in which the group would guess a number from one to ten and the robot would give a response after each guess. After the first demo, we controlled

the robot to display three different emotions (anger, sad, and happy). We picked those three emotions because we wanted the robot to display anger emotion if the communicative partners were being rude such as disrupting or ignoring the AAC users; sad emotion if the communicative partners unintentionally misbehaved such as looking at the floor while waiting because they did not know whether they should look at the AAC users; and happy emotion if the communicative were attentive and the AAC users wanted to reinforce those behaviors. Then, the moderator would give out the Impression toward Robot Survey to the participants (Appendix C). Next, the moderator would do the same things for the second robot design. The demo and the emotional expressions of both robots that used in this study were made by the original creator. We used the original version from the creator instead of our own version because we wanted to avoid our bias to either the designs. Between the focus group, we tried to counterbalance the order of our demo based on the number of attendees in each group.

The discussion about what the participants like or dislike about each of the designs started after the two demos. We stopped the discussion twenty minutes before the end of our ninety-minute study. Then, we gave the participant the Ranking Characteristics of Robots survey (Appendix C). The survey was used to cover certain characteristics of the robot that we missed during the discussion. After the participants had finished the survey, we showed them a video demo of the phrase-chunking feature for the AAC device. The moderator reminded the participant to give feedback on the feature using the online survey that would be sent out later in the day. The online survey was created in case we did not have

time to talk about the program. We discussed the feature until the end of the study.

5.1 Materials

5.1.1 Surveys

We incorporated four different surveys in this study. The first survey was a demographic survey, which collected basic information about the participants. Next, we wanted to measure the impression of the participants toward our robots. We looked at three difference scale: the Negative Attitudes towards Robots Scale (NARS) [50], the Godspeed Questionnaire [51], and the Robotic Social Attributes Scale (RoSAS) [52]. The NARS looks at the negative attitudes toward 1) situations and interactions with robots, social influence of robots, and emotions in interaction with robots. The Godspeed Questionnaire measures perceived safety, perceived intelligence, likeability, animacy, and anthropomorphism. Lastly, warmth, competence, and discomfort are measured in the RoSAS. Because we did not have enough time in our study for the participant to do all three surveys, we chose to use the RoSAS in our study. We believed that the constructs in the RoSAS were closely related to what we wanted to know from the users. Additionally, our population included people that had lost fine motor skills and would need more time to fill out a survey than typical participants. As the result, we decided to use only three items from each subcategory in the RoSAS. The total scale included ten items: emotion, social,

and compassionate for the warmth subcategory; interactive, reliable, and capable for the competence subcategory; scary, awkward, and aggressive for the discomfort subcategory; and an extra question asking about the usefulness of the robots. Our third survey was a ranking survey which listed various of characteristics of the robot and asked participants to rank them. Finally, we had a short online survey to gather final thoughts about our feature for the AAC device (Appendix B).

5.1.2 Video

We combined the first thirty or forty-five seconds of the five clips that were in the prior focus group together. The first two clips displayed the behaviors that the prior group indicated as rude behaviors. The communicator partner in the next two clips displayed typical behaviors. The focus group thought the some of partners′ behaviors, such as looking down for too long, were inappropriate but acceptable because they knew the partner did not mean to be rude for the AAC users. Finally, the last clip was the one that the focus group enjoyed the behaviors of the communication partner.

5.1.3 Prototype

5.1.3.1 The emotions displayed by robot behaviors

For the Blossom robot, the robot displayed anger behaviors by turning sharply to one side and the other. The sadness was portrayed by the robot slowly looking

down. Finally, The robot moved its head up and down rapidly as a display of happiness.

Figure 5.1 was a snapshot of the three emotions displayed by the avatar.

Figure 5.1: *Facial expression of the avatar*

5.1.3.2 The feature for the AAC device

For this project, we used a CRF++[49] for our CRF chunking algorithm. The algorithm is set to speak after fifteen seconds. We prerecorded how the program worked to make sure that our demo would be consistent across multiple focus groups [1].

5.2 Participants

For this study, we also explicitly recruited participants who had experiences with AAC devices through either using or interacting with it. There were thirteen participants (5 males and 8 females, $M_{age} = 53.8$) came to four focus groups: two people in the first focus, three people in the second and third focus group, and five

[1] This is the link for the video demo: `ttps://drive.google.com/file/d/1rP6vq9Yn8F4vBZMUgGo6ZDREcz6dPicR/view?usp=s`aring

people in the last focus group. The population included five people with ALS, six caretakers, and two people that were neither of them at the moment but had experiences with speech synthesizer devices. Participants were recruited from an ALS support group in Salem and Eugene through the group organizer. As they have to drive to Oregon State University from either Salem or Eugene to participate in the study, we compensated them $100 for each family group, which tends to consist of an ALS patient and a caretaker.

5.3 Procedure

The study took place in a conference room at Oregon State University. Similar to the second study, the study had the support group organizer as an assistant. When all the participants arrived at the room, the moderator gave a little introduction about the study. The short video was showed to summarize the first focus group. Afterward, the moderator brought out the first robot prototype and started the demo programs. The program consists of an interactive demo from the original creator and the robot expresses three emotions: anger, sad, and happy. The Impression toward Robot Survey was given to the participants afterward. When the participant finished the survey, the moderator started the demo for the other robot prototype. Like the first demo, the Impression toward Robot Survey was given to the participants after the demo. Then, the experimenter started the discussion on what participants like or dislike about each prototype and how it could be improved. The discussion was stopped twenty minutes before the end of the

study and the experimenter gave the participants the Ranking Characteristics of Robots Survey. The video on the feature for the AAC device was showed after the survey. The moderator reminded the participants to write down feedback on the feature using the online survey that would be sent out later in the day, and then discussed the feature until the end of the study. Before the participants left, the experimenter gave the participants their compensations and thanked them for sharing their thoughts.

5.4 Pre-analysis for the RoSAS surveys

The RoSAS in this study did not have all of the items and should be checked for the relationship between each item in the subcategory using principal component analysis. However, it is recommended from the principal component analysis chapter in *A step-by-step approach to using SAS for factor analysis and structural equation modeling* that "the minimal number of subjects providing usable data for the analysis should be the larger of 100 subjects or five times the number of variables being analyzed" [53]. We only had 26 responses total (each participant did the survey two times) for a total of nine variables in our study. Therefore, we decided to only calculate the correlations between items in each category. This would allow us to have a rough estimate of whether the items were still measured the same construct.

Warmth:

The Table 5.1 shows us the descriptive statistics of the items in the warmth

Table 5.1: *Descriptive statistics on items in warmth component*

	N	Mean	SD
Emotion	26	5.08	1.65
Social	26	4.88	1.80
Compassion	26	4.00	1.70

Table 5.2: *Pearson (r) correlations between items in warmth component* ($N = 26$)

	Emotion	Social	Compassion
Emotion	1		
Social	.72***	1	
Compassion	.59***	.56***	1
	** p <.1; ** p <.05; *** p <.01*		

subcategory. From Table 5.2, it showed that all of the components were highly correlated to each other. Therefore, those items seemed to belong in the same construct.

Competence:

Table 5.3: *Descriptive statistics on items in competence component*

	N	Mean	SD
Interactive	26	4.88	1.86
Reliable	26	5.00	1.41
Capable	26	5.35	1.55

The Table 5.3 displays the descriptive statistics of the items in the competence subcategory. From Table 5.4, reliable item was strongly correlated with the capable item, however, those two prior items were only correlated with the interaction item moderately. As the average correlation was still high (around .59), we believed that

Table 5.4: *Pearson (r) correlations between items in competence component* ($N = 26$)

	Interactive	Reliable	Capable
Interactive	1		
Reliable	.46**	1	
Capable	.53**	.79***	1
	** p <.1; ** p <.05; *** p <.01*		

the three items were still related to each other and measured the same component.

Discomfort

Table 5.5: *Descriptive statistics on items in discomfort component*

	N	Mean	SD
Scary	26	2.04	1.61
Awkward	26	3.54	1.75
Aggressive	26	2.08	1.38

Table 5.6: *Pearson (r) correlations between items in discomfort component* ($N = 26$)

	Scary	Awkward	Aggressive
Scary	1		
Awkward	.55***	1	
Aggressive	.81***	.38*	1
	** p <.1; ** p <.05; *** p <.01*		

The descriptive statistics of the items in the competence subcategory is showed in Table 5.5. From Table 5.6, scary item was significantly correlated with the awkward item and the aggressive item. There existed a weak correlation between

the awkward and aggressive item. As the average correlation was still high (around .58), those three items could be treated as items in the same component.

