Intonation in Robot Speech
Does it work the same as with people?

SUBMITTED IN PARTIAL FULLFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE

P.C. (Ella) Velner
11360666

MASTER INFORMATION STUDIES
Information Systems

25 June 2019

1st Examiner
Prof. dr. P.P.G. (Paul) Boersma
Phonetic Sciences, UvA

2nd Examiner
Dr. M.M.A. (Maartje) de Graaf
Information and Computing Sciences, UU

3rd Examiner
Dr. F.M. (Frank) Nack
Informatics Institute, UvA

Intonation in Robot Speech
Does it work the same as with people?

P.C. Velner
ellavelner@gmail.com
11360666
Universiteit van Amsterdam
Amsterdam

P.P.G. Boersma
p.p.g.boersma@uva.nl
Supervisor
Universiteit van Amsterdam
Amsterdam

M.M.A. de Graaf
m.m.a.degraaf@uu.nl
Supervisor
Universiteit Utrecht
Utrecht

ABSTRACT

In HRI the aim is often to make it as natural as possible. Researchers see how the interaction is done in the human world and apply that to the HRI. However, it is not studied if this is the best aim, therefore it is useful to look into each aspect of the HRI. One of these aspects is the intonation in robot speech and how this affects the interaction. In an experimental study, 120 participants were asked to have a conversation with a robot (NAO) in one of four conditions: (1) no intonation, (2) focus intonation, (3) end-of-utterance intonation, (4) full intonation. After the conversation they also filled in a questionnaire on naturalness, eeriness and sociability. The conversations were recorded and objective data on naturalness and sociability was extracted from these. Results showed that human-robot interaction did not work the same as human-human interaction, since the regular turn-taking system did not apply and the types of intonation did not have the same effect as in conversations between people. This was mainly due to inappropriately long silences between the turns of the robot and the participant. Future research needs to be done to support these claims and to apply it to other languages.

KEYWORDS

Human-Robot Interaction, HRI, Intonation, Robot Speech, Turn Organization, Naturalness, Emotion, Eeriness, Uncanny, Linguistics, Conversation Analysis

Contents

1 INTRODUCTION

Robots are coming into our everyday lives more and more. Whether it is a robotic vacuum cleaner or a social robot dog, people are getting used to having them around. The common thing with these robots is that they are all based on something we already know, like a vacuum cleaner or a pet. This makes it easier to interact with these otherwise strange and complex things [7]. Researchers are constantly investigating how to make a robot seem more natural, in looks [36], behaviour [6, 8, 46], and speech [21, 29, 36]. They look into how conversing is done in our world and apply that to the robots.

However, this can lead to some problems. Robots can be natural and human-like in one way, but lack this humanness in other ways, which results in an eerie and uncanny outcome [35]. A second problem is that humans know when they are dealing with robots most of the time, and they have certain thoughts and expectations when interacting with them [14]. This affects the way they interact with the robot. For example, with chatbots, people tend to use simpler language [24]. Another issue is that, while we try to make robots more human-like, robots are not the same as humans. Therefore, some mechanisms and theories that apply to the human world might not apply when robots are involved. So research is necessary to investigate if interaction between humans works the same way as interaction between humans and robots.

One of the key features of the robot in human-robot interaction (HRI) is speech. Robot engineers are constantly working on making robot speech as human-like as possible, so that talking with robots feels natural to the human [12, 21, 27, 36]. But speech has a lot of different aspects to take into account. When speaking, humans make decisions on six levels, often at the same time: phonetically, phonologically, morphologically, syntactically, semantically, and pragmatically [31]. And these decisions all differ for each language and culture [9, 25]. To implement all these decisions in a robot is a lot of work, and ideally, every aspect should be researched separately and tested to be useful. As a follow-up it should furthermore be studied if these aspects influence each other.

One aspect within speech that has been thoroughly researched within human interaction, but not so much in HRI, is intonation [1]. This is the pattern of pitch within utterances [31]. Intonation can alter the meaning of a sentence, give stress and therefore more importance to a certain aspect within the utterance, or prepare the listener for turn-taking, by implicitly letting them know they are done speaking [9, 22]. But do listeners also pick up on these cues when the speaker is a robot? Therefore the following research question is proposed: What effect has intonation on naturalness in robot speech in human-robot interaction?

To answer this question, there are some other questions that need to be answered first: What is intonation? What are the effects of intonation within human-human interaction? What is naturalness? What are other variables that could be affected by intonation? What is the current state within human-robot interaction? These will be answered with literature and previous research in the next section. Then, in the methodology section, the resources, approach, and method of evaluation will be discussed. In the result section the results will be discussed and hypotheses posed in the literature section will be accepted or rejected. After this, the results will be connected to the literature and limitations are discussed in the discussion. Also, some future work is proposed. The paper ends with a general conclusion.

2 LITERATURE REVIEW

2.1 Intonation

To be able to see what effect intonation has on naturalness, it is first necessary to find a consensus on what intonation is and does within HHI. Intonation is a combination of three prosodic features: pitch (talking high or low), loudness (shouting vs. whispering), and segmental duration (fast vs. slow) [22]. By varying in intonation, a speaker can highlight or phase (part of) their utterance to give focus points for the listener [5]. Other functions are expression of emotion and marking sentence modality [9, 22]. Intonation has certain patterns to express this sentence modality, but also to make clear to the listener when it is an appropriate time to switch turns, i.e. when there is a transition-relevance place (TRP) [11]. These patterns differ for each language, which can lead to bad interpretations of the intonation when interacting in a multicultural environment [25]. For instance, while in British English a high pitch means friendliness, in Dutch it means emphasis. This is because Dutch has a smaller pitch range than British English [9]. Since the experiment is run in Dutch, all functions of intonations that are discussed here apply to the Dutch language, unless specified otherwise. There are two main types of intonation: intonation on focus points and intonation at the end of an utterance. Together, this creates full intonation. Since intonation is mostly done subconsciously, people normally talk with full intonation [34].

2.1.1 *Focus Points.*

Intonation on focus points is the intonation on specific words in utterances. When speakers speak with a higher pitch at moments within their utterances, they emphasise this, and therefore communicate to the listener that this is important information. This is called 'focalisation' [25]. With focalisation, a contrast can also be made clear. Another function of adding intonation on focus points is expressing how the speaker feels about what they are saying. When speaking in a monotonous voice (without varying pitch or speed), no emotion is attributed to this speech. However, when a speaker varies in tone, they can communicate to the listener that they are happy, sad, or surprised, for example. Also, emotion evokes emotion [48]. When one of the conversation partners shows more emotion, the other one will mimic this by showing emotions themselves.

