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

Ella Velner
p.c.velner@utwente.nl
University of Amsterdam, NL

Paul P.G. Boersma
paul.boersma@uva.nl
University of Amsterdam, NL

Maartje M.A. de Graaf
m.m.a.degraaf@uu.nl
Utrecht University

ABSTRACT

Human-robot interaction (HRI) research aims to design natural interactions between humans and robots. Intonation, a social signaling function in human speech investigated thoroughly in linguistics, has not yet been studied in HRI. This study investigates the effect of robot speech intonation in four conditions (no intonation, focus intonation, end-of-utterance intonation, or combined intonation) on conversational naturalness, social engagement, and people's humanlike perception of the robot collecting objective and subjective data of participant conversations (n = 120). Our results showed that humanlike intonation partially improved subjective naturalness but not observed fluency, and that intonation partially improved social engagement but did not affect humanlike perceptions of the robot. Given that our results mainly differed from our hypotheses based on human speech intonation, we discuss the implications and provide suggestions for future research to further investigate conversational naturalness in robot speech intonation.

KEYWORDS

Conversation Analysis; Human-Robot Interaction; Linguistics; Speech Intonation; Turn Taking

ACM Reference Format:
Ella Velner, Paul P.G. Boersma, and Maartje M.A. de Graaf. 2020. Intonation in Robot Speech: Does it work the same as with people?. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20), March 23–26, 2020, Cambridge, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3319502.3374801

1 INTRODUCTION

One of the goals of human-robot interaction (HRI) research is to make robots appear more natural, in looks [44], behaviour [8, 12, 57], and speech [27, 38, 44], commonly done by applying theories from (social) psychology to a robot context. However, such an approach may raise some challenges. First is to avoid the uncanny valley [43] in which the naturalness and familiarity plunges into a surge of strangeness or eeriness due to subtle deviations from human norms. For example, a robot can move very humanlike but talk mechanically. Second is to manage people's expectations of their robotic partners given these expectations affect the interaction [19]. For example, people tend to use simpler language when interacting with artificial agents (compared to human agents) [30]. Further research is necessary to investigate to what extent the psychological mechanisms in human-human interaction align with or deviate from the interactions we have with robots.

A key feature of the robot in HRI is speech. Similar to the general goal of social robotics, speech engineers focus on making robot speech more natural thereby increasing conversational fluency between humans and robots [17, 27, 33, 44]. Optimizing robot speech has shown to be a demanding task given the various aspects that need to be taken into account. When speaking, humans simultaneously make speech decisions on six levels: phonetically, phonologically, morphologically, syntactically, semantically, and pragmatically [40]. Implementing all these decisions in a robot is a complex endeavor, especially given the cultural and language-based dependency of these decisions [14, 31]. Ideally, the effects of each speech decision should be studied separately as well as in (cor)relation to each other.

Intonation is one aspect of speech that has been thoroughly researched within human-human interaction [1, 34, 48], but has been somewhat missed in HRI. Intonation is the pattern of pitch within utterances [40], and can alter the meaning of a sentence by putting the emphasis on different words, 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 [14, 28]. Such intonation effects and interaction patterns flow naturally in human-human interactions [31]; but will human listeners pick up on these cues when interacting with a robot partner? In this paper, we examine the effect of intonation on robot speech on the naturalness of verbal interactions between humans and robots.

2 LITERATURE REVIEW

2.1 Intonation

Intonation, in the broad sense of the term, is a combination of three prosodic features: pitch (talking high or low), loudness (shouting vs. whispering), and segmental duration (fast vs. slow) [28]. By varying in intonation, a speaker can: (1) differentially divide their utterance into phrases or emphasize different parts of their utterance so that these receive focus for the listener [6]; (2) express emotion or mark sentence modality (i.e., indicating whether the sentence is a statement or question) [14, 28]; and (3) signal the listener when it is an appropriate time to switch turns [16]. Intonation patterns differ by language [14, 31]. Given that the present study has been conducted in The Netherlands, this paper focuses on the two main loci of intonation in Dutch: at focus points and at the end of an utterance.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HRI '20, March 23–26, 2020, Cambridge, United Kingdom
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6746-2/20/03…$15.00
https://doi.org/10.1145/3319502.3374801

2.1.1 Focus Points. Intonation at focus points is the intonation on specific words in utterances and may have three different functions. First, when a conversation partner speaks with a higher pitch on a certain part of their utterance (*focalization*), this can mean that they aim to *emphasize* this part, i.e. to indicate to the listener that particularly this part contains important information [31]. Second, focalization can highlight information that is *new*, or information that *contrasts* with something else [36]. A third function of adding intonation at focus points is to express how the speaker feels about what they are saying: when a speaker uses a monotonous voice (without varying pitch or speed), listeners will attribute no emotion (or perhaps sadness) to this speech, whereas when a speaker varies in tone, they can communicate to the listener that they are happy, or surprised [4, 10]. Additionally, emotion evokes emotion [59]: when one of the conversation partners shows more emotion, the other partner will mimic this by showing emotions themselves.