5.5 Analysis and hypotheses

5.5.1 RoSAS survey and useful questionnaire

The warmth, competence, and discomfort categories were calculated by taking the average of the components in them. As the robot had a more friendly appearance than the avatar, we hypothesized that the robot would be perceived as warmer and less discomfort than the avatar. However, because facial expressions and verbal cues are easier to read than nonverbal cues from the body, we hypothesized that the avatar would be rated higher than the robot on competence category and useful category.

5.5.2 Ranking survey

We picked out the top five characteristics that the participants mentioned in each of the four categories: Most important (IM), wanted (WA), most undesired (UD), and unwanted (UW). We also created two weighted categories which were good characteristics (G) and bad characteristics (B). The formulas for them are below:

$$G = 2 * IM + WA$$

$$B = 2 * UD + UW$$

5.5.3 Online survey and transcript

The online survey would give us information for our discussion about people′s opinions on our feature for the AAC device. Due to the lack of time, we were unable to complete all the transcripts from the audio of the study and would rely mostly on observation notes from the moderator at the end of each focus group to give additional information in the discussion section.

5.6 Result

5.6.1 Impression toward robots: RoSAS and the Useful questionnaire

Table 5.7 and Figure 5.2 show a general statistical description of the components in our Impression toward Robots survey.

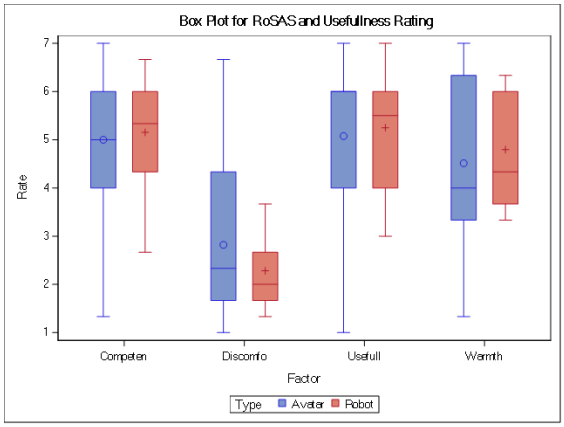


Figure 5.2: *Graph for the impression survey toward robot items*

From the paired t-test, there were not any differences in the warmth, com-

Table 5.7: *Descriptive statistics on items in impression toward the robot survey*

		N	Mean	SD
Robot	Warmth	13	4.79	1.16
	Competence	13	5.15	1.18
	Discomfort	13	2.28	0.77
	Useful	13	5.25	1.29
Avatar	Warmth	13	4.51	1.79
	Competence	13	5.00	1.59
	Discomfort	13	2.82	1.72
	Useful	13	5.08	1.89

Table 5.8: *Paired t-test on items in the impression toward robot survey between the robot and the avatar*

	Mean	SD	DF	t-value	p
Warmth (Robot - Avatar)	0.28	1.60	12	0.63	.538
Competence (Robot - Avatar)	0.15	1.55	12	0.36	.728
Discomfort (Robot - Avatar)	-0.58	1.93	12	-1.00	.335
Useful (Robot - Avatar)	0.33	2.02	12	0.57	.578

petence, discomfort, and useful rating between the robot and the avatar (Table 5.8).

5.6.2 Characteristics ranking

This section displayed the ranking results of the most important characteristics (Table 5.9), the wanted characteristics (Table 5.10), the weighted good characteristics (Table 5.11), the most undesired characteristics (Table 5.12), the unwanted characteristics (Table 5.13), and the weighted bad characteristics (Table 5.14).

Table 5.9: *Ranking for most important characteristics*

Rank	Characteristics	Frequency
1	Easy to use	7
2	Be able to talk	4
3	Easy to troubleshoot	3
4	Inexpensive	3
5	Durable	3

From Table 5.9, the group thought that it was important for the robot to be easy to use and trouble, be able to talk, durable, end inexpensive.

Table 5.10: *Ranking for wanted characteristics*

Rank	Characteristics	Frequency
1	Portable	7
2	Long battery time	5
3	Durable	4
4	Be able to hear command	4
5	Easy to use	3

The group wanted their robot to be portable and durable, have long battery time, be able to hear verbal commands and be easy to use (Table 5.10).

Table 5.11: *Ranking for weighted good characteristics*

Rank	Characteristics	Weighted Frequency
1	Easy to use	17
2	Portable	11
3	Durable	10
4	Be able to talk	10
5	Easy to trouble shoot	9

Overall, the participants highly prioritized the easy-to-use, portability, dura-

bility, ability-to-talk, and easy to troubleshoot (Table 5.11)

Table 5.12: *Ranking for most undesired characteristics*

Rank	Characteristics	Frequency
1	Break easily	12
2	Expensive	5
3	Manual control	2
4	Short battery time	1
5	Not have professional appearance	1

From Table 5.12, the group would not purchase the robot if it was easy to break, expensive, manually controlled, and have short battery time and unprofessional appearance.

Table 5.13: *Ranking for unwanted characteristics*

Rank	Characteristics	Frequency
1	Casual/Childish appearance	4
2	Hard exterior	3
3	Soft exterior	3
4	Expensive	2
5	Manual Control	2

The group did not want their robot to have child appearance, manual control, or high price tag. They had different opinions about the hard and soft exterior of the robot (Table 5.13).

Overall, the participants did not want their robot to break easily, be expensive, have a childish appearance, manual control, nor hard exterior (Table 5.14).

Table 5.14: *Ranking for weighted bad characteristics*

Rank	Characteristics	Weighted Frequency
1	Break easily	25
2	Expensive	12
3	Casual/childish appearance	6
4	Manual Control	6
5	Hard exterior	5

5.7 Discussion

5.7.1 Design of the robots

The data showed that the people in the group perceived the two robots similar to each other. There were two possible explanations for this phenomenon. The first explanation would be that our survey did not work as we intended. The RoSAS is designed and validated with only avatar robots and has not been validated with physical robots [52]. To test our theory, a post hoc analysis was done in Appendix E. Instead of looking at the correlation between items for both robot and avatar condition, this analysis calculated the correlation between the items for the robot condition and avatar condition separately. The data suggested that while the avatar condition was able to maintain a high correlation between the elements in each subcategory, there were almost no correlations between some items in those subcategories. This suggested that the RoSAS might perform differently with physical robots that did not have a facial display. Another explanation would be the prior positive feeling of the participants toward the prototypes. As the goal of the study was to improve their lives, the participants might look at the robots

more positive than normal. Table 5.7 showed that the participants rated the *good* quality (warmth and competence) above the average of the scale, and rated the *bad* quality (discomfort) below the average of the scale.

However, from our memos after each focus group, it was suggested that the participants who had trouble speaking seemed to prefer the avatar and its ability to talk while the other participants prefer the approachable appearance of the physical robot. A participant mentioned that she "would prefer to have the avatar now than wait for a perfect avatar". Those participants, who could not speak clearly, seemed to lose more control of their facial muscles than the other, and therefore, saw the ability to express facial emotions of the avatar more than just a mediator in their interaction. They perceived it as a way for them to express their emotions again. In contrast, most of the care-takers or other ALS patient could not overlook the uncanny appearance of the avatar. Finally, the group agreed that the nonverbal behaviors of the robot were harder to read than the facial cues of the avatar. We thought it could be because human is better at reading facial cues than body language. A further research on the behaviors of a physical robot would be needed before it could fully be applied in the actual settings.

For the characteristics of the robots, the group highly prioritized the durability and portability of the robot. Additionally, as the participants were in Oregon, they also mentioned that the robot should be weather-proof. They preferred the robot to be autonomous, and easy to use and troubleshoot. The participants had different preferences about the exterior of the robot, but they would prefer the robot to not have a childish appearance.

5.7.2 Feedback on the feature for the AAC device

For the phrase chunking program, all the participants said that they would use this feature. However, they preferred the ability to turn on and off this feature. The focus groups also commented on the accuracy of the phrases that spoke out.