2.1.2 *End of Utterance.*

Another form of intonation is the intonation people use at the end of their utterances. This lets the listener know two things: what the modality of the utterance is and when it is appropriate to switch turns. A mode shows what the speaker wants to do with their utterance. For example, they can state their utterance as a fact, or ask a question. An utterance can have one of four modes: declarative, interrogative, exclamatory, and imperative [19]. The speaker chooses which one is appropriate and adjusts their intonation to that mode. In Dutch, a declarative utterance usually ends with low intonation. An interrogative ends with a rising intonation. An exclamatory is mostly presented by varying loudness, mainly at the end. An imperative shares its intonation pattern with the declarative mode, but is expressed more loudly [45] [9]. One thing to keep in mind is that other factors might play a role in clarifying the utterance modality, the main one being the word order in the interrogative mode, since questions usually have an inversed subject and verb [19].

Turn-taking is a process that people do unconsciously and therefore naturally, but when it goes wrong (because of misinterpreted end intonation), inappropriate interruptions take place, repairs are necessary and longer pauses occur [30, 41]. End-of-utterance intonation signals to the listener when it is appropriate to switch turns (a TRP, transition-relevance place). In an average conversation these TRPs create a silence between 0 and 200 milliseconds [44]. Longer means that there either is a need for some thinking, or a misunderstanding in turn-taking. When the switching of turns occurs within the utterance of one person, an interruption occurs, which is also not preferred, unless it is a continuer, like "hmm" of "ok". These continuers are signals of continuation where the listener that uses this, does not take a full turn, but wants to let the speaker know that they got their attention [40].

2.2 Naturalness

As stated before, the interaction between robots and humans is aimed to be as natural for the humans as possible [12, 21, 27, 36]. But what is natural? A natural conversation within HRI is an interaction that has the same conversational flow as humans have interacting with each other [18, 27]. This means the same rules of turn-taking occur and the actors in the conversation do not have to think about how they speak with each other [39]. However, there is a difference between perceived naturalness and objective naturalness [27]. Perceived naturalness is how the actor, in this case the human actor talking with the robot, feels like the interaction was easy and usual. Objective naturalness is the factual similarity with human-human interaction (HHI), where there are the same amount of turns between actors, and there is not much disagreement or miscommunication between them. These two forms of naturalness are not necessarily in agreement with each other, since humans bring their thoughts and expectations about the other actor. This can have an influence on the level of naturalness they expect [14]. Therefore, it is useful to measure both types of naturalness, using conversation analysis as the method to assess the objective naturalness, and a questionnaire to assess the subjective naturalness [43].

Other aspects to achieve naturalness in HRI are already researched extensively. For example, researchers have investigated what appearance a robot should have [36]. There also have been researchers looking into human behaviours in speech, which are still quite problematic for robots to do and comprehend, like emotion [7, 12], humor [46] and laughter [6]. These types of behaviour are linked to intonation, because they are often used in human speech to convey emotions. However, since these types are all based on the functions of intonation, only part of the intonation is researched and there is an assumption that the same rules of intonation apply to HRI as to HHI. This may not always be the case. For example, turn-taking is not done in the same way in quasi-synchronous computer-mediated communication as in oral conversation [20]. For instance, placement of the utterance was not deemed as important by the participants as clarity of the utterance. This could indicate that humans do not expect the same rules to apply when the communication is different than their usual HHI.

Since full intonation is most like how people talk, it was hypothesized that participants who talked to a robot with fully implemented intonation would have a more natural conversation, both objectively and subjectively. When comparing intonation on focus points and intonation on the end of utterances, end of utterances is hypothesised to have most effect on naturalness, since this should help the turn-taking system and therefore make the conversation more natural.

H1a: Full intonation is subjectively most natural.
H1b: Full intonation is objectively most natural.
H2a: Intonation on end of utterances has most effect on subjective naturalness.
H2b: Intonation on end of utterances has most effect on objective naturalness.

2.3 Sociability

Emotion measures the attitude towards the robot. When the person shows more emotion, the robot is perceived as more human-like [42]. This creates another variable to be worth measuring: sociability. This can then be measured by the number of emotions a person shows, which is objective data. Also the gaze of the person is interesting, since looking at your conversation partner shows interpersonality [3]. According to Mirenda et al. [32], an adult looks at the other not more than 60% of the conversation, on average. Therefore, it also can be said that when the person shows more emotion and looks more at the robot, the robot is seen more as a social being [43]. Since emotion evokes emotion, and intonation on focus points shows emotion, it can be hypothesised that this type of intonation will have most effect on sociability. Also, it is expected that full intonation results in the robot being the most social, because that is what we expect in HHI.

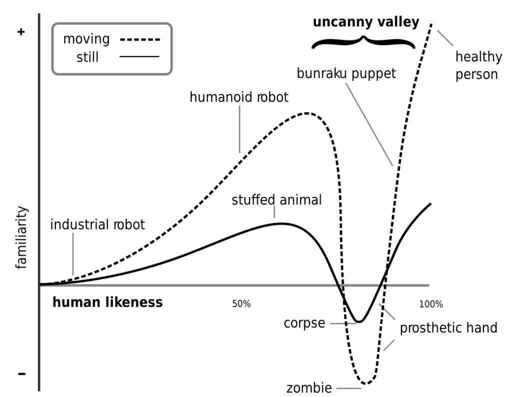
H3a: Intonation on focus points has most effect on subjective sociability.
H3b: Intonation on focus points has most effect on objective sociability.
H4a: Full intonation is subjectively most sociable.
H4b: Full intonation is objectively most sociable.

2.4 Eeriness

Also, some precaution is needed when trying to achieve naturalness in HRI. When robots behave human-like in one aspect, but lack this humanness in other aspects, there is a mismatch and the robot will be perceived as uncanny. This is what Mori described in 1970 as the Uncanny Valley [35]. It states that we are able to make robots human-like only to a certain extent, after which we will find ourselves in the Uncanny Valley. This means people do not want to interact with the robots, since they are perceived as creepy (see figure 1). To get to the other side of the valley is still a long way to go. This has been shown more recently, when chatbots were made more complex and animated, and people found them eerie and uncomfortable to talk with [10]. Full intonation is hypothesised to be least eerie, since this is closest to human speech, whereas when only one of the two intonation types is implemented, it creates a mismatch. A conversation partner talking without intonation is also not usual for the person, and will probably make the interaction more uncomfortable as well.

H5: Full intonation is least eerie.

Figure 1: The Uncanny Valley, as described by Mori [35].



To summarize, intonation has two main parts in human speech: intonation on focus points [5] and intonation on the end of utterances [41]. This research will then investigate how these types of intonation influence the naturalness of the HRI (in Dutch). Together with naturalness, the perceived eeriness of the robot will be questioned, to see if the robot does not belong in the Uncanny Valley. How this research will be done, is discussed in the next section.

3 METHODOLOGY

In a between subjects design, 130 participants interacted in a brief, casual conversation with a NAO robot in one of four conditions (4x: no intonation vs. intonation on focus points vs. intonation at the end of utterances vs. both intonation on focus points and at the end) to investigate the naturalness of the conversation, and the perceived eeriness of the robot. To test these variables, both quantitative and qualitative data were collected.