2.1.2 End of Utterance. The intonation people use at the end of their utterances, on the other hand, informs the listener of two things: (1) what the modality of the utterance is; and (2) when it is appropriate to switch turns. A *mode* shows the functional intention of the speaker's utterance. An utterance can have one of four modes: declarative (making a statement), interrogative (asking a question), exclamatory (conveying a strong emotion), or imperative (giving a command or advice) [25]. The speaker chooses which one is appropriate and adjusts their intonation to that mode. In Dutch, a declarative utterance usually ends with low intonation, and an interrogative utterance with rising intonation [56]. An exclamation is usually presented by varying loudness, predominantly at the end of an utterance [56]. An imperative shares its intonation pattern with the declarative mode, but is expressed more loudly [14, 56]. One thing to keep in mind is that other factors may play a role in clarifying utterance modality, the main one (in Dutch) being word order, since Dutch questions usually have the subject–verb order reversed as compared to statements [25].

Turn-taking is a subconscious and therefore natural process for humans. When the turn-taking process fails (due to misinterpreted end intonation), inappropriate interruptions take place, repairs are necessary, and longer pauses occur [39, 52]. End-of-utterance intonation signals when it is appropriate for the listener to switch turns (i.e, a TRP, transition-relevance place). Listeners tend to avoid interrupting a continuous utterance by another person, unless their interruption is a continuer, such as "hmm" or "ok". These continuers are actually signals of continuation, where the listener does not take a full turn but wants to let the speaker know that they have their attention [51].

2.2 Evaluating Intonation in HRI

While not often, intonation has been a topic within the field of HRI before. Aarestrup et al. found out that different intonation contours, while only hearing either 'hi' or 'hello', elicited different interpretations for people [1]. Also, when only looking at the expressive qualities of intonation, people have personal preferences [34]. The automation of intonation generation by machine learning has also been studied, and works quite well [48]. However, to the best of our knowledge, the effects of intonation in HRI and the need of intonation have not yet been studied. The current study focuses on the effects of robot speech intonation on conversational naturalness, social engagement, and the humanlike perception of the robot.

2.2.1 Conversational naturalness. As stated before, HRI research aims to build robots capable of natural interactions with human users [17, 27, 33, 44]. A natural conversation within HRI is considered here as an interaction that has the same conversational flow as humans have interacting with each other [24, 33]. This implies the occurrence of similar rules of turn-taking as a subconscious and therefore natural process for the actors in the conversation [49]. However, there is a difference between perceived naturalness and objective naturalness in conversational fluency [33]. Perceived naturalness is how an actor, in this case a human actor talking with a robot, evaluates the conversation as easy and familiar. Objective naturalness is the factual similarity between human-human and human-robot conversations in terms of turns and disagreement or miscommunication between actors. Given that people's expectations of robots affect evaluations of human-robot interactions [19], these two types of conversational naturalness may not be in agreement with each other as a consequence of anticipated expectations about the level of naturalness in robot speech.

Broader aspects to achieve naturalness in HRI have already been studied. Researchers have investigated human behaviors in speech that are still quite problematic for robots to express and comprehend, such as emotion [11, 17], humor [57] and laughter [8]. These types of speech behavior are linked to intonation; intonation is often used in human speech to convey emotions. However, making direct claims for the effects of intonation in robot speech based on these studies is problematic in two ways. First, since these types of speech behaviors are all based on the functions of intonation, only part of the intonation is often researched. Second, these studies assume the direct application of established human rules of intonation to robot contexts. Yet, deviations from this assumption may arise. For example, turn-taking occurs somewhat differently in quasi-synchronous computer-mediated communication as in oral conversation [26]. In their study, the placement of the utterance was not deemed as important by the participants as clarity of the utterance. This indicates that humans may not expect similar rules when communicating with artificial conversational partners.

As a starting point to investigate the individual aspects of robot speech, this study will follow common HRI research practices by evaluating robot speech intonation along the lines of how this occurs in human speech. Based on human intonation research, we have formulated the following hypotheses for robot speech in which combined intonation is defined as having both focus point and end-of-utterance intonation:

H1a: Combined intonation is subjectively most natural.

H1b: Combined intonation is objectively most natural.

H2a: Intonation on end of utterances, as a facilitator of the turn-taking mechanism, contribute to a larger extent to subjective naturalness than focus point intonation.

H2b: Intonation on end of utterances contribute to a larger extent to objective naturalness than focus point intonation.

2.2.2 Social Engagement. Social engagement in HRI is often considered in terms of gaze and turn-taking [46]. Gazing at your conversation partner shows interpersonality (i.e., the quality of being

interpersonal) [3, 13]. Another way to evaluate people's engagement with a robot is by observing expressed emotions during the conversation. People who are more emotionally expressive during human-robot interactions and gazed more directly at the robot are more likely to perceive that robot as humanlike [53, 54]. Moreover, when a robot appropriately conveys emotion, a human partner will become more emotionally expressive [37] indicating that robot emotion evokes emotional mimicry similar as in human conversations [59]. Combining this with the knowledge that one of the functions of intonation on focus points in human speech is the convey emotion [4, 10], we have formulated the following hypotheses:

H3a: Intonation on focus points contribute to a larger extent to subjective social engagement than end-of-utterance intonation.

H3b: Intonation on focus points contribute to a larger extent to objective social engagement than end-of-utterance intonation.

H4a: Combined intonation evokes the most subjective social engagement.

H4b: Combined intonation evokes the most objective social engagement.