5.7.3 Limitation and future work

One of our potential drawbacks was the possible inconsistency of our robots. We did not experience any technical issues in all of our focus group. However, because it was a live demo instead of a video clip, there was a possibility that the demo was different from each other. We originally planned to let the participant interact with the robots because we wanted them to have a full experience with the robots. Our study confirmed the importance of a live demo as one of the participants mentions that they like the "3-D" physical body of the Blossom Robot.

Our result suggests that future work that validates the RoSAS survey for physical robots would be needed. It is essential that the validation study should recruit a wider range of population than this study to eliminate any prior biases. Because we have not done a full analysis of the transcripts, an analysis that incorporates information from the transcripts would be needed. Finally, additional research and development on the robot behaviors would be needed to create better nonverbal cues for the robots.

Part IV

Summary and Future Work

Chapter 6 Summary

From the first study, we have learned that the conversation partners tend to look at the AAC user less frequently but get distracted longer when they do not feel interested in interactions that have short pauses. However, for interactions that are made up of long pauses, the inattentive time is the sole indicator for the lack of interest toward the conversations. Additionally, it seems that the AAC users are unaffected by the gaze behaviors of their partner. Interestingly, the second study suggests that the AAC users are often not aware of their partner′ behaviors while they are typing because they have to focus on the monitor. The few moments that the AAC users notice their partners is when they finished typing.

The second study suggests that the AAC users want a technology that can prevent bad behaviors and encourage good behaviors from the communication partners, to make the waiting time shorten and less awkward, and to signal the communicative partners before the speech synthesizer starts to speak.

Finally, it is suggested in our third study that most of the people who have trouble speaking prefer the avatar robot than the Blossom robot. As they start to lose their ability to control their facial muscle, they see the ability to express facial emotions of the avatar more than just a mediator in their interaction. They think it as a way for them to express their emotions again. In contrast, most of the care-takers or other ALS patient cannot overlook the uncanny appearance of

the avatar and prefer the Blossom robot which is more approachable.

Chapter 7 Future Work

For the first study, we need a future analysis on the correlations between gaze behaviors of the communication partners and how the AAC users feel toward the interaction when the speech synthesizer started to verbalize, and while the AAC users are typing. For the third study, an in-deed analysis on the transcript of the study is needed because the author was not able to analyze on completed transcripts, instead he used his notes at the end of each focus group for the analysis.

For developing better behaviors for the Blossom robot, a meta-analysis on robot behaviors is needed to understand the different behaviors have been implemented and its impact on a person. For the feature in the speech synthesizer, we need to try different state-of-the-art learning algorithm such as deep learning to improve our prediction. Additionally, to test the effect of our system with a communication partner, we would want to reproduce the first study to include the robot and speech synthesizer. Finally, the end goal of this research would be to have a completed system for the AAC users and examine its impact on their daily conversation in a long-term study.

Bibliography

[1] National Institute of Neurological Disorders and Stroke. Amyotrophic lateral sclerosis (als) fact sheet. 2018. URL `https://www.ninds.nih.gov/disorders/amyotrophiclateralsclerosis/detail_ALS.htm`.

[2] The ALS Association. Facts you should know. 2016. URL `http://www.alsa.org/about-als/facts-you-should-know.html`.

[3] Jayanti Ray. Real-life challenges in using augmentative and alternative communication by persons with amyotrophic lateral sclerosis. *Communication Disorders Quarterly*, 36(3):187–192, 2015.

[4] Assistive devices for people with hearing, voice, speech, or language disorders. *U.S. Department of Health and Human Services*, 2017. URL `https://www.nidcd.nih.gov/health/assistive-devices-people-hearing-voice-speech-or-language-disorders`.

[5] Tobii dynavox. *Tobii Dynavox*, 2019. URL `https://www.tobiidynavox.com/`.

[6] Laura J Ball, David R Beukelman, and Gary L Pattee. Acceptance of augmentative and alternative communication technology by persons with amyotrophic lateral sclerosis. *Augmentative and Alternative Communication*, 20 (2):113–122, 2004.

[7] John W Mullennix, Steven E Stern, Stephen J Wilson, and Corrie-lynn Dyson. Social perception of male and female computer synthesized speech. *Computers in Human Behavior*, 19(4):407–424, 2003.

[8] Steven E Stern. Computer-synthesized speech and perceptions of the social influence of disabled users. *Journal of Language and Social Psychology*, 27(3): 254–265, 2008.

[9] Steven E Stern, Muriel Dumont, John W Mullennix, and M Lynn Winters. Positive prejudice toward disabled persons using synthesized speech: Does the effect persist across contexts? *Journal of Language and Social Psychology*, 26 (4):363–380, 2007.

[10] Rupal Patel. Synthetic voices, as unique as fingerprints. 2013. URL `https://www.ted.com/talks/rupal_patel_synthetic_voices_as_unique_as_fingerprints/up-next?language=en`.

[11] Adam Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63, 1967.

[12] Mattias Heldner and Jens Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.

[13] Abigail J Sellen. Speech patterns in video-mediated conversations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 49–59. ACM, 1992.

[14] Ann Wennerstrom and Andrew F Siegel. Keeping the floor in multiparty conversations: Intonation, syntax, and pause. *Discourse Processes*, 36(2):77–107, 2003.

[15] Carmen Basil. Social interaction and learned helplessness in severely disabled children. *Augmentative and Alternative Communication*, 8(3):188–199, 1992.

[16] Janice Light, Barbara Collier, and Penny Parnes. Communicative interaction between young nonspeaking physically disabled children and their primary caregivers: Part ii—communicative function. *Augmentative and Alternative Communication*, 1(3):98–107, 1985.

[17] Shelley K Lund and Janice Light. Long-term outcomes for individuals who use augmentative and alternative communication: Part ii–communicative interaction. *Augmentative and Alternative Communication*, 23(1):1–15, 2007.

[18] Frank J Bernieri. The expression of rapport. *The sourcebook of nonverbal measures: Going beyond words*, 347:359, 2005.

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

[20] Judith H Langlois, Lisa Kalakanis, Adam J Rubenstein, Andrea Larson, Monica Hallam, and Monica Smoot. Maxims or myths of beauty? a meta-analytic and theoretical review. *Psychological bulletin*, 126(3):390, 2000.

[21] Frank Bernieri and Robert Rosenthal. Coordinated movement in human interaction. *Fundamentals of nonverbal behavior*, pages 401–431, 1991.

[22] E Bruce Goldstein. *Cognitive psychology: Connecting mind, research and everyday experience*. Nelson Education, 2014.

[23] Daniel Kahneman. *Attention and effort*, volume 1063. Citeseer, 1973.

[24] Carol L Colby, Jean-René Duhamel, and Michael E Goldberg. Oculocentric spatial representation in parietal cortex. *Cerebral cortex*, 5(5):470–481, 1995.

[25] Carrie J McAdams and R Clay Reid. Attention modulates the responses of simple cells in monkey primary visual cortex. *Journal of Neuroscience*, 25 (47):11023–11033, 2005.

[26] Daniel Kahneman, W Scott Peavler, and Linda Onuska. Effects of verbalization and incentive on the pupil response to mental activity. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 22(3):186, 1968.

[27] Chris L Kleinke. Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78, 1986.

[28] Jie Sui and Chang Hong Liu. Can beauty be ignored? effects of facial attractiveness on covert attention. *Psychonomic Bulletin & Review*, 16(2):276–281, 2009.

[29] Michael Argyle and Roger Ingham. Gaze, mutual gaze, and proximity. *Semiotica*, 6(1):32–49, 1972.

[30] Hee Rin Lee, Selma Šabanović, Wan-Ling Chang, Shinichi Nagata, Jennifer Piatt, Casey Bennett, and David Hakken. Steps toward participatory design of social robots: Mutual learning with older adults with depression. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 244–253. ACM, 2017.

[31] Alan W Black, Paul Taylor, Richard Caley, and Rob Clark. The festival speech synthesis system, 1999.

[32] John Kominek, Alan W Black, and Ver Ver. Cmu arctic databases for speech synthesis. 2003.

[33] R Rosenthal and RL Rosnow. Essential of behavioral research . ny, 2008.

[34] Joseph A Maxwell. *Qualitative research design: An interactive approach*, volume 41. Sage publications, 2012.