3.1 The Robot

The available robot was a Dutch-speaking NAO from Softbank Robotics, equipped with the NaoQi 1.14.1 software development kit (see figure 2) and the Choregraphe Suite. This robot also has extensive documentation online [47]. NAO's intonation was manipulated using Acapela tags [17] and Python code within the Choregraphe Suite.

3.2 Approach

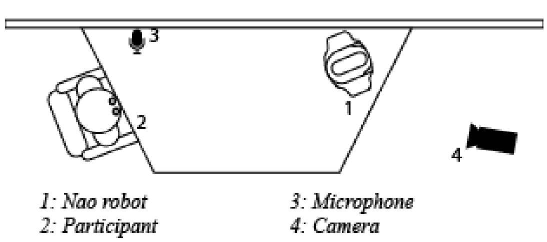
Participants had a conversation with NAO, which was recorded (both audio and video). Each participant talked with NAO in one

Figure 2: NAO (Softbank Robotics)

of four conditions: (1) NAO without intonation, (2) NAO with intonation at the end of utterances, (3) NAO with intonation on focus points, and (4) NAO with both intonation at the end as on focus points. The conditions were randomly assigned. Each conversation was between two and six minutes (*M*=3.26, *SD*=0.99) and in the form of a casual conversation about movies and games. This topic had been chosen, because it is often discussed among people in this age group and therefore increases naturalness, and because it has been used in previous research [33]. It allows questions, anecdotes, and opinions, and is not bound to constraints, which is, for example, the case in interviews. The participants were not informed about the goal of the conversation.

The setup can be seen in figure 3. The NAO was on the table in sitting position, with the robot and the participant facing each other. The camera was placed behind the robot to see the participant's face. A microphone was placed near the participant to record the audio separately from the video. The researcher was not in the room during the experiment.

Figure 3: Setup of the experiment.

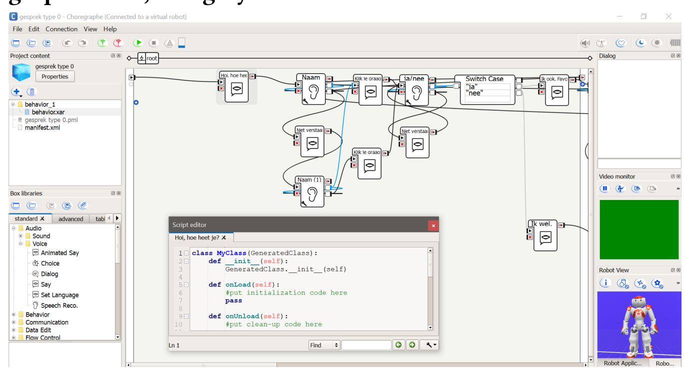


After the conversation with the robot, the participant was seated at a different table to fill in the questionnaire on a laptop. This was to limit the presence of the robot during the evaluation, so this would minimize the influence on the participant's answers. Again, the researcher was not in the room during this.

3.3 Manipulation

To create the four different conditions, first a conversation was constructed in the Choregraphe Suite, using Python boxes (see figure 4. Then, the intonation that was already integrated in the system was stripped away as much as possible to create the first condition: no intonation. This was done by leaving out the interpunction, having only short pauses, maintaining flat pitch and flat speed, and de-emphasise words that were automatically emphasised (stressed). The pause, pitch, speed and emphasis manipulations were done using Acapela tags, respectively \pau\, \rpit\, \rspd\, and \emph\ [17]. An example of code to make the robot speak without intonation (condition 1) is given in listing 1.

Figure 4: A snapshot of the baseline conversation in Choregraphe Suite, using Python boxes.



Listing 1: Example of Python code in the no intonation condition

```
def onInput_onStart(self, p):
    tts = ALProxy("ALTextToSpeech")
    tts.setLanguage("Dutch")
    tts.say("\\rspd=100\\ \\rpit=70\\
    \\emph=0\\ Hoi, \\pau=250\\ \\emph=0\\
    ik \\emph=0\\ ben \\emph=0\\ Robin.
    \\pau=250\\ \\emph=0\\ Hoe \\rpit=50\\
    \\emph=0\\ heet \\emph=0\\ jij?")

    self.onStopped() #activate the output
    of the box
```

Condition 2 was with intonation on focus points. This was done by taking the first condition and adding the pitch and emphasis on these focus points. For instance, when the robot and participant were already talking about films, the robot said:

```
Mijn favoriete film is Spirited Away.
focus points: mijn, Away

My favorite film is Spirited Away.
focus points: my, Away
```

Of course, it is important here to take the context into account. If they had not been talking about films before, but about preferences, the focus point would shift from "favorite" to "film".

The third condition was with intonation at the end of utterances, or contour intonation. Again, this was added on top of the first condition to keep the same baseline, and used interpunction. The interpunction automatically added some appropriate intonation and pauses in the utterances. Sometimes an additional emphasis was necessary when interpunction did not achieve its goal.

The fourth condition was a combination of the second and third condition, where they were both implemented. An overview of the manipulation techniques are shown in table 1.

Table 1: Manipulation techniques of the intonation conditions.

Condition	Choregraphe handling	Acapela tags
no intonation	no interpunction	
	short pauses	\pau = 150\ or \pau = 250\
	keep flat pitch	\rpit = 70\
	keep flat speed	\rspd = 100\
	remove emphasis	\emph = 0\
focus points	varying pitch	\rpit\
	varying speed	\rspd\
	varying emphasis	\emph = 1\ or \emph = 2\
end of utterance	add interpunction	
both	add interpunction	
	varying pitch	\rpit\
	varying speed	\rspd\
	varying emphasis	\emph = 1\ or \emph = 2\

Some other manipulations concerned all intonation types. To keep the speech recognition during the conversations as human-like as possible, its visual expression was turned off (otherwise it would flash its eyes) and the "beep"-sound it made when initiating speech recognition and when it recognized a word, was also turned off. Furthermore, the AutonomousLife mode was enabled to have the robot blinking and making small movements with its head and arms, to resemble some human movement. Also, if words were not pronounced correctly by the robot, they were rewritten using the Dutch phonetic alphabet that Acapela Group has also provided [2].

3.4 Measurements

There are two types of data that were used for evaluation: factual data of the interviews and questionnaires. The first is to analyse the objective naturalness, the second to also include the subjective naturalness.

3.4.1 Objective Data. To analyse the objective data, transcripts and annotations of the conversations were used. These can be found in appendix B For each conversation, the following information was extracted from the transcripts to test naturalness, based on previous research [27]: (1) the number of turns between actors, (2) the number of re-prompts, (3) the number of interruptions, and (4) the average length of silence between turns. Also, to address whether the participant saw the robot as a social being, (5) the valence of the facial expressions of the participant and (6) the attention the participant paid to the robot by looking at it during the conversation were extracted from the video material. The first three items were counted, while transcribing the conversations. The fourth is the average length of all silences between when the robot stopped and the participant started talking. Emotions were noticed, counted, and given a valence of either positive or negative. An example of these is shown in figure 5

Figure 5: Example of a positive and negative emotion.