2.2.3 Humanlike perception of the Robot. Some precaution is advised when trying to achieve naturalness in HRI. When robots behave humanlike in one aspect, but lack this humanness in other aspects, a mismatch might occur and the robot will be perceived as strange or eerie (i.e., Mori's Uncanny Valley [43]). This phenomenon indicates that robots can be humanlike only to a certain extent, after which we will find ourselves in the Uncanny Valley. As a result, people may not want to interact with such robots given they are perceived as creepy. Indeed, people deem more complex and animated chatbots eerie and feel uncomfortable conversing with them [15]. Given that combined intonation is most humanlike, a robot implemented with such speech intonation is expected to be perceived least eerie whereas the implementation of only one of the two intonation types or none at all should create an eerie mismatch of what people expect[10]. Combined intonation could then also lead to more humanization of the robot [20]. Humanlikeness is greater when voice is combined with gestures [50], so a robot should not be motionless. Based on this knowledge, we formulated the following hypotheses:

H5: Combined intonation makes the robot most humanlike.

3 METHODOLOGY

In a between-subjects design, 130 participants interacted in a brief, casual conversation with a NAO robot in one of four conditions of robot speech intonation: no intonation, intonation on focus points, intonation at the end of utterances, or intonation both on focus points and at the end (combined). Both quantitative and qualitative data was collected to investigate the differential effects of these robot intonation types on three dependent variables: conversational naturalness, the user's social engagement, and the perceived humanlikeness of the robot.

3.1 The Robot and Speech Intonations

A Dutch-speaking NAO robot from Softbank Robotics, equipped with the NaoQi 1.14.1 software development kit and the Choregraphe Suite was deployed. This robot has extensive documentation online [58]. NAO's intonation was manipulated using Acapela tags [23] and Python code within the Choregraphe Suite. To create the four different conditions, first a conversation was constructed in the Choregraphe Suite, using Python boxes. Then, the default intonation already integrated in the system was stripped away as much as possible to create the first condition: no intonation. Specifically, all interpunction was erased, having only short pauses, maintaining flat pitch and flat speed, and de-emphasize words that were automatically emphasized (stressed). The pause, pitch, speed and emphasis manipulations were implemented using Acapela tags, respectively \pau\, \rpit\, \rspd\, and \emph\ [23]. An example of code to make the robot speak without intonation (condition 1) is given in listing 1.

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
```

For condition 2, the robot was implemented with intonation on focus points by taking the first condition and adding the pitch and emphasis on these focus points as suggested in the literature [10, 31, 47]. For instance, when the robot and participant were already talking about films, the robot said:

```
(I like that movie a lot.)
My favorite film is Spirited Away.
focus point: my
```

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 "my" to "film" [36].

For the third condition, the robot was implemented with intonation at the end of utterances, or contour intonation, using interpunction as suggested in the literature [14, 56]. Again, this was added on top of the first condition to keep the same baseline. Adding interpunction in the Choregraphe environment automatically adds some appropriate intonation and pauses in the utterances (like rising intonation when writing a question mark). Sometimes an additional emphasis was necessary when interpunction did not achieve its goal. For example, when the intonation was not strong enough on the end of a question by only adding a question mark, /*emph* = 1/ was added to the last word to signify that the robot asked a question.

In the fourth condition, the robot was implemented with a combination of focus intonation (condition 2) and end of utterance intonation (condition 3) which we refer to as combined intonation.

An overview of all four conditions and the applied 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\

Other signals in the default settings of NAO that may have affected the conversational fluency were disabled for all four conditions. These include visual expressions such as its eye color to indicate the robot was activated as well as its "beep"-sound to initiate speech recognition. This ensured us that participants would not get distracted nor would be able to derive information about the robot's state from those signals. The AutonomousLife mode, on the other hand, was deliberately enabled to have the robot blinking and making small movements with its head and arms as an attempt to resemble lifelike movement. This was expected to trigger greater perceived humanlikeness [50]. Additionally, if words were not pronounced correctly by the robot, they were rewritten using the Dutch phonetic alphabet that Acapela Group provides [2] to ensure words were not misunderstood.

3.2 Procedure

During the experiment, the robot was on the table in sitting position, with the robot and the participant facing each other (see Figure 1). A camera was placed behind the robot to record the participant's face, and a microphone was placed near the participant to record the audio separately from the video. The participants were not informed about the goal of the conversation beforehand.

After giving consent and the experimenter had left the room, participants were randomly assigned to have a casual conversation with the robot in one of the four conditions. Their conversations were recorded (both audio and video) and lasted between two and six minutes (M=3.26, SD=0.99). The conversational topic was on movies and games (see Figure 2); a popular topic among adolescents that allows questions, anecdotes, and opinions without being bound to constraints, and that has been successfully applied in previous HRI research focusing on conversations [42]. When the robot did not understand the participant, it asked for rephrasing.

After the conversation with the robot, the experimenter briefly reentered the room only to guide the participant to a different table

Figure 1: Setup of the experiment.

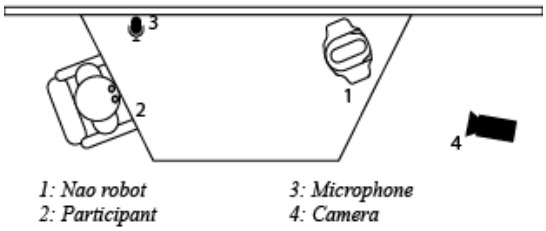


Figure 2: The conversational structure was identical for every condition.

Introduction: robot (R) introduced itself, participant (P) introduced themselves.
Movies: R asked if P likes movies. P answered they (dis)like movies, then R either:
(like) asked what their favorite movie was and answered with its opinion on that movie. Then R stated their favorite movie and asked if P knew it. R stated some facts about the movie if not known, otherwise it skipped this part.
(dislike) R said they like movies and tell their favorite movie including some facts about the movie.
Games: R asked if P liked games and whether they preferred board or video games. P replied positively/negatively, then R either:
(likes board) R answered approving.
(likes video) R stated they prefer board games.
(dislike games) R answered they like board games.
Chess: R stated they like chess and asked if P can play, then R either:
(yes) R stated they could play someday.
(no) R stated they could teach them someday.
Goodbye: after this, R said goodbye.

to complete a questionnaire on a laptop. With the robot out of sight at this stage, we aimed to minimize the influence of the robot's presence on the participant's answers. After the participant had completed the questionnaire, the experimenter would reenter the room once again to debrief the participant.

3.3 Measurements

We collected both objective and subjective data to examine the effect of robot speech intonation on conversational naturalness, the user's social engagement, and the perceived humanlikeness of the robot.

3.3.1 Objective Data. We used audio transcripts and video annotations of the conversations to capture objective data. To measure conversational naturalness, we annotated (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 [33]. To measure the user's social engagement, we also annotated (5) the number and valence of the facial expressions of the participant [13, 17] (for example, see Figure 3) and (6) the participant's direct gaze at the robot during the conversation [3, 13].

3.3.2 Subjective Data. The post-conversation questionnaire consisted of 39 items (in random order) to capture the participant's subjective evaluation of the conversation and the robot. All items were presented on a 7-point Likert scale, and a back-translation process was applied to validate our Dutch translations of the items. To measure the conversational naturalness we used the conversational fluency scale used by Mirnig et al.[42] (α = .772; after dropping one of five items). Social engagement was measured using two scales

Figure 3: Example of a positive and negative emotion.

Positive *Negative*