[35] Jennifer Kent-Walsh and Cathy Binger. Fundamentals of the impaact program. *Perspectives on Augmentative and Alternative Communication*, 22(1): 51–61, 2013.

[36] Ann Kaiser and Courtney Wright. Enhanced milieu teaching: Incorporating aac into naturalistic teaching with young children and their partners. *Perspectives on Augmentative and Alternative Communication*, 22(1):37–50, 2013.

[37] Marynel Vázquez, Aaron Steinfeld, and Scott E Hudson. Maintaining awareness of the focus of attention of a conversation: A robot-centric reinforcement learning approach. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, pages 36–43. IEEE, 2016.

[38] Marynel Vázquez, Elizabeth J Carter, Braden McDorman, Jodi Forlizzi, Aaron Steinfeld, and Scott E Hudson. Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 42–52. ACM, 2017.

[39] Guy Hoffman, Oren Zuckerman, Gilad Hirschberger, Michal Luria, and Tal Shani Sherman. Design and evaluation of a peripheral robotic conversation companion. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 3–10. ACM, 2015.

[40] Jaeeun Shim, Ronald Arkin, and Michael Pettinatti. An intervening ethical governor for a robot mediator in patient-caregiver relationship: Implementation and evaluation. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 2936–2942. IEEE, 2017.

[41] Michael J Pettinati and Ronald C Arkin. Towards a robot computational model to preserve dignity in stigmatizing patient-caregiver relationships. In *International Conference on Social Robotics*, pages 532–542. Springer, 2015.

[42] Nahum A Torres, Nathan Clark, Isura Ranatunga, and Dan Popa. Implementation of interactive arm playback behaviors of social robot zeno for autism spectrum disorder therapy. In *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments*, page 21. ACM, 2012.

[43] Michelle J Salvador, Sophia Silver, and Mohammad H Mahoor. An emotion recognition comparative study of autistic and typically-developing children using the zeno robot. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 6128–6133. IEEE, 2015.

[44] Evan Ackerman. Blossom: A handmade approach to social robotics from cornell and google. *IEEE Spectrum*, 2017.

[45] Gabriel Skantze. Iristk: Intelligent real-time interactive systems toolkit. 2018. URL `http://www.iristk.net/installation.html`.

[46] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics, 2002.

[47] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics, 2003.

[48] Taku Kudo and Yuji Matsumoto. Chunking with support vector machines. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

[49] Taku Kudo. Crf++: Yet another crf toolkit (2005). *Software available at http://crfpp. sourceforge. net*, 2014.

[50] Dag Sverre Syrdal, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. *Adaptive and Emergent Behaviour and Complex Systems*, 2009.

[51] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81, 2009.

[52] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. The robotic social attributes scale (rosas): Development and validation.

In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*, pages 254–262. ACM, 2017.

[53] Norm O'Rourke, R Psych, and Larry Hatcher. *A step-by-step approach to using SAS for factor analysis and structural equation modeling*. Sas Institute, 2013.

[54] Breck Baldwin and Bob Carpenter. Lingpipe. *Available from World Wide Web: http://alias-i. com/lingpipe*, 2003.

[55] Taku Kudoh and Yuji Matsumoto. Use of support vector learning for chunk identification. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 142–144. Association for Computational Linguistics, 2000.

[56] Jason Baldridge. The opennlp project. *URL: http://opennlp. apache. org/index. html,(accessed 2 February 2012)*, 2005.

[57] Erik F Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 127–132. Association for Computational Linguistics, 2000.

APPENDIX

Chapter A Surveys in First Study

DEMOGRAPHIC INFORMATION

1. How old are you? _____ years _____ months

2. Sex(Please circle one) FEMALE MALE OTHER

3. Is English your first language? Yes No

4. Please indicate which of the following race and ethnicity group best describe you

American Indian or Alaskan Native	Asian or Pacific Islander
African American	Hispanic
Caucasian/White	Other

5. How many years have lived in the United States? _____years

6. How much experience do you have with a computer keyboard?

   a. I've never used one before.

   b. I've used one a few times in my life.

   c. I use one every couple of months.

   d. I use one weekly.

   e. I use one daily.

7. How much experience do you have with X-box type controllers?

   a. I've never used one before.

b. I've used one a few times in my life.

c. I use one every couple of months.

d. I use one weekly.

e. I use one daily.

Interaction Assessment

This next section does not apply to you or your partner as individuals. Instead, we'd like to get your assessment of the conversational event. Please rate the interaction between you and your partner on the following characteristics. Circle the number that you think best describes the quality of the interaction.

NOT AT ALL								EXTREMELY	
0	1	2	3	4	5	6	7	8	WELL-COORDINATED
0	1	2	3	4	5	6	7	8	BORING
0	1	2	3	4	5	6	7	8	COOPERATIVE
0	1	2	3	4	5	6	7	8	HARMONIOUS
0	1	2	3	4	5	6	7	8	UNSATISFYING
0	1	2	3	4	5	6	7	8	UNCOMFORTABLY PACED
0	1	2	3	4	5	6	7	8	COLD
0	1	2	3	4	5	6	7	8	AWKWARD
0	1	2	3	4	5	6	7	8	ENGROSSING
0	1	2	3	4	5	6	7	8	UNFOCUSED
0	1	2	3	4	5	6	7	8	INVOLVING
0	1	2	3	4	5	6	7	8	INTENSE
0	1	2	3	4	5	6	7	8	UNFRIENDLY
0	1	2	3	4	5	6	7	8	ACTIVE
0	1	2	3	4	5	6	7	8	POSITIVE
0	1	2	3	4	5	6	7	8	DULL
0	1	2	3	4	5	6	7	8	WORTHWHILE
0	1	2	3	4	5	6	7	8	SLOW

FINAL QUESTIONS

1. How well did you know the person you interacted with today before showing up for today's study? (Circle 1)

1. Never met them before today.
2. I've seen them but we've never talked.
3. We've talked but I don't know them well.
4. We are well acquainted.
5. We are friends and know each other well.

Chapter B Surveys in Second Study

DEMOGRAPHIC INFORMATION

1. Are you a caregiver, a patient with ALS, or other?

CAREGIVER	PATIENT WITH ALS	OTHER

2. How old are you?

YEARS	
.	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9

3. Sex

FEMALE	MALE	OTHER

4. Are you fluent in English?

YES	NO

5. If you are a patient with ALS, how long ago were you diagnosed?

YEARS		MONTHS
	0	0
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6
7	7	7
8	8	8
9	9	9
		10
		11
		12

6. Have you had experiences with speech synthesizer devices (AAC device, text-to-speech device, Tobii device, etc) from either interacting with or using it?

YES	NO

7. In your consent form, you have consented to be audio recorded throughout the focus group. Would you be willing to be video recorded as well in order to aide in the collection of data? All audio and video recording will remain confidential and only used for the purpose of this study.

YES	NO

Chapter C Surveys in Third Study

DEMOGRAPHIC INFORMATION

1. Are you a caregiver, a patient with ALS, or other?

CAREGIVER	PATIENT WITH ALS	OTHER

2. How old are you?

YEARS	
.	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9

3. Sex

FEMALE	MALE	OTHER

4. Are you fluent in English?

YES	NO

5. If you are a patient with ALS, how long ago were you diagnosed?

YEARS		MONTHS
	0	0
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6
7	7	7
8	8	8
9	9	9
		10
		11
		12

6. Have you had experiences with speech synthesizer devices (AAC device, text-to-speech device, Tobii device, etc) from either interacting with or using it?

YES	NO

7. In your consent form, you have consented to be audio recorded throughout the focus group. Would you be willing to be video recorded as well in order to aide in the collection of data? All audio and video recording will remain confidential and only used for the purpose of this study.

YES	NO

8. Have you been to the previous support group?

YES	NO

IMPRESSION TOWARD ROBOT

Using the scale provided, how closely are the words below associated with the robot?

1 = definitely not associated to 7 = definitely associated

a. Emotion

1	2	3	4	5	6	7

b. Social

1	2	3	4	5	6	7

c. Compassionate

1	2	3	4	5	6	7

d. Interactive

1	2	3	4	5	6	7

e. Reliable

1	2	3	4	5	6	7

f. Capable

1	2	3	4	5	6	7

g. Scary

1	2	3	4	5	6	7

h. Awkward

1	2	3	4	5	6	7

i. Aggressive

1	2	3	4	5	6	7

j. Useful

1	2	3	4	5	6	7

IMPRESSION TOWARD AVATAR

Using the scale provided, how closely are the words below associated with the avatar?