3.4.2 Subjective Data. The subjective data was retrieved from a questionnaire, that participants filled out online after their conversation with the robot. This questionnaire consisted of 39 'content'-questions about how they had perceived the conversation and what they thought of the robot, and 7 more questions to gather some information about the participants themselves (age, gender, education, previous robot experience, etc.). The full questionnaire can be found in appendix A. In the appendix the questions are in order, with the questions grouped per scale. However, when the participants filled in the questionnaire, they were presented in random order.

The 39 content-questions were divided into three categories: naturalness (10), sociability (21), and eeriness (8). All questions were put on a 7-point Likert scale and were based on validated scales from previous research. If they were not available in Dutch yet, the items went through a process of back-translation to validate the translations. Naturalness was measured using the conversational fluency scale from Mirnig et al. [33] to look at the conversation itself and the Godspeed I scale from Bartneck et al. [4] to see if the perceived naturalness of the robot itself was affected. Sociability was measured using the social presence scale and the perceived sociability scale, both scales from Heerink et al. [23], with one question inversed, and the warmth and competence scales of the Stereotype Content Model [16]. Eeriness was measured using the "anxiety towards discourse with robots" scale of the RAS (Robot Anxiety Scale) [37] [13]. However, this scale has been criticized for being only focussed on the negative [38]. To include a balanced scale that addresses both positive and negative by having antonyms opposed on a Likert scale, it was decided to include the eeriness scale from Ho and MacDorman [26].

The internal consistency between items of each construct was tested by calculating Cronbach's alpha, with the reversed coded items recoded. To increase Cronbach's alpha from .659 to .770, the second question of the conversational fluency scale was dropped, so four items remained to test it. The Godspeed I already had a Cronbach's alpha of .711 with all of its five items and was therefore sufficient. The social presence scale was also acceptable (5 items;

α=.755), and the perceived sociability had a reliability of α=.812, after dropping one item and having three remaining items. The warmth and competence scales were both reliable, respectively α=.728 and α=.803, both with six items. The RAS scale, unfortunately, was unreliable, with a Cronbach's alpha of .506, which did not improve by removing one or more items. The eerie scale from Ho and MacDorman (2017) was reliable, however, with a Cronbach's alpha of .785, after deleting two of the four items.

Table 2: Used measurements for objective and subjective naturalness.

Type of measurement	Items
Objective data	
naturalness	number of turns, re-prompts, interruptions, length of silence between turns
sociability	valence of emotion, gaze
Subjective data	
naturalness	conversational fluency, Godspeed I
sociability	social presence, perceived sociability, warmth and competence
eeriness	RAS S3, eeriness

3.5 Analysis

The recordings of the conversations were annotated using ELAN, a linguistic annotating software. A coding scheme was developed for gaze and emotion and employed by two coders, the second being a fellow student. An intercoder reliability test (Krippendorff's α) was performed on the first 30 recordings to check if the annotations were reliable between the coders. The results of this can be found in table 3. Krippendorff's α is seen as acceptable when $\alpha \geq .800$ [28]. Every item met this criterium.

Table 3: Krippendorff's alpha per objective measurement.

Measurement	Krippendorff's α
Participant's turns	.982
Robot's turns	.931
Re-prompts	.924
Interruptions	.858
Average silence	.971
Positive emotions	.853
Negative emotions	.847
Total emotions	.890
Percentage of no gaze	.968

After the annotation, the results of each condition were compared to each other, using statistical analysis. The test that was performed is a One-Way ANOVA, since each type of measurement was tested individually on the four categorical independent variables, namely the intonation levels [15].

3.5.1 Participants. 130 students were recruited on the campus of Utrecht University, and volunteered to participate in a study on interaction with robots. Due to technical problems, 10 of them had to be discarded, therefore in further analysis the data of 120 remaining participants is used, 30 per intonation type. They ranged in age from 17 to 28 years (1:M= 23.75, SD= 3.53; 2:M= 22.7, SD= 1.68; 3:M= 23.4, SD= 2.34; 4:M= 22.6, SD=1.34) and 52.5% were male (0: M=12, F=18; 1: M=20, F=10; 2: M=16, F=14; 3: M=15, F=15). 18.3% of all participants stated in the questionnaire that they had never seen a robot before, and 88.3% had never interacted with a robot before. Most of the students were either studying Natural Sciences (46,7%) or Mathematics & Informatics (33.8%).

4 RESULTS

In this section the results of the statistical analyses will be discussed. First, the data obtained from the questionnaire will be examined, after that the objective data will be discussed. With each result, the hypotheses that were based on the literature in chapter 2 will be reviewed.

4.1 Subjective Data

The questionnaire tested naturalness, and how the robot was perceived, with the constructs eeriness and sociability. One-way ANOVA tests were used to test to compare the types of intonation on these constructs.

4.1.1 Naturalness. Naturalness was tested with two measurements, the conversational fluency scale and the Godspeed I. The Godspeed I scale showed no significant differences between the different conditions (*M*=2.69, *SD*=0.78). However, the conversational fluency did differ significantly ($F(3,116)=3.91$, $p=.011$, $\eta^2=.092$). Next, a Tukey test was performed to look at the differences between the intonation types. Participants found the conversation with focus intonation (*M*=3.82, *SD*=1.18) more fluent than with no intonation (*M*=3.03, *SD*=0.86, $p=.016$). End-of-utterance intonation (*M*=3.71, *SD*=0.98) also made the conversation more fluent for participants ($p=.049$). However, there was no significant difference between end-of-utterance intonation and focus intonation, therefore, H2a is not confirmed. However, no significant difference was found between no intonation and full intonation (*M*=3.32, *SD*=0.96), therefore H1a is not confirmed. Figure 6 shows the distribution of the conversational fluency over all four conditions.

4.1.2 Perception of the Robot. No significant result was found that the robot was perceived as more sociable on either the perceived sociability scale (*M*=4.04, *SD*=1.16) or the social presence scale (*M*=3.08, *SD*=1.07), when the conditions were compared to each other with a One-Way ANOVA test. It was also not detected that the robot was warmer (*M*=5.14, *SD*=0.70) or more competent (*M*=4.55, *SD*=0.85). This means H3a and H4a could not be proved. However, a trend was observed of participants finding the robot talking with no intonation more eerie than the other conditions ($F(3,116)=2.18$, $p=.094$, $\eta^2=.053$), which can also be seen in figure 7. Thus, H5 cannot be confirmed, but the trend should not be ignored, since it was almost significant.

