**Universiteit Leiden**
**Opleiding Informatica**

# The impact of the naturalness of a robot's voice on Human-Robot-Interaction

Name: Luit Verschuur
Date: 14/07/2020
1st supervisor: Dr. Joost Broekens

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

# The impact of the naturalness of a robot’s voice on Human-Robot-Interaction

**Abstract.** In recent years a lot of research has been done to explore possibilities to increase the human perception of a social robot. In 2008 research has been done about the impact of a humanized voice used by a robot. This research concluded that ‘approaches to the robot with synthesized voice were found to induce significantly larger approach distances’ (Walters et al, 2008). Nowadays, new conceptions about the preferability of humanlike robots have stimulated the discussion. Recent research suggested that more solid knowledge is needed about positive and negative consequences of humanization of robot voices (Giger et al, 2019). In this research we focus on the impact of naturalness in the voice of a robot, based on consequences on approach distance, the perception of the robot, the interaction quality and the task performance. Even though our voice manipulation was successful, we found no significant effects of robot voice manipulation. These results are in conflict with previous work and implicate that more research is needed in order to find the ‘best’ voice for a robot.

## 1 Introduction

In recent years a lot of research has been done on the impact of the voice of a social robot on humans. First of all in 2008: ‘Approaches to the robot with synthesized voice were found to induce significantly further approach distances’ (Walters et al, 2008). This was one of the first studies that showed that there is a possibility that a natural voice of a social robot is preferred over a synthesized voice. In addition, another research concluded that people interacting with a robot with a less robotic voice had more positive emotions during the interaction than people who interacted with a robot with a more robotic voice (Tamagawa et al, 2011). This research strengthened the possibility of a preference for a more natural voice within a robot. Furthermore, research about the perception of robots has been conducted by testing a reception robot. In this research Ana was a humanlike robot in appearance and in voice and Kobiana is a mechanical robot that does not look or sound like a human. This research concluded that, after comparing both on different parameters, Ana was preferred over Kobiana (Trovato et al, 2015). However not only the voice differed in this study. Also the appearance of the robots differed. A logical conclusion after combining these three articles would be that people prefer a more natural humanized voice over a mechanic synthesized voice in a robot. However, recent research questions these conclusions by focusing on different types of consequences (positive and negative) (Giger et al, 2019). Therefore in this research we investigate positive consequences but also the negative consequences of natural voices in robots. In addition, we do not question the outcomes of these studies but try to show more types of consequences that are relevant to consider within this field. Hence in this paper, the experiments of the articles above are replicated while adding more types of consequences (positive and negative).

### 1.1 Motivation and related work

Nowadays research in the field of human–robot interaction (HRI) is often based on the design goal of creating robots that act and think like humans. Over the last 20 years, social robots have become increasingly humanlike. A key assumption for developers is that humanlike social robots will improve HRI and facilitate their acceptance (Giger et al, 2019). In the case that this assumption is true, new research will keep focusing on improving robots to become more humanlike. To quote Robert (2017): 'one thing that seems to unite many scholars that study robots is the goal of one day creating a fully autonomous human-like robot capable of mimicking all human behaviors and emotions'. On the other hand, he does not agree with these scholars. Therefore in that same article he states that we should not aim for this goal. He claims that in some cases, when robots would become too humanlike, humans do not know how to appropriately use the robot anymore. Robert (2017) describes the situation where love and friendship will play a role within the relationship between a human and a robot. In this case the appropriate way to interact with the robot is questionable. In addition to the hesitation of Robert, Giger et al. ask themselves two questions. The first question is: If humanizing robots is a means to an end, then when and how is that end achieved? The second question is: Is partial humanization enough? (Giger et al, 2019). Answers to these valid questions are not yet provided but are very relevant for the optimal goal within this field. Nevertheless, studies about the development of new human features for robots are on the rise. Therefore these studies say that we should start by focusing on our knowledge of the different types of consequences (Giger et al, 2019). This implies that before we implement an increasing number of human features in robots, we should look more closely at their consequences. Therefore it is important that more different types of consequences are studied and considered thoroughly.

For this reason we will take a broad look to investigate the different types of consequences between voices. In the work of Walters et al. (2018) the approach distance is studied. The approach distance contains the value of the actual distance to the robot that is preferred by the participant. They concluded that the approach distance to the robot with more synthesized voices were significantly further. The outcomes of their study are valid, but the interest of this study is to consider more types of consequences. Therefore we will take the approach distance as a part of the measures that will be taken.

Furthermore, research is done on the number of positive emotions during the interaction between a robot and a human. This research concluded that people who were randomized to a less robotic voice had more positive emotions during the interaction than people randomized to a more robotic voice (Tamagawa et al, 2011). The added value of this research is about the positive feelings of the participant. It shows that user feeling and in addition the users' perception are interesting within this field. Furthermore, the research of Tamagawa et al. (2011) does not only focus on the level of naturalness in voices but also on the accent used by the robot. They used two accents where the less robotic voice was a New Zealand accent while the other more robotic voice was a United States accent. Therefore the outcomes of the study are not only based on differences between naturalness within the voice but is also based on the difference in accent. This means that it cannot be concluded that more positive emotions were only present because of the difference in how robotic the voice was. Furthermore, the study of Trovato et al. (2015) was also not only based on a difference in voice. In this study they concluded that a more humanlike robot in appearances combined with a more humanlike voice was preferred over a more mechanical robot. So, this research is not purely on the difference in the voice of the robot because also the physical appearance of the robot was manipulated.

To summarize, most of the conclusions drawn above are not purely based on a difference in voice. In addition, the articles above are only a small selection of all the studies that analyze the voices robots use. However, like said before, very little of these studies focus only on the differences between voices that the robot uses. The lack of research on this specific topic makes that there is little known about the direct impact of the naturalness of the voice in a robot. Therefore more and more precise work is needed to investigate the direct consequences on the interaction of implementing a different voice within a robot.

As mentioned before, studies about the development of new human features for robots are on the rise (Giger et al, 2019). For example, research about the voice pitch is