1 = definitely not associated to 7 = definitely associated

a. Emotion

1	2	3	4	5	6	7

b. Social

1	2	3	4	5	6	7

c. Compassionate

1	2	3	4	5	6	7

d. Interactive

1	2	3	4	5	6	7

e. Reliable

1	2	3	4	5	6	7

f. Capable

1	2	3	4	5	6	7

g. Scary

1	2	3	4	5	6	7

h. Awkward

1	2	3	4	5	6	7

i. Aggressive

1	2	3	4	5	6	7

j. Useful

1	2	3	4	5	6	7

RANKING CHARACTERISTICS OF ROBOTS

Please read the following characteristics of the robots

A. Easy to troubleshoot	B. Be able to talk
C. Animal-like appearance	D. Easy to use
E. Humanoid appearance	F. The robot is autonomous
G. Price-how expensive the robot is	H. The robot is manual control
I. Professional appearance	J. Having a physical body
K. Casual/childish appearance	L. Portable
M. Soft exterior (yarn, wool)	N. Durable
O. Hard exterior (plastic, wood, metal)	P. Friendly appearance
Q. Break easily	R. Having an avatar display
S. Battery time	T. Be able to hear command

Please select the 2 most important characteristics in your opinion (You will not buy the robot if it does not have it)

Please select 3 important characteristics (it would be nice to have it)

Please select the 2 most undesired characteristics in your opinion (You will not use the robot if it have it)

Please select 3 undesired characteristics (it would be nice to not having it)

Extra comment on next page

If you have one other characteristic that is most important to use, please list them below and explain why (optional)

If you have one other characteristic that is most undesired for you, please list them below and explain why (optional)

ONLINE SURVEY

Do you want the speech synthesizer autonomously speak for you?

- Yes
- No

Why do you choose the answer above (optional)?

Do you think that you will use this function on your speech synthesizer?

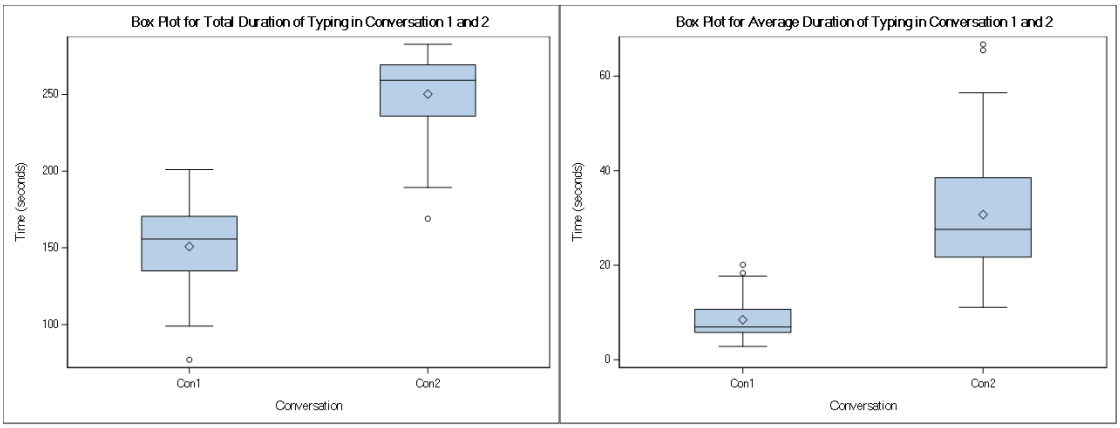
- Yes
- No

Why do you choose the answer above (optional)?

Any additional comments?

Chapter D Analysis for Normal Time in Study 1

D.1 Typing



(a) Total duration of typing (b) Average time of typing

Figure D.1: *Normal time graph for typing behavior*

Table D.1: *Descriptive statistics on typing behavior in the whole interaction (normal time)*

Interaction	Duration	N	Mean	SD
1	Total Duration	59	150.9	26.78
	Average Duration	59	8.44	3.94
2	Total Duration	59	250.2	24.52
	Average Duration	59	30.71	13.25

Table D.2: *Paired t-test on typing between the interaction (normal time)*

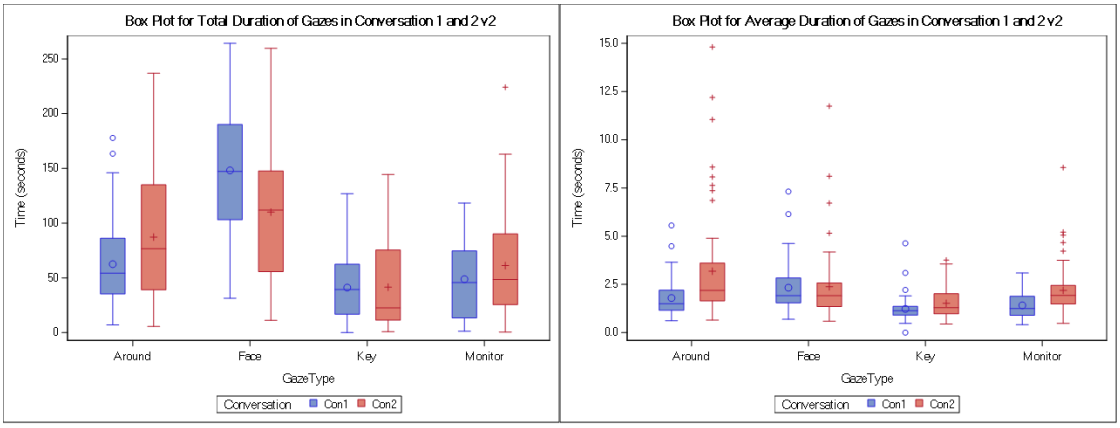
	Mean	SD	DF	t-value	p
Total Duration (Dyad 2 - Dyad 1)	99.38	28.21	58	27.05	<.001
Average Duration (Dyad 2 - Dyad 1)	22.26	11.35	58	15.06	<.001

Table D.3: *Pearson (r) correlation on typing between the interaction (normal time)*

	N	r	p
Total Duration (Dyad 1 and Dyad 2)	59	0.40	=.001
Average Duration (Dyad 1 and Dyad 2)	59	0.60	<.001

D.2 Gaze Behavior

D.2.1 The whole dyad



(a) Total Duration of Typing (b) Average Time of Typing

Figure D.2: *Normal time graph for gaze behaviors in the whole 5 minutes interaction*

Table D.4: *Descriptive statistics on gaze behavior in the whole interaction (normal time)*

Gaze Type	Interaction	Duration	N	Mean	SD
Face gaze	1	Total Duration	59	148.20	59.28
		Average Duration	59	2.33	1.25
	2	Total Duration	59	110.00	60.25
		Average Duration	59	2.38	1.85
Around gaze	1	Total Duration	59	62.55	37.81
		Average Duration	59	1.79	0.89
	2	Total Duration	59	87.27	60.36
		Average Duration	59	3.19	2.87

Face Gaze

Table D.5: *Paired t-test on face gaze behaviors between the interaction (normal time)*

	Mean	SD	DF	t-value	p
Total Duration (Dyad 2 - Dyad 1)	-38.14	39.46	58	-7.42	<.001
Average Duration (Dyad 2 - Dyad 1)	0.05	1.02	58	0.37	.710

Table D.6: *Pearson (r) correlation on face gaze behaviors between the interaction (normal time)*

	N	r	p
Total Duration (Dyad 1 and Dyad 2)	59	0.78	<.001
Average Duration (Dyad 1 and Dyad 2)	59	0.85	<.001