from Heerink et al. [29]: the social presence scale (α=.755; five items) and the perceived sociability scale (α=.811; after dropping one of four items). The participant's humanlike perception of the robot was obtained with the humanlikeness scale from Bartneck et al. [5] (α = .771; five items), and the warmth and competence scales from the Stereotype Content Model [22] (respectively α=.728 and α=.803, both with six items). To measure the participant's perceived eeriness, we used the scale from Ho and MacDorman [32] (α = .506.; deemed unreliable and thus excluded from further analysis), and the anxiety towards discourse with robots scale of the RAS (Robot Anxiety Scale) [18, 45] (α = .682; after dropping one of four items). The questionnaire ended with some general information about the participant (age, gender, educational background, previous robot experience).

3.4 Data Analysis

The recordings of the conversations were annotated using ELAN, a linguistic annotating software. A coding scheme was developed for the six objective measures (see above), and 25% was coded by two independent coders with an acceptable intercoder reliability (Krippendorff's $\alpha \geq .80$ for all annotated items). After the annotation, the results of each condition were compared to each other, using MANOVAs since each type of measurement was tested individually on the four categorical independent variables, namely the four intonation types in each condition [21].

3.4.1 Participants. We recruited 130 students on a Dutch university campus to participate in a study on interaction with robots. Due to technical problems, data of 10 participants had to be discarded. Further analysis was performed on the data of 120 remaining participants, 30 in each condition. The participants' age ranged from 17 to 28 years (M = 20.87, SD = 2.30) and 52.5% were male. 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

This section presents the results of the statistical analyses on conversational naturalness (H1 and H2), the users' social engagement (H3 and H4), and the humanlike perception of the robot (H5).

4.1 Conversational Naturalness

To explore the effect of robot intonation type on conversational naturalness, we performed a MANOVA including the conversational fluency scale together with the number of turns, re-prompts, interruptions, and average silence duration. The mean values and standard deviations are shown in table 2. The MANOVA showed a significant multivariate effect of robot intonation type on the naturalness of the conversation between humans and robots (Pillai's trace $F(18, 339) = 4.276$, $p < .001$, partial $\eta^2 = .185$). Each independent variable was subjected to a further ANOVA to show whether this trend was the same for each of the separate dependent variables. The subjective naturalness of the conversation differed significantly between intonations ($F(3, 116) = 3.905$, $p = .011$, partial $\eta^2 = .092$). Posthoc tests using Bonferroni adjustment revealed that the conversation with focus intonation (M = 3.82, SD = 1.18) as well as with end-of-utterance intonation (M = 3.71, SD = 0.98) were perceived as more natural than the conversation with no intonation (M = 3.03, SD = 0.86), with both pairs statistically significant (respectively, p = .016 and p = .049; see Figure 4. The number of robot turns also differed significantly between intonation types ($F(3, 116) = 6.092$, p = .001, partial η^2 = .136). Posthoc tests using Bonferroni adjustment revealed that the robot had more turns in conversations with no intonation (M = 12.93, SD = 2.62) compared to conversations with end-of-utterance intonation (M = 11.63, SD = 2.95) and combined intonation (M = 10.67, SD = 2.16), with both pairs statistically significant (respectively, p = .001 and p = .001; see Figure 5). Moreover, the number of interruptions ($F(3, 116) = 20.708$, $p < .001$, partial η^2 = .349) was different between intonation types. Posthoc tests using Bonferroni adjustment revealed that the robot interrupted participants more often in conversations with end-of-utterance intonation (M = 2.73, SD = 2.03) and with combined intonation (M = 3.40, SD = 2.24) compared to conversations with focus intonation (M = 0.97, SD = 1.27) and no intonation (M = 0.50, SD = 0.77), which were all statistically significant ($p < .001$ for these pairs; see Figure 6). However, the participants' turns ($F(3, 116) = 1.044$, *ns*, partial η^2 = .026), reprompts ($F(3, 116) = 0.252$, *ns*, partial η^2 = .006), and average silence duration ($F(3, 116) = 0.637$, *ns*, partial η^2 = .016) failed to reach statistical significance.

Based on these results, we found no support for hypotheses H1a+b and H2a+b since the data did not support our hypothesized superiority of combined intonation over no intonation nor the dominance of end-of-utterance intonation over focus point intonation in terms of conversational naturalness. Against our expectations, end-of-utterance and combined intonation evoked more interruptions than no intonation or intonation on focus points. The number of turns of the robot, however, was significantly less in the condition of combined intonation or end-of-utterance intonation than when the robot was talking with no intonation.

4.2 Social Engagement

To investigate the effect of intonation on the participants' social engagement during the conversation, we performed a MANOVA including the social presence and sociability scales together with the number of positive and negative emotions, and time gazing at the robot. The mean values and standard deviations are shown in table 3.

	Conversational Fluency		Robot Turns		Participant Turns		Interruptions		Reprompts		Avg. Silence Length	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
None	3.03*†	0.86	12.97*†	2.59	16.97	3.31	0.50*†	0.78	3.27	2.53	0.91	0.37
Focus	3.82*	1.18	11.63	2.95	15.57	4.32	0.97#‡	1.27	3.10	1.99	0.95	0.35
End	3.71†	0.98	10.67*	2.16	16.20	4.30	2.27*#	2.03	2.80	1.83	0.85	0.25
Combined	3.32	0.96	10.67†	1.81	17.13	3.50	3.40†‡	2.04	2.97	2.24	0.88	0.19

Table 2: Mean values and standard deviations for the items used to measure Conversational Naturalness. *, †, ‡, and # show significantly different pairs from post-hoc testing using Bonferroni, at the $p < .05$ level.

	Social Presence		Sociability		Positive Emotions		Negative Emotions		% Gaze	
	M	SD	M	SD	M	SD	M	SD	M	SD
None	3.03	1.02	3.70	0.92	3.77*	2.45	1.87*	1.87	87.79	9.77
Focus	3.12	1.09	4.13	1.17	5.33	3.85	3.43	3.31	86.78	6.76