Figure 6: **Boxplot of the conversational fluency for each condition.**

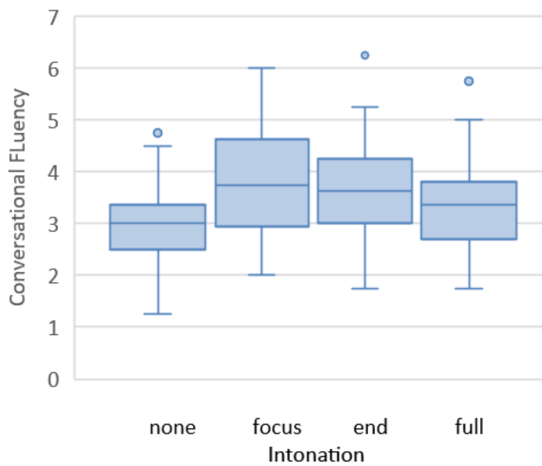
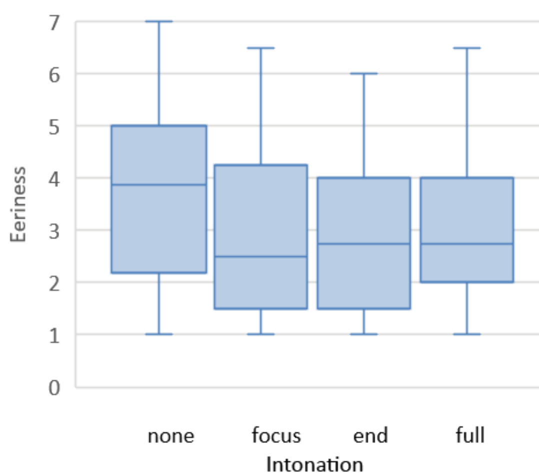


Figure 7: Boxplot of the eeriness for each condition.

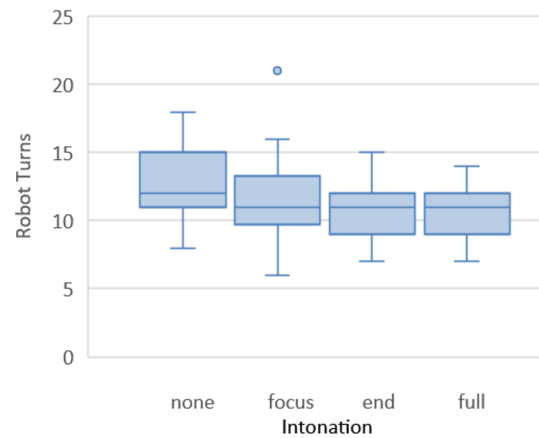


4.2 Objective Data

To measure the objective naturalness four measurements were used: the number of turns, the number of re-prompts, the number of interruptions, and the average length of silence after the robot has stopped talking. To address the sociability of the robot, two more measurements were used: the gaze of the participant towards the robot, and the number of positive and negative emotions. These were all tested by performing a One-Way ANOVA-test to compare each condition with each other for each type of measurement.

4.2.1 Naturalness. There were no significant results regarding the number of turns of the participant between the four intonation types (*M*=16.47, *SD*=3.89). However, the number of robot turns differed significantly (*F*(3,116)=5.88, *p*=.001, η^2=.132), with a quite large effect. When performing a Tukey test to see where this difference was exactly, it was found that when more turns of the robot were necessary when there was no intonation (*M*=12.93, *SD*=2.62) in comparison with end-of-utterance intonation (*M*=11.63, *SD*=2.95, *p*=.001) and full intonation (*M*=10.67, *SD*=2.16, *p*=.001). This can also be seen in the boxplot in figure 8.

Figure 8: Boxplot of the number of turns of the robot for each condition.

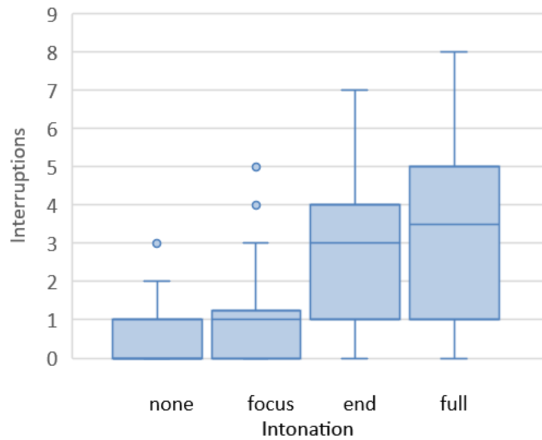


This means that the robot had to ask the participant to repeat what they had said, because the robot had not understood them. End-of-utterance intonation, thus, made the conversation more effective. However, no significant result was found on re-prompts by the participant between intonation types (*M*=3.03, *SD*=2.14), which means no difference was found for the number of times participants had to restate their answer. Another result was that there was a difference in number of interruptions (*F*(3,116)=20.71, *p*<.001, η^2=.349), with a very large effect. As can also be seen in figure 9, more interruptions occurred with end-of-utterance intonation (*M*=2.73, *SD*=2.03) and full intonation (*M*=3.40, *SD*=2.24), than when there was none (*M*=0.50, *SD*=0.77) or focus intonation (*M*=0.97, *SD*=1.27). Thus, participants did not know when the robot gave a proper transition-relevance place.

The length of silence between turns was not significant (*M*=0.90, *SD*=0.30), although on average quite long, since this is only 0 to 0.20 in HHI. This indicates that the end-of-utterance intonation did not improve the naturalness of the conversation much, only on the robot understanding the participants quicker, and it even influenced the naturalness negatively when looking at the turn-taking system. Therefore, H1b and H2b are not confirmed.

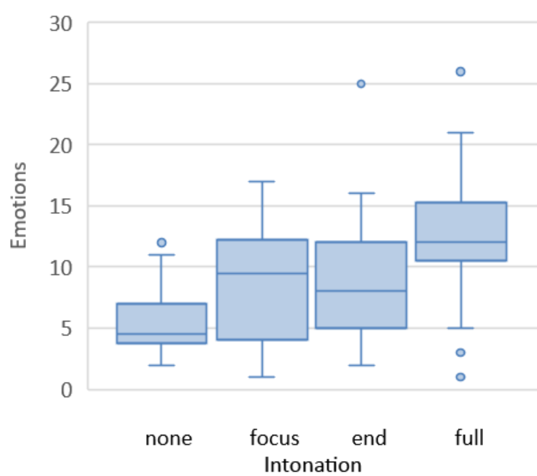
4.2.2 Sociability. To test if the participants saw the robot as a social being, the percentage of looking away from the robot by the participant and the number of emotions and their valence of the participant were measured. First, the amount that the participants looked away from the robot showed no significant differences among the conditions (*M*=13.09, *SD*=9.42). So, in general, the participants looked more at the robot than they would normally do in HHI (which is up to 60%, according to Mirenda et al. [32]), but no effect was found by the intonation. However, the number of emotions showed by participants did significantly differ, with a large effect,

Figure 9: Boxplot of the number of interruptions for each condition.