Inattentive Gaze

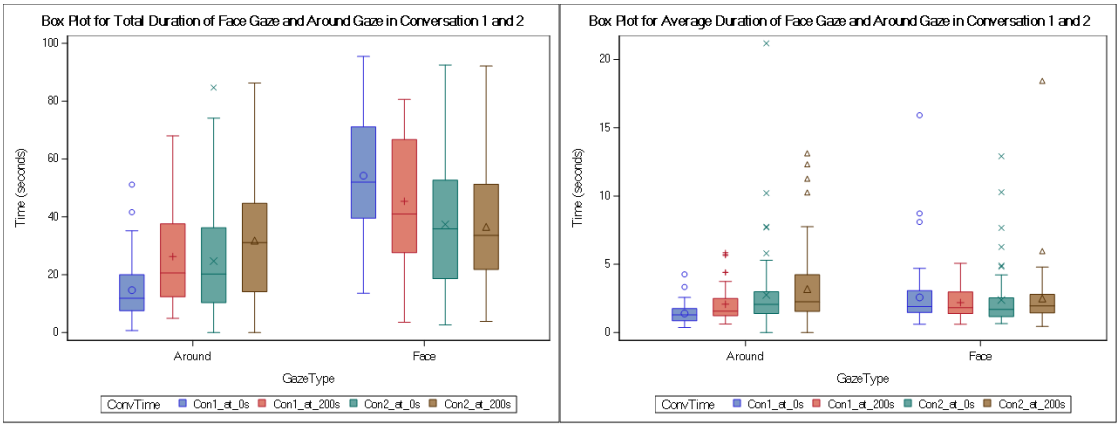
Table D.7: *Paired t-test on inattentive gaze behaviors between the interaction (normal time)*

	Mean	SD	DF	t-value	p
Total Duration (Dyad 2 - Dyad 1)	24.72	43.75	58	4.34	<.001
Average Duration (Dyad 2 - Dyad 1)	1.39	2.31	58	4.62	<.001

Table D.8: *Pearson (r) correlation on inattentive gaze behaviors between the interaction (normal time)*

	N	r	p
Total Duration (Dyad 1 and Dyad 2)	59	0.69	<.001
Average Duration (Dyad 1 and Dyad 2)	59	0.72	<.001

D.2.2 Within the dyad



(a) Total Duration of Typing (b) Average Time of Typing

Figure D.3: *Normal time graph for gaze behaviors at the beginning 100s and at the end 100s of the interaction*

Table D.9: *Descriptive statistics on gaze behavior within the first interaction and second interaction (normal time)*

Int.	Gaze Type	Moment	Duration	N	Mean	SD
1	Face gaze	First 100s	Total Duration	59	54.24	19.61
			Average Duration	59	2.58	2.29
		Last 100s	Total Duration	59	45.40	21.52
			Average Duration	59	2.20	1.09
	Around gaze	First 100s	Total Duration	59	14.68	10.70
			Average Duration	59	1.40	0.70
		Last 100s	Total Duration	59	26.28	16.58
			Average Duration	59	2.08	1.23
2	Face gaze	First 100s	Total Duration	59	37.43	21.53
			Average Duration	59	2.39	2.22
		Last 100s	Total Duration	59	36.56	20.82
			Average Duration	59	2.50	2.39
	Around gaze	First 100s	Total Duration	59	24.72	19.99
			Average Duration	59	2.76	3.10
		Last 100s	Total Duration	59	31.83	20.88
			Average Duration	59	3.19	2.79

Face Gaze

Table D.10: *Paired t-test on face gaze behaviors within the first interaction and second interaction (normal time)*

Int.	Duration	Mean	SD	DF	t-value	p
1	Total (Last 100s - First 100s)	-8.84	11.32	58	-6.00	<.001
	Average (Last 100s - First 100s)	-0.38	1.95	58	-1.50	.140
2	Total (Last 100s - First 100s)	-0.87	10.78	58	-0.62	.540
	Average (Last 100s - First 100s)	0.11	1.97	58	0.44	.659

Table D.11: *Pearson (r) correlation on face gaze behaviors within the first interaction and second interaction (normal time)*

Int.	Duration	N	r	p
1	Total (Last 100s and First 100s)	59	0.85	<.001
	Average (Last 100s and First 100s)	59	0.53	<.001
2	Total (Last 100s and First 100s)	59	0.87	<.001
	Average (Last 100s and First 100s)	59	0.64	<.001

Inattentive Gaze

Table D.12: *Paired t-test on inattentive gaze behaviors within the first interaction and second interaction (normal time)*

Int.	Duration	Mean	SD	DF	t-value	p
1	Total (Last 100s - First 100s)	11.60	11.19	58	7.96	<.001
	Average (Last 100s - First 100s)	0.68	0.92	58	5.65	<.001
2	Total (Last 100s - First 100s)	7.11	12.35	58	4.42	<.001
	Average (Last 100s - First 100s)	0.43	2.47	58	1.34	.186

Table D.13: *Pearson (r) correlation on inattentive gaze behaviors within the first interaction and second interaction (normal time)*

Int.	Duration	N	r	p
1	Total (Last 100s and First 100s)	59	0.74	<.001
	Average (Last 100s and First 100s)	59	0.67	<.001
2	Total (Last 100s and First 100s)	59	0.82	<.001
	Average (Last 100s and First 100s)	59	0.65	<.001

Chapter E Post Analysis for Study 3

This is the post analysis between items for each component in the RoSAS for the robot condition and the avatar condition.

Items in Warmth Components

Table E.1: *Pearson (r) correlations between items in warmth component for both robot and avatar* ($N = 13$)

Robot				Avatar			
	Emo.	Soc.	Com.		Emo.	Soc.	Com.
Emotion	1			Emotion	1		
Social	.49*	1		Social	.85***	1	
Compassion	.41	.40	1	Compassion	.67**	.66**	1
				** p <.1; ** p <.05; *** p <.01*			

Items in Competence Components

Table E.2: *Pearson (r) correlations between items in competence component for both robot and avatar* ($N = 13$)

Robot				Avatar			
	Int.	Rel.	Cap.		Int.	Rel.	Cap.
Interactive	1			Interactive	1		
Reliable	.20	1		Reliable	.70***	1	
Capable	.14	.76***	1	Capable	.80***	.83***	1
				** p <.1; ** p <.05; *** p <.01*			

Items in Discomfort Components

.

Table E.3: *Pearson (r) correlations between items in discomfort component for both robot and avatar* ($N = 13$)

Robot				Avatar			
	Sca.	Awk.	Agg.		Sca.	Awk.	Agg.
Scary	1			Scary	1		
Awkward	.26	1		Awkward	.65**	1	
Aggressive	.70***	.19	1	Aggressive	.81***	.44	1
				** p <.1; ** p <.05; *** p <.01*			

Chapter F Design for the Phrase Chunking Algorithm

In this project, we investigate how well previous chunking approaches to complete sentence chunking translate to the incomplete sentence domain. Following the pattern set by Sha and Pereira[47], we consider 1.) a HMM implementation to represent classical, generative probabilistic models, 2.) a Support Vector Machine (SVM) implementation to represent the best of classical, discriminative classifiers, and 3.) a CRF implementation to represent more recent approaches as proposed by Sha and Pereira. While chunking prediction needs to be performed in real-time for our purposes, we note that the models can be pre-trained. While model training often takes a considerable amount of time for the approaches we consider, prediction is performed significantly faster, and the prediction time for a single example as found in our context is negligible. Additionally, we note that the chunking task often involves a pre-processing step of tagging the sentence tokens with part of speech (POS) markers. Since POS tagging is an easier task with many robust existing implementations, we assume POS tagging has already been performed in our experiments (in fact, our training data includes the tokens' POS tags).

This document is structured as follows. Section F.1 presents our general framework and contains background information on the chunking models we consider. In Section F.2, we briefly describe the dataset we use and discuss our experimental approach. Results and discussion is presented in Section F.3. Finally, we end with some conclusions and considerations for future work in Section F.4.

F.1 Methodology

In this section, we present our approach for investigating the efficacy of popular chunking models in the incomplete sentence domain. Specifically, we test whether chunking performance on incomplete sentences improves when the proposed models are trained on incomplete rather than complete sentences. We also compare the performance from the different models to see which performs best for incomplete sentences. For our models, we test the classical approaches of using a generative probabilistic model or a discriminative classifier as well as a more modern approach of using a CRF. To represent classical approaches, we use a HMM chunker implementation found in the LingPipe toolkit[54] and a SVM chunker implementation presented by Kudo and Matsumoto[55]. For the more modern approach, we use a CRF chunker implementation found in the OpenNLP toolkit[56].

In the following subsections, we first propose our mathematical formalization for measuring chunking performance on incomplete sentences. Subsequently, we describe how the HMM, SVM, and CRF models are adapted to the chunking domain. The details of our experimental evaluation are presented in Section 3.