End	3.05	0.93	3.84	1.23	4.83	3.51	3.70	3.14	89.56	6.01
Combined	3.11	1.26	3.80	1.23	7.10*	4.18	5.30*	3.09	83.52	12.92

Table 3: Mean values and standard deviations for the items used to measure Social Engagement. * show significantly different pairs from post-hoc testing using Bonferroni, at the $p < .05$ level.

Figure 4: Mean Conversational Fluency per intonation type. Error bars show standard error.

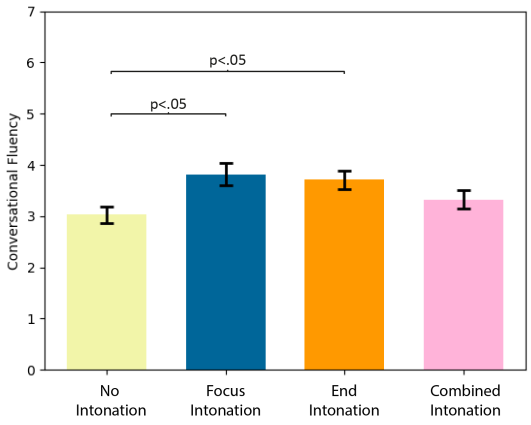
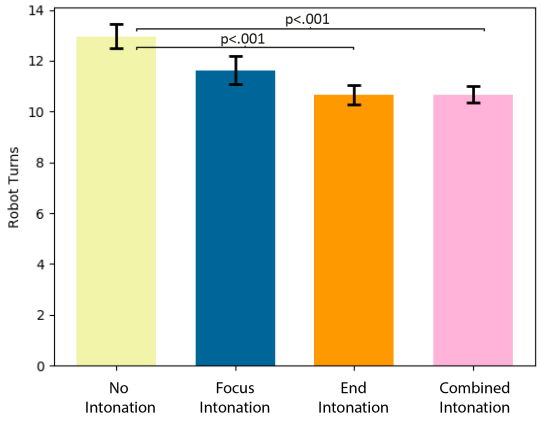


Figure 5: Mean Number of Robot Turns per intonation type. Error bars show standard error.



MANOVA showed a significant multivariate effect of intonation in robot speech on participants' social engagement while interacting with the robot (Pillai's trace $F(15, 342) = 2.656$, $p = .001$, partial $\eta^2 = .104$). Each independent variable was subjected to a further ANOVA to show whether this trend is the same for each of the separate dependent variables. A significant effect was observed for intonation type on expressed positive emotions by participants ($F(3, 116) = 4.583$, $p = .005$, partial $\eta^2 = .106$). Posthoc tests using the Bonferroni adjustment revealed that participants conveyed significantly more positive emotions ($p = .002$) during conversations with combined intonation ($M = 7.10$, $SD = 4.18$) compared to conversations with no intonation ($M = 3.77$, $SD = 2.45$; see Figure 7). A significant effect was also observed for intonation type on expressed negative emotions by participants ($F(3, 116) = 6.994$, $p < .001$, partial $\eta^2 = .153$). Similarly, posthoc tests using the Bonferroni adjustment revealed that participants also conveyed significantly more negative emotions ($p < .001$) during the conversation with combined intonation ($M = 5.30$, $SD = 3.09$) compared to the conversation with no intonation ($M = 1.87$, $SD = 1.87$; see Figure 8. However, the perceived social presence of the robot ($F(3, 116) = 0.053$, *ns*, partial $\eta^2 = .001$), its perceived sociability ($F(3, 116) = 0.795$, *ns*, partial $\eta^2 = .020$), and the time the participant gazed at the robot ($F(3, 116) = 0.869$, *ns*, partial $\eta^2 = .022$) failed to reach statistical significance.

Based on these results, our data do not provide support for (or against) H3a+b and H4a, i.e., our hypothesized superiority of combined intonation over no intonation and the dominance of focus point intonation over end-of-utterance intonation in terms of social engagement. The data did, however, partially support H4b given that the robot talking with combined intonation evoked more positive as well as negative emotions than the one with no intonation at all.

Figure 6: Mean Number of Interruptions per intonation type. Error bars show standard error.

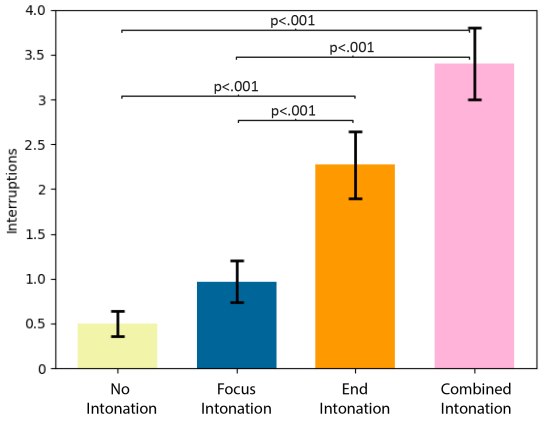
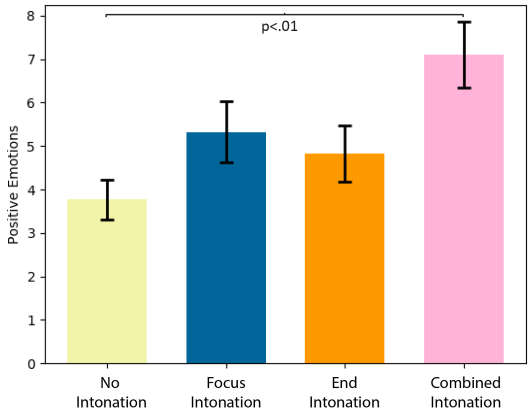


Figure 7: Mean Number of Positive Emotions per intonation type. Error bars show standard error.



4.3 Humanlike Perception

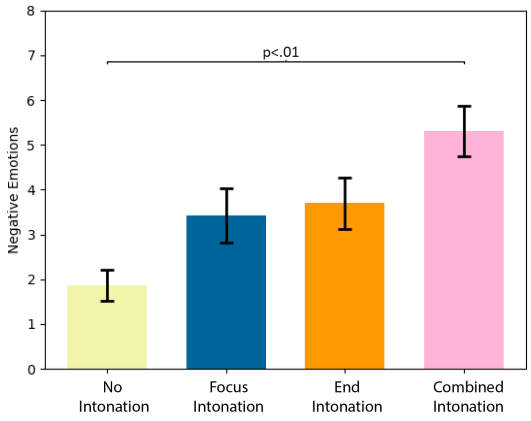
To examine the effect of intonation on the participants' humanlike perception of the robot, we performed a MANOVA including the Godspeed I, RAS, and warmth and competence scales. The mean values and standard deviations are shown in table 4.

MANOVA results show that the effect of intonation of the participants' humanlike perception of the robot did not reach statistical significance, Pillai's trace $F(12, 345) = 1.097$, *ns*, $\eta^2 = .037$. Therefore, we have no support for (or against) H5, given that the data does not support our hypothesized superiority of combined intonation over all other types of intonation in terms of the participants' humanlike perception of the robot.

5 GENERAL DISCUSSION

To investigate the effect of intonation type in robot speech on the conversational naturalness, the users' social engagement, and the humanlike perception of a robot, we have collected objective and subjective data from 120 interactions on a Dutch university campus.

Figure 8: Mean Number of Negative Emotions per intonation type. Error bars show standard error.



The combined results indicate that at least some form of intonation would be beneficial for the naturalness of the conversation between humans and robots. The robot implemented with no intonation was, both subjectively and objectively, less natural to converse with. Additionally, we observed that this robot tended to ask participants more often to repeat what they had said. One explanation could be that participants answered too quickly (a trend observed while annotating video recordings) before the robot had activated its speech recognition (which takes a while to start up after the robot's last utterance). Yet, the number of interruptions (by either the human or the robot) that occurred was not that high in the case of no intonation or focus intonation. This may be a result of the participants' lower level of objective social engagement (i.e., fewer emotions displayed, both positive and negative). Given that people who are more emotionally expressive during interactions with robots perceive that robot as more humanlike[37, 53], the lack of objective social engagement observed in our current study might explain our non-significant results on humanlike perceptions of the robot between the conditions.

Be that as it may, combined intonation has not shown to be most optimal either. While the robot implemented with combined intonation did evoke more emotions in participants, we could not detect that such conversations were perceived as more conversationally fluent than conversations with the robot with no intonation. Moreover, we observed an increase in interruptions during conversations with the robot with combined intonation, and such a robot also evoked more negative emotions in the participants. An explanations could be that participants got more irritated with a robot implemented with combined intonation due to the failing turn-taking system, where the TRPs did not seem to be evident enough. For example, when the robot fell silent for a while, participants seemed not to know whether they could take the turn or whether the robot was still processing. Some participants noted after the experiment that they were disappointed, which may indicate that the robot's abilities did not meet their expectations. Social engagement was objectively better with combined intonation, shown by the increased number of conveyed emotions. Although the literature states that focus intonation adds emotion [31] and emotion evokes emotion