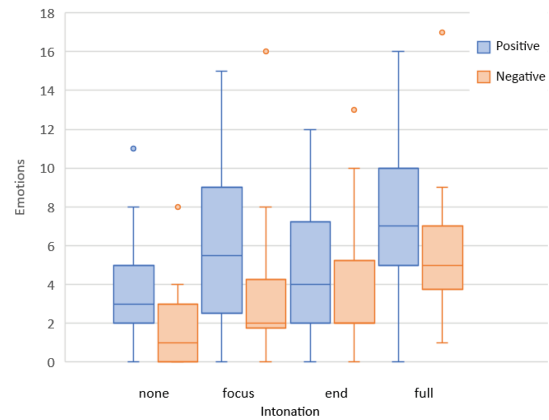
($F(3,116)$=11.21, p=.042, η^2=.225) among the conditions. When looking at the post-hoc test, participants showed more emotions when focus intonation was implemented (M=8.77, SD=4,69) than when no intonation was implemented (M=5.63, SD=2.98, p=.042). Also, the condition of no intonation evoked significantly less emotions than full intonation (M=12.40, SD=5.31, p<.001). Furthermore, focus intonation evoked less emotions than full intonation (p=.013). End-of-utterance intonation (M=8.53, SD=4.82) also evoked less emotions than full intonation (p=.007). So where no intonation was implemented, participants showed the least emotion, and full intonation evoked the most emotions, therefore hypothesis 4b was proved. As was hypothesised, focus intonation has the most influence on the number of emotions participants show (see also the boxplot in figure 10), and therefore H3b was confirmed.

Figure 10: Boxplot of the number of emotions for each condition.



When looking at the valence of these emotions, there was also a significant difference between intonation types, for both positive ($F(3,116)$=4.58, p=.002, η^2=.106) and negative emotions ($F(3,116)$=6.99, p<.001, η^2=.153). Looking at the post-hoc test, this difference was in both cases significant comparing the condition of no intonation (positive: M=3.77, SD=2.45, negative: M=1.87, SD=1.87) with full intonation (positive: M=7.10, SD=4.18, p=.002, negative: M=5.30, SD=3.09, p<.001). This can be seen in figure 11.

Figure 11: Boxplot of the number of positive and negative emotions for each condition.



5 DISCUSSION

To see if the type of intonation in robot speech has an effect on the naturalness, eeriness and sociability of the interaction and the robot itself, objective and subjective data was gathered from 120 interactions. Looking at the results, it can be concluded that at least some intonation is in the benefit of the conversation. No intonation performed poorly on both subjective and objective naturalness, since it scored low on conversational fluency and the robot needed to ask the person to repeat what they had said more often. This could be, for example, because the participant answered too quickly without the robot having activated their speech recognition, since this takes a while after having said something. However, not many interruptions occurred, but this could be explained by the fact that participants were not very engaged in the conversation, since they did not show many emotions, neither positive, nor negative. Also, a trend was revealed where a robot speaking with no intonation is perceived as more eerie than when the robot speaks with intonation. Since in the end the people's perception is the most important part (they need to be willing to talk with the robot), this trend should not be ignored.

Be that as it may, full intonation is also not ideal. While the robot does evoke a lot of emotions, the robot talking with no intonation did not have a more fluent conversation than the robot talking with full intonation, according to the participants. This is supported by the fact that a lot of interruptions occurred, and this also evoked more negative emotions. So, it seems people got more irritated, since the turn-taking system was not working properly, but they did care about it, and were then disappointed. So H1a and b, about full intonation being most natural, could not to be confirmed. Sociability was only objectively better with full intonation, therefore H4b is accepted, but H4a was not confirmed.

Another conclusion that can be drawn, is that the two different intonations do have different effects, although not always as expected. While a robot speaking with focus intonation does evoke more emotions than one with no intonation, this is also the case with the robot speaking with end-of-utterance intonation. Therefore, it cannot be said that focus intonation has the most effect on objective sociability, and H4b was not proved. Although the literature said focus intonation adds emotion [25] and emotion evokes emotion [48], the results show that end intonation already gives some emotion as well. Since the subjective sociability scales showed no significant results, H4a was also not confirmed.

Another unexpected outcome was that the robot with end-of-utterance intonation yielded more interruptions than the robot with no or focus intonation. Literature from the conversation analysis in HHI suggested that end-of-utterance intonation would show the conversation partner when it was the right time to switch turns. However, this seemed not to be the case when people talk with a robot. This can be explained by the fact that the silences between turns were on average 900 milliseconds, where in HHI pauses are usually between 0 and 200 milliseconds [44]. This throws the turn-taking system off guard and it cannot work properly anymore. These pauses are integrated within the interpunction used in Choregraphe, which is the standard way to implement this type of intonation with this software. To make a conversation in HRI as natural as possible, this should be adjusted to create shorter, more natural pauses, so the turn-taking system can hopefully work properly. However, for now, the results mean H2b was not confirmed. The same goes for H2a, since there was no difference found between the robot talking with focus intonation and it talking with end-of-utterance intonation. This could be explained by the fact that sometimes intonation was very persistent. This means the software did not allow the intonation to be completely stripped away on some words where it was necessary, which made the difference in intonation types a bit unclear at times.

One more finding was the gaze. No significant differences were found, so no effect was detected of the type of intonation on the sociability in terms of gaze. However, people tended to look at the robot more than they would do in HHI (around 87% instead of up to 60%) [32]. This could suggest that people are uncertain or uncomfortable while talking with the robot, and therefore keep an eye on the robot at (almost) all times.

5.1 Limitations & Future Work

Looking at all these results together, it seems intonation in HRI does not work the same as it does in HHI, which was already suggested for computer-mediated communication [20]. However, further research is needed to support this claim, since this was still an exploratory study. Also, some things did not work as expected and need more attention when setting up the experiment in future studies. For example, participants often returned with the remark that they felt they had to yell at the robot for it to understand them. So maybe the volume of the microphone can be adjusted. Another problem was that sometimes the intonation was quite persistent in the robot and could not be adjusted properly, which made the difference between focus and end-of-utterance intonation a bit cloudy at times and this may have affected the results.

For future research, a different approach might be better, like the Wizard-of-Oz-method, which is used in many HRI studies. This limits the misunderstandings of the robot and hereby decreases frustrations of participants where the robot gave wrong answer. It might also be interesting to use a different robot to see if that would make a difference, like a non-humanoid robot or an even more humanoid robot. Also, performing this study with the Pepper, the larger sibling of the NAO, could be interesting, since it has different intonation tags (NUANCE instead of Acapela) and these seem to be more straightforward and easier to implement. This robot might have less difficulty with persistent intonation. Furthermore, since intonation is language-bound, the results are only applicable to Dutch and so this study should be repeated with other languages and their functions and forms of intonation to see the effects of intonation on these languages.