F.1.1 Formalization

In this subsection, we formalize the learning model for our chunking task. Generally, we try to maximize the following likelihood function:

$$\arg\max_{ChunkingSequence} P(ChunkingSequence|ObservedSentence) \tag{F.1}$$

However, this objective must be adapted to fit the model for each approach that we investigate.

Classical classification algorithms can only output a single classification label. Therefore, direct inference and sequential prediction is difficult. We modify the objective as shown below:

$$\arg\max_y \prod_{y=1}^{T} P(y_i|\phi(\mathbf{x})) \tag{F.2}$$

Note that the best prediction sequence is not equivalent to the product of best predicted labels, y_i. Such an approach is limited as compared to a structured prediction approach.

Since HMM's support exact inference of the joint likelihood, we use the joint likelihood as our objective. The joint likelihood is proportional to the conditional likelihood as shown in equation F.3.

$$\arg\max_{ChunkingSequence} P(ChunkingSequence, ObservedSentence) \tag{F.3}$$

Due to increased complexity in the model, CRF's on the otherhand are able to maximize the likelihood function found in equation F.1 directly.

F.1.2 SVM

We use a classical Support Vector Machine (SVM) model adapted for the chunking context, as proposed by Kudoh and Matsumoto[55]. The model was introduced for the CoNLL 2000 Shared Task on Chunking[57].

Generally, the goal of a SVM is to maximize the margin of support vectors close to the decision boundary. The input format for a SVM is a vector and the output is

a single label. In this implementation, chunk classification for each token is performed sequentially, and the input features for each classification include local word tokens, POS tags, and previous classifications provided by the model. The input vector is defined as follows:

$$\mathbf{x} = \begin{bmatrix} x_1 : \text{token:} -2 \\ x_2 : \text{token:} -1 \\ \vdots \\ x_5 : \text{token:} +2 \\ x_6 : \text{pos:} -2 \\ x_7 : \text{pos:} -1 \\ \vdots \\ x_{10} : \text{pos:} +2 \\ x_{11} : \text{chunk:} +1 \end{bmatrix} \tag{F.4}$$

The SVM implementation provided by the authors is an open-source package named Yet Another Multipurpose CHunk Annotator (YamCha). We note that this implementation can be trained with any polynomial kernel and can use either pairwise comparison or a one verses all approach for its multi-class classification. This model also performed best in the CoNLL 2000 Shared Task.

F.1.3 HMM

One argument for using a structured model is that the most likely sequence of predictions is not equivalent to the best set of predictions for each latent variable. Classical approaches ignore the correlation between latent variables.

A Hidden Markov Model (HMM) is a generative structured graphic model, which, similarly to Naive Bayes, predicts the joint probability of observations and a latent sequence. One serious limitation of this model is its strong assumption of conditional independence amongst the observations, which is often not true.

Despite this, HMM model is often used to perform sequential prediction because it is easy to implement. Training a HMM is straightforward using the Expectation Maximization algorithm, and model inference is similarly easy using the Viterbi algorithm, or more generally the max-product algorithm.

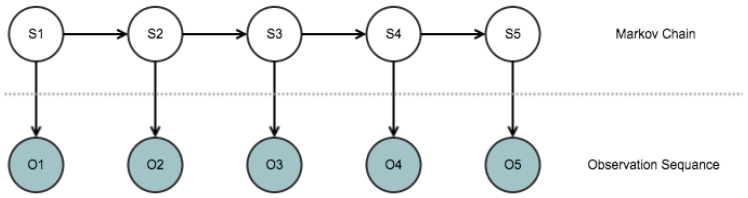


Figure F.1: *HMM model*

The joint likelihood of a HMM can be expressed as follows:

$$P(O,S) = P(S_1)\prod_{i=2}^{T} P(S_t|S_{t-1})\prod_{j=1}^{T} P(O_t|S_t)$$

The HMM chunking task is described with two line variables as shown in Figure F.1. The observations are POS tags and the latent variables are chunking tags. Here, in Figure F.1, to simplify, each latent variable has only one observation. However, it is possible to have multiple observations for single latent variable.

Training this model is normally achieved by using the EM algorithm. First we need to guess an initial parameter sequence that determines the transition probability and

fix it to maximize latent variables. Then, we fix learned latent variables to optimize parameters.

The flexibility of the HMM model is strictly limited by Markov assumptions, which means it is unable to take more than one observation into consideration at each time step. It is thus more of a baseline algorithm for our work.

For our analysis, we use a HMM chunking implementation provided in the LingPipe Java Natural Language Processing package[54] .

F.1.4 CRF

In the Natural Language Processing field, Conditional Random Fields (CRF) have become more common in tackling POS tagging and chunking problems. The relation between a CRF and HMM is similar to that of Logistic Regression and a Naive Bayes Classifier.

A CRF is a deterministic model which tries to maximize the conditional probability of $P(Y|X)$ directly rather than assuming the probability of observations.

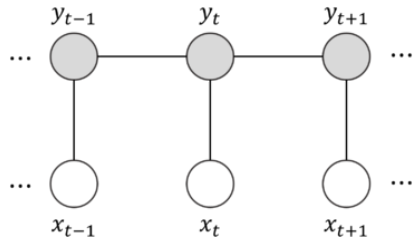


Figure F.2: *CRF model*

Factorization of an undirected graphical model is more focused on compatibility among variables. Therefore, the complexity of the CRF model depends on the number

of variables considered.

$$V_{\mathrm{x}} = \frac{1}{Z} \prod_{t=1}^{T} \exp\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\} \tag{F.5}$$

Computing the partition function of the CRF is difficult, which makes using the undirected graphical model to compute the joint likelihood intractable. However, computing the conditional probability $P(Y|X)$ will directly eliminate the partition function.

$$\begin{aligned} P(\mathbf{y}|\mathbf{x}) &= \frac{P(\mathbf{y}, \mathbf{x})}{\sum_{y'} P(\mathbf{y}', \mathbf{x})} \\ &= \frac{\prod_{t=1}^{T} \exp\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t)\}}{\sum_{y'} \prod_{t=1}^{T} \exp\{\sum_{k=1}^{K} \theta_k f_k(y'_t, y'_{t-1}, x_t)\}} \end{aligned} \tag{F.6}$$

In this project, we use a Java implementation of a CRF chunker provided in the OpenNLP toolset[56]. The package provides code for training and testing a CRF chunker.

F.2 Experiment

F.2.1 Data

We decide to use data provided by the CoNLL 2000 Shared Task on Chunking[57]. This task provides a standard metric for assessing the performance of chunkers on complete sentences. This data is furthermore useful since the YamCha and OpenNLP implementations we use are constructed to accept data following its format. The task provides a corpus of 211,727 token-POS-chunk instances used for training and a corpus of 47,377 instances for testing.

The structure for CoNLL 2000 data is shown in F.1. The data is contained in a single

Table F.1: *Data sample*

Confidence	NN	B-NP
in	IN	B-PP
the	DT	B-NP
pound	NN	I-NP
is	VBZ	B-VP
widely	RB	I-VP
expected	VBN	I-VP
to	TO	I-VP
take	VB	I-VP
another	DT	B-NP
sharp	JJ	I-NP

file and includes three columns: word token, POS tag, and chunk tag. Chunk tags begin with either a "B" or "I", specifying whether the given token begins a new chunk or continues an existing chunk. The prefix is followed by the chunk identifier. A separate "O" tag is used to identify punctuation. For our purposes, we say that a word is chunked correctly if it is correctly assigned the correct tag (prefix and chunk identifier) as found in the test set.

We use the CoNLL training and test sets as bases for generating datasets with incomplete sentences. Specifically, we bootstrap examples from the original corpus and randomly select a word in the sentence to split on. The first half of the sentence is then treated as an incomplete sentence example. We take this bootstrapping approach to randomly generate incomplete sentences from the original data. In order to construct similar training sets for complete and incomplete sentences, for every bootstrapped example, we save the unaltered version of the example in a complete sentence training set and the randomly split version in an incomplete sentence training set. We construct a testing set of incomplete sentences in a similar manner. We choose to bootstrap a number of

examples equal to a chosen multiplier multiplied by the size of the original data set.