	Godspeed I		Eeriness (RAS)		Warmth		Competence	
	M	SD	M	SD	M	SD	M	SD
None	2.64	0.77	4.12	1.4	5.03	0.69	4.49	0.85
Focus	2.75	0.74	3.47	1.52	5.27	0.66	4.74	0.82
End	2.56	0.66	3.44	1.04	5.14	0.75	4.62	0.86
Combined	2.81	0.93	3.64	1.2	5.13	0.72	4.35	0.88

Table 4: Mean values and standard deviations for the items used to measure the Humanlike Perception of the robot.

[59], our results show that a robot speaking with combined intonation evokes greater engagement in the conversation in terms of expressed emotions.

Another unexpected outcome was that the robot with end-of-utterance intonation yielded more interruptions than the robot with no or focus intonation. Literature on conversation analysis in human-human interaction suggests that end-of-utterance intonation signals the conversation partner of the right time to switch turns [49]. However, our results show that this turn-taking signaling function of end-of-utterance intonation seems to work differently when people talk with a robot. We observed that the silences between turns were on average 900 milliseconds in our study, while in human-human interaction such pauses are commonly between 0 and 200 milliseconds [55]. The long pauses in interactions with robots may have deranged the natural turn-taking system as a social signaling function. The pauses we implemented in our set-up were the default interpunctions in Choregraphe, which we deem an appropriate approach given that our study serves as an initial, exploratory study. For future research aiming to increasing conversational naturalness, however, these default pauses should be adjusted to create shorter, more natural pauses to prompt the natural social signaling process of the turn-taking system. The absence of a significant difference in conversational naturalness between focus point and end-of-utterance intonation could be explained by persistent intonation in some cases (i.e., the software did not allow the intonation to be completely stripped away on some words where it was necessary, lowering the clear difference in intonation types at times). We recommend future research to not rely solely on the Acapela tags and interpunction, and should employ for example Praat [9].

A final remarkable finding was the participants' gaze behavior at the robot. Although we found no significant differences between intonation types in participants' social engagement, participants tended to look at the robot much longer (approximately 87% of the time) than people commonly would in human-human interaction (up to 60%, [41]). This finding may imply that people feel uncertain at some level while talking with the robot, and therefore prefer to keep an eye on the robot.

5.1 Limitations & Future Research

A simple conclusion may be that intonation in HRI works differently from what is commonly observed in human-human interaction, as has been suggested for computer-mediated communication [26]. Given that people tend to anthropomorphize robots to a greater extent than other interactive technologies, we still believed the common practices in human-human interaction would be a promising starting point to investigate robot speech. Clearly, additional research is needed to further unravel our initial findings. Considering that some alterations of our robot speech intonation did provide the expected results, we recommend future researchers to reconsider parts of our experimental set-up. For example, some participants stated during the debriefing that they felt they had to yell at the robot for it to understand them. Adjusting the gain of the microphone might solve this issue. Another limitation of our findings was the persistence in intonation when creating the condition without intonation; something that could not be adjusted properly given the system's limitations regarding its settings. Subsequently, the difference between focus and end-of-utterance intonation might not have been as clear as it ideally should have been, which in turn may have affected our results. To overcome the issues encountered in our set-up, an alternative approach for future research might be the Wizard-of-Oz-method; a method used in many HRI studies [7] such as in [29, 35]). This significantly reduces the misunderstandings during the interactions and, as a result, may alleviate frustrations experienced by participants when the robot answers incorrectly. Another direction for future research is to deploy a different robot, either a non-humanoid robot or a zoomorphic robot, to explore any effects of a robot's morphology on robot speech intonation and conversational naturalness. For example, redoing this study with the Pepper, the larger sibling of the NAO, could be interesting given that it uses different intonation tags (NUANCE instead of Acapela) which would make the implementation of persistent intonation more strait-forward. Finally, since intonation is language-bound [14, 31], our current results may apply only to the Dutch language (and perhaps other Germanic languages). Therefore, future research should replicate our study using other languages including their functions and forms of intonation to explore the effects of intonation of robot speech in these languages.

5.2 Conclusion

By investigating the effects of different type of intonation in robot speech, our results suggest that a robot should at least have some intonation when talking with a human being for people to perceived the conversation as natural. A robot without any type of intonation is perceived as less natural and hinders people to engages with it on a social level (as compared to a robot implemented with any type of intonation). Contrary to what is stated in the literature on intonation in human-human interaction [4, 31, 36, 49, 56], we could not confirm that combined intonation is the most natural or social way for a robot to talk. Additional research is therefore necessary to further investigate the extent to which psychological mechanisms in human-human interaction align with or deviate from the interactions we have with robots.

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*, 1 (1980), 44–54.

[4] Banse, R., and Scherer, K. R. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology 70*, 3 (1996), 614.

[5] 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.

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

[7] Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., and Belpaeme, T. From characterising three years of HRI to methodology and reporting recommendations. In *Eleventh ACM/IEEE International Conference on Human Robot Interaction* (2016), pp. 391–398.

[8] 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.

[9] Boersma, P., and Weenink, D. Praat: Doing phonetics by computer [computer program], 2019.

[10] Botinis, A., Granström, B., and Möbius, B. Developments and paradigms in intonation research. *Speech communication 33*, 4 (2001), 263–296.

[11] 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.

[12] 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.

[13] Castellano, G., Pereira, A., Leite, I., Paiva, A., and McOwan, P. W. Detecting user engagement with a robot companion using task and social interaction-based features. In *Proceedings of the 2009 International Conference on Multimodal Interfaces* (2009), ACM, pp. 119–126.

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

[15] 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.

[16] Clayman, S. E. *Turn–constructional units and the transition–relevance place.* 2012, pp. 151–166.

[17] 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.

[18] 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.

[19] 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.

[20] Eyssel, F., Kuchenbrandt, D., and Bobinger, S. Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (2011), ACM, pp. 61–68.

[21] Field, A. P. *Discovering statistics using SPSS.* SAGE, London, England, 2015.

[22] 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.

[23] Florent, G. *Acapela TTS for Windows, Mac and Linux – User's Guide*, 40 ed. Acapela Group, Mons, Belgium, June 2015.

[24] Fong, T., Nourbakhsh, I., and Dautenhahn, K. A survey of socially interactive robots. *Robotics and autonomous systems 42*, 3-4 (2003), 143–166.

[25] 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.

[26] 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.

[27] Ghosh, S., and Pherwani, J. Enabling naturalness and humanness in mobile voice assistants., 2015.

[28] Grice, M., and Baumann, S. *An introduction to intonation–functions and models.* 2007, pp. 25–52.

[29] 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.

[30] 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.

[31] Hirst, D., and Di Cristo, A. E. *Intonation systems: a survey of twenty languages.* Cambridge University Press, 1998.

[32] Ho, C. C., and MacDorman, K. F. Measuring the uncanny valley effect. *International Journal of Social Robotics 9*, 1 (2017), 129–139.

[33] 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.

[34] Igic, A., Watson, C., Teutenberg, J., Broadbent, E., Tamagawa, R., and MacDonald, B. Towards a flexible platform for voice accent and expression selection on a healthcare robot. In *Proceedings of the Australasian Language Technology Association Workshop 2009* (2009), pp. 109–113.

[35] Kim, Y., Kwak, S. S., and Kim, M. S. Am i acceptable to you? effect of a robot's verbal language forms on people's social distance from robots. *Computers in Human Behavior 29*, 3 (2013), 1091–1101.

[36] Krahmer, E., and Swerts, M. On the alleged existence of contrastive accents. *Speech Communication 34*, 4 (2001), 391–405.

[37] Leite, I., Castellano, G., Pereira, A., Martinho, C., and Paiva, A. Long-term interactions with empathic robots: Evaluating perceived support in children. In *International Conference on Social Robotics* (2012), Springer, pp. 298–307.

[38] Mavridis, N. A review of verbal and non–verbal human–robot interactive communication. *Robotics and Autonomous Systems 63* (2015), 22–35.

[39] Mazeland, H. J. *Inleiding in de conversatieanalyse.* Coutinho, 2003.

[40] Mihalicek, V., and Wilson, C. E. *Language Files: Materials for an Introduction to Language and Linguistics.* Columbus, OH: The Ohio State University Press, 2011.

[41] 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.

[42] 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.

[43] Mori, M., MacDorman, K. F., and Kageki, N. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine 19*, 2 (2012), 98–100.

[44] 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.

[45] 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.

[46] Rich, C., Ponsler, B., Holroyd, A., and Sidner, C. L. Recognizing engagement in human-robot interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2010), IEEE, pp. 375–382.

[47] Rodriguez, I., Martínez-Otzeta, J. M., Lazkano, E., and Ruiz, T. Adaptive emotional chatting behavior to increase the sociability of robots. In *International Conference on Social Robotics* (2017), Springer, pp. 666–675.

[48] Ronanki, S., Henter, G. E., Wu, Z., and King, S. A template-based approach for speech synthesis intonation generation using lstms. In *INTERSPEECH* (2016), pp. 2463–2467.

[49] 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.

[50] Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., and Joublin, F. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics 5*, 3 (2013), 313–323.

[51] 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.

[52] Schegloff, E. A. Turn organization: One intersection of grammar and interaction. *Studies in interactional sociolinguistics 13* (1996), 52–133.

[53] 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.

[54] 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.

[55] 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.

[56] '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.

[57] 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.

[58] Van Straten, C. L. Looks good, sounds nice: Intonation and bodily appearance in robot–mediated communicative treatment for children with autism., 2016.

[59] 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.