6 CONCLUSION

To summarize, this study showed that a robot should at least have some intonation when talking with a human being. When a robot does not use intonation, it is less natural and less sociable than when it does use intonation. Also, people seemed to find the robot eerier when this happens, but this could not be concluded, since only a trend was spotted. However, full intonation is not most natural or sociable with respect to the other types of intonation. It is possible that this is because of how end-of-utterance intonation was implemented in this study, since this caused long pauses between the utterances of the robot. This was probably the cause of the many interruptions occurring when the robot spoke with the end-of-utterance or full intonation. Therefore, in future research, this type of intonation should be manipulated in a different way. Focus intonation did achieve its expected goal, namely evoking more emotions, and therefore making the interaction more sociable.

REFERENCES

[1] Aarestrup, M., Jensen, L. C., and Fischer, K. The sound makes the greeting: Interpersonal functions of intonation in human-robot interaction. *2015 AAAI Spring Symposium Series* (2015).

[2] Acapela Group. *Language Manual: HQ and HD Dutch.* Mons, Belgium, March 2011.

[3] Andersen, P. A., and Coussoule, A. R. The perceptual world of the communication apprehensive: The effect of communication apprehension and interpersonal gaze on interpersonal perception. *Communication Quarterly 28* (1980).

[4] Bartneck, C., Croft, E., Kulic, D., and Zoghbi, S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics 1*, 1 (2009), 71–81.

[5] Baumann, S., and Grice, M. The intonation of accessibility. *Journal of Pragmatics 38*, 10 (2006), 1636–1657.

[6] Becker-Asano, C., Kanda, T., Ishi, C., and Ishiguro, H. How about laughter? perceived naturalness of two laughing humanoid robots. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (2009), pp. 1–6.

[7] Breazeal, C. Emotive qualities in robot speech. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems.* (2001), vol. 3, pp. 1388–1394.

[8] Breazeal, C. Social interactions in HRI: The robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 34*, 2 (2004), 181–186.

[9] Chen, A. J. *On the universal and language-specific perception of paralinguistic intonational meaning.* PhD thesis, University of Utrecht, 2005.

[10] Ciechanowski, L., Przegalinska, A., Magnuski, M., and Gloor, P. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems 92* (2019), 539–548.

[11] Clayman, S. E. *Turn–constructional units and the transition–relevance place.* 2012, pp. 151–166.
[12] Crumpton, J., and Bethel, C. L. A survey of using vocal prosody to convey emotion in robot speech. *International Journal of Social Robotics 8*, 2 (2016), 271–285.
[13] De Graaf, M. M., and Allouch, S. B. The relation between people's attitude and anxiety towards robots in human-robot interaction. In *2013 IEEE RO-MAN* (2013), pp. 632–637.
[14] De Graaf, M. M. A., and Allouch, S. B. The influence of prior expectations of a robot's lifelikeness on users' intentions to treat a zoomorphic robot as a companion. *International Journal of Social Robotics 9*, 1 (2017), 17–32.
[15] Field, A. P. *Discovering statistics using SPSS.* SAGE, London, England, 2015.
[16] Fiske, S. T., Cuddy, A. J., Glick, P., and Xu, J. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology 82*, 6 (2002), 878–902.
[17] Florent, G. *Acapela TTS for Windows, Mac and Linux – User's Guide*, 40 ed. Acapela Group, Mons, Belgium, June 2015.
[18] Fong, T., Nourbakhsh, I., and Dautenhahn, K. A survey of socially interactive robots. *Robotics and autonomous systems 42*, 3-4 (2003), 143–166.
[19] Gå rding, E. Intonation in Swedish. In *Intonation systems: a survey of twenty languages*, D. Hirst and A. Di Cristo, Eds. Cambridge University Press, 1998, pp. 96–111.
[20] Garcia, A. C., and Baker Jacobs, J. The eyes of the beholder: Understanding the turn-taking system in quasi-synchronous computer-mediated communication. *Research on language and social interaction 32*, 4 (1999), 337–367.
[21] Ghosh, S., and Pherwani, J. Enabling naturalness and humanness in mobile voice assistants., 2015.
[22] Grice, M., and Baumann, S. *An introduction to intonation–functions and models.* 2007, pp. 25–52.
[23] Heerink, M., Krose, B., Evers, V., and Wielinga, B. Measuring acceptance of an assistive social robot: a suggested toolkit. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication* (September 2009), pp. 528–533.
[24] Hill, J., Ford, W. R., and Farreras, I. G. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior 49* (2015), 245–250.
[25] Hirst, D., and Di Cristo, A. E. *Intonation systems: a survey of twenty languages.* Cambridge University Press, 1998.
[26] Ho, C. C., and MacDorman, K. F. Measuring the uncanny valley effect. *International Journal of Social Robotics 9*, 1 (2017), 129–139.
[27] Hung, V., Elvir, M., Gonzalez, A., and DeMara, R. Towards a method for evaluating naturalness in conversational dialog systems. In *2009 IEEE International Conference on Systems, Man and Cybernetics* (October 2009), pp. 1236–1241.
[28] Kripendorff, K. Reliability in content analysis: Some common misconceptions. *Human Communications Research 30* (2004), 411–433.
[29] Mavridis, N. A review of verbal and non–verbal human–robot interactive communication. *Robotics and Autonomous Systems 63* (2015), 22–35.
[30] Mazeland, H. J. *Inleiding in de conversatieanalyse.* Coutinho, 2003.
[31] Mihalicek, V., and Wilson, C. E. *Language Files: Materials for an Introduction to Language and Linguistics.* Columbus, OH: The Ohio State University Press, 2011.
[32] Mirenda, P. L., Donnellan, A. M., and Yoder, D. E. Gaze behavior: A new look at an old problem. *Journal of Autism and Developmental Disorders 13*, 4 (1983), 397–409.
[33] Mirnig, N., Weiss, A., Skantze, G., Al Moubayed, S., Gustafson, J., Beskow, J., Granström, B., and Tscheligi, M. Face-to-face with a robot: What do we actually talk about? *International Journal of Humanoid Robotics 10*, 01 (2013), 1350011.
[34] Mitrofanova, Y. Raising EFL students'awareness of English intonation functioning. *Language Awareness 21*, 3 (2012), 279–291.
[35] Mori, M., MacDorman, K. F., and Kageki, N. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine 19*, 2 (2012), 98–100.
[36] Nishio, S., Ogawa, K., Kanakogi, Y., Itakura, S., and Ishiguro, H. Do robot appearance and speech affect people's attitude? Evaluation through the ultimatum game. *Geminoid Studies: Science and Technologies for Humanlike Teleoperated Androids* (2018), 263–277.
[37] Nomura, T., Kanda, T., Suzuki, T., and Kato, K. Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE transactions on robotics 24*, 2 (2008), 442–451.
[38] Prakash, A., and Rogers, W. A. Why some humanoid faces are perceived more positively than others: effects of human-likeness and task. *International journal of social robotics 7*, 2 (2015), 309–331.
[39] Sacks, H., Schegloff, E. A., and Jefferson, G. *A simplest systematics for the organization of turn taking for conversation.* Academic Press, 1978, pp. 7–55.
[40] Schegloff, E. A. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. *Analyzing discourse: Text and talk 71* (1982), 93.
[41] Schegloff, E. A. Turn organization: One intersection of grammar and interaction. *Studies in interactional sociolinguistics 13* (1996), 52–133.
[42] Sirithunge, H. C., Muthugala, M. V. J., Jayasekara, A. B. P., and Chandima, D. P. A wizard of oz study of human interest towards robot initiated human–robot interaction. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO–MAN)* (2018), pp. 515–521.
[43] Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., and Goodrich, M. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction* (March 2006), ACM, pp. 33–40.
[44] Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J., Yoon, K., et al. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences 106*, 26 (2009), 10587–10592.
[45] 't Hart, J. Intonation in Dutch. In *Intonation systems: a survey of twenty languages*, D. Hirst and A. Di Cristo, Eds. Cambridge University Press, 1998, pp. 96–111.
[46] Tay, B. T., Low, S. C., Ko, K. H., and Park, T. Types of humor that robots can play. *Computers in Human Behavior 60* (2016), 19–28.
[47] Van Straten, C. L. Looks good, sounds nice: Intonation and bodily appearance in robot–mediated communicative treatment for children with autism., 2016.
[48] Wild, B., Erb, M., and Bartels, M. Are emotions contagious? Evoked emotions while viewing emotionally expressive faces: quality, quantity, time course and gender differences. *Psychiatry Research 102*, 2 (2001), 109–124.