F.2.2 Approach

To test the performance of our proposed models on chunking incomplete sentences, we first generate two separate training sets containing complete or incomplete sentences as described in the previous subsection. For each of the proposed models, we train a separate version on each of the training sets. To assess the robustness of each model to the size of the training set, we use a variety of bootstrap multipliers for training.

The test set was generated by bootstrapping the CoNLL 2000 Shared Task test data with a bootstrap multiplier of 5, splitting each example at a random index to produce incomplete sentences. This produced a test set of 139,084 token-POS-chunk instances in a similar format to that of the CoNLL 2000 shared task, but with incomplete rather than complete sentences. We chose a bootstrap multiplier of 5 since it was high enough such that we are expected to sample almost all (99%) of the examples at least once.

F.2.3 Parameter Tuning

While the HMM and CRF chunker tools did not have tunable parameters, the SVM implementation had the option of specifying the polynomial order of the kernel, C value for slack weighting, and multi-class classification strategy (pairwise or one versus rest). To perform this tuning, we split off 80% of the original CoNLL training data as training data and the other 20% as validation data for testing the effects of parameter variation. From each set, we then bootstrapped training examples using a bootstrap multiplier of 5. After determining the best parameters by comparing performance on the bootstrapped

validation set, we retrained the SVM model using the optimal parameters on data bootstrapped from the whole CoNLL training dataset for our model comparison evaluation. We note that we do not perform cross-validation due to time constraints and the large time cost of training the SVM models.

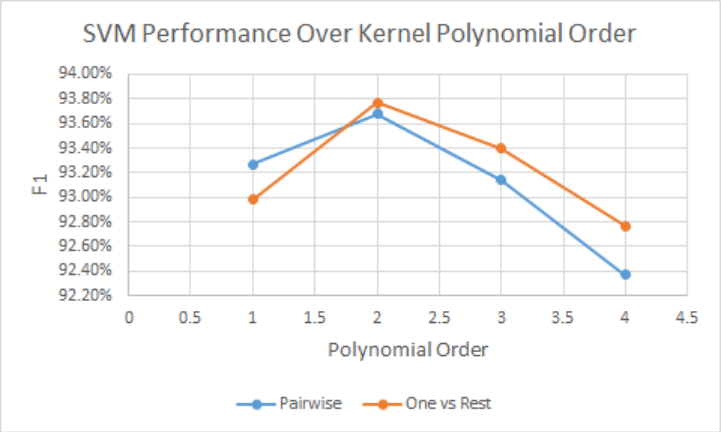


Figure F.3: *F_1 performance of SVM when trained with a polynomial kernel of various orders. Performance using both pairwise and one verses rest multi-class classification strategy is shown.*

Figures F.3 and F.4 show the results of performing parameter tuning on the SVM model. Due to the large training time overhead, we used a greedy approach by first tuning the polynomial order and then tuning the C value using the optimal polynomial parameter value. We found that using a polynomial order of 2 and a C value of 0.1 was optimal for both multi-class classification strategies. Furthermore, we chose to use the one versus rest approach since it performed optimally in the best case.

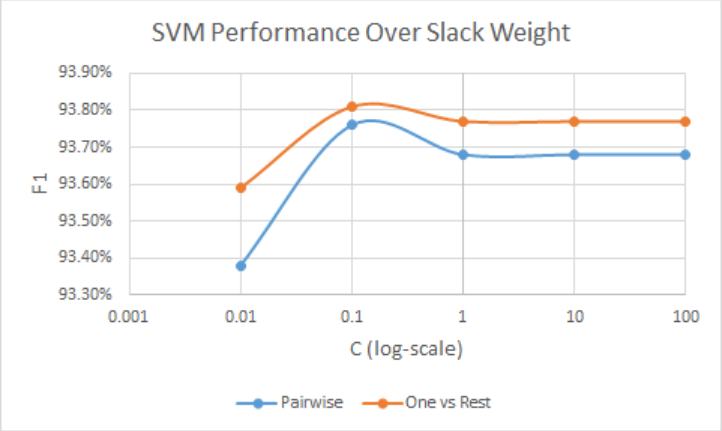


Figure F.4: *F_1 performance of SVM when trained with different C values. Performance using both pairwise and one verses rest multi-class classification strategy is shown.*

F.3 Results and Analysis

Figure F.5 shows the test performance of each of the proposed models after training on either complete or incomplete sentences. To evaluate the robustness of each model to the size of the training data, we vary the bootstrap multiplier used to generate the training data. We choose to represent performance using the F_1 score, a standard metric in natural language processing for combining the information provided by precision and recall. For each model, we see that the version trained on incomplete sentences performs better than the version trained on complete sentences in almost all cases. This breaks down slightly for the HMM model, where the version trained on complete sentences performs better for smaller amounts of data. Still, on average, we see a 0.34% point increase in performance for the SVM, 3.86% point increase for the HMM, and 0.81% point increase for the CRF when the model is trained on incomplete rather than complete sentences. This suggests that training a chunking model on incomplete sentences will

lead to increased performance in chunking incomplete sentences over a model trained on complete sentences.

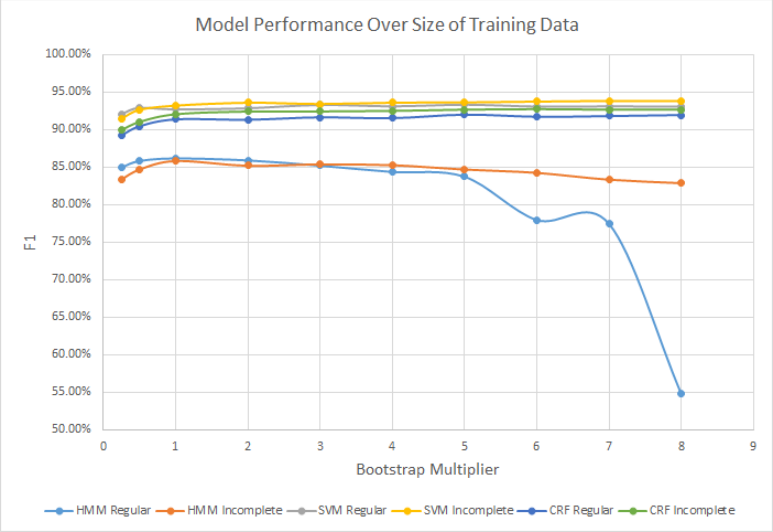


Figure F.5: *F_1 performance on incomplete sentences for HMM, SVM, and CRF models. A version of each model was trained on complete sentences and a separate version was trained on incomplete sentences.*

We also see that both versions of the SVM model perform better than the CRF models, which in turn perform significantly better than the HMM models. While it is surprising that the SVM model outperforms the CRF model, we note that this particular implementation of a SVM chunker was developed specifically for the CoNLL 2000 shared task (it in fact won the competition), and thus was probably tuned specifically for this data set. We also note Sha and Perier mention in their paper that their CRF chunker did not outperform the YamCha implementation on complete sentences[47]. In any case, we note that both the SVM and CRF significantly outperform the HMM models, suggesting that these more advanced models are preferred for the chunking task. We also note that HMM performance (especially when trained on complete sentences) drops off with

too much data. This might be because of the fact that the HMM assumes conditional independence between the features, which breaks down with too much data (particularly since all examples are generated from the same original data set).

F.4 Conclusions

In this project, we examined how three different approaches (namely SVM, HMM, and CRF) of complete sentence chunking translated to the incomplete sentence domain. The experimental results suggest that training on incomplete rather than complete sentences leads to improved performance for all models. We also find that the CRF and SVM implementations have comparable performance, with the SVM chunker performing slightly better. Both models, however, significantly outperform the HMM chunker.

We do, however, note a number of limitations in our approach. First, the SVM tool that we used was tailored specifically for our data set and, due to time constraints and the large time cost of model training, we used only that data set for our experiments. This might have led to unfair bias in favor of the SVM model. In any case, testing on multiple datasets or even generating multiple training sets from the source text would lead to an increase in the validity of our results. Second, for each method, we only used one implementation in our experiments. There might be other SVM, HMM, or CRF tools that outperform our models, but we only considered available open-source implementations. Additionally, we note that advanced structured prediction models such as a structured SVM (which had been used to great effect for tasks like parsing) were not considered in our experiments. Adapting such advanced models could lead to improved performance in the incomplete chunking domain.