Appendices

A QUESTIONNAIRE

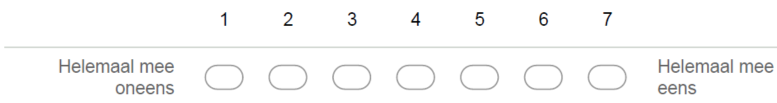
Praten met een robot

U heeft net een gesprekje gehad met een robot. De volgende vragen gaan over uw ervaring gedurende en na dit gesprekje. Aan het eind van deze vragenlijst volgen wat vragen omtrent uw demografische achtergrond.

Nu volgen een aantal stellingen over uw ervaringen van het gesprek dat u zonet met de robot heeft gehad. Geeft u a.u.b. per stelling aan in hoeverre u het er mee eens of oneens bent. Het gaat hierbij om uw persoonlijke mening. Er is dus geen goed of fout antwoord. (*Every statement was followed by the same 7-point Likert scale as in figure 12.)*

(1) De robot reageerde snel genoeg.

Figure 12: 7-point Likert scale for statements.



(2) De robot viel mij in de rede.
(3) De robot heeft verstaan wat ik zeg.
(4) Het gesprek met de robot verliep vloeiend.
(5) Het gesprek met de robot verliep gemakkelijk.
(6) Ik vreesde voor de manier waarop ik met de robot moest praten.
(7) Ik vreesde voor de manier waarop ik moest reageren wanneer de robot tegen mij sprak.
(8) Ik vreesde dat de robot niet zou begrijpen waarover ik sprak.
(9) Ik vreesde dat ik niet zou begrijpen waarover de robot sprak.
(10) Tijdens het gesprek met de robot, voelde het alsof ik met een echt persoon praatte.
(11) Soms voelde het alsof de robot echt naar mij keek.
(12) Ik kan me de robot voorstellen als een echt persoon.
(13) Ik dacht vaak dat de robot geen echt persoon was.
(14) Soms leek het alsof de robot echte gevoelens had.
(15) Ik vind de robot een prettige conversatiepartner.
(16) Ik vind de robot prettig in de omgang.
(17) Ik heb het gevoel dat de robot begrip voor me heeft.
(18) Ik vind de robot aardig.
(19) De robot was bekwaam.
(20) De robot was zelfverzekerd.
(21) De robot was kundig.
(22) De robot was efficiënt.
(23) De robot was intelligent.
(24) De robot was vaardig.
(25) De robot was vriendelijk.
(26) De robot was sympathiek.
(27) De robot was betrouwbaar.
(28) De robot was warm.
(29) De robot was goedaardig.
(30) De robot was oprecht.

Geeft u a.u.b. per tegenstelling aan welke indruk u van de robot heeft. Nogmaals, het gaat hierbij om uw persoonlijke mening. Er is dus geen goed of fout antwoord. *(Every opposition was put on a 7-point Likert scale as in figure 13.)*

(31) saai – bizar
(32) voorspelbaar – ongemakkelijk
(33) gewoon – eigenaardig
(34) alledaags – vreemd

Figure 13: 7-point Likert scale for oppositions.

	1	2	3	4	5	6	7	
saai	○	○	○	○	○	○	○	bizar

(35) onecht – natuurlijk
(36) lijkend op een machine – lijkend op een mens
(37) onbewust – heeft een bewustzijn
(38) kunstmatig – levensecht
(39) houterige bewegingen – vloeiende bewegingen

Demografische gegevens

(40) Wat is uw leeftijd?
____

(41) Wat is uw geslacht?
- ○ Man
- ○ Vrouw
- ○ Anders

(42) Wat is uw hoogst afgeronde opleiding?
- ○ Geen, Basisonderwijs, Lager beroepsonderwijs (LBO, VMBO), Middelbaar algemeen voorbereidend onderwijs (MAVO)
- ○ Middelbaar beroepsonderwijs (MBO), Hoger algemeen voorbereidend, wetenschappelijk onderwijs (HAVO, VWO)
- ○ Hoger beroepsonderwijs (HBO), kandidaats/propedeuse wetenschappelijk onderwijs, (post) wetenschappelijk onderwijs
- ○ Weet ik niet / Zeg ik liever niet

(43) Wat doet u in het dagelijks leven?
- ○ Ik studeer
- ○ Ik werk
- ○ Allebei, maar vooral studeren
- ○ Allebei, maar vooral werken
- ○ Geen van beide

(44) Wat is uw werk/studieveld?
- ○ Bedrijf en economie
- ○ Filosofie en religie
- ○ Gedrag en maatschappij
- ○ Gezondheid en beweging
- ○ Informatica en wiskunde
- ○ Kunst, cultuur en geschiedenis
- ○ Natuurwetenschappen
- ○ Recht en bestuur
- ○ Talen en communicatie
- ○ Anders: ____________________

(45) Heeft u al eens eerder een robot gezien?
- ○ Ja
- ○ Nee

(46) Heeft u al eens eerder met een robot gepraat?
- ○ Ja
- ○ Nee

(47) Heeft u nog vragen of opmerkingen? Dan kunt u deze hier plaatsen.

____

____

____

Einde van de vragenlijst.

Dit is het einde van de vragenlijst. Dank u voor uw deelname.

B TRANSCRIPTS

The transcripts are sent via a separate pdf file (appendixB.pdf), since it consists of 334 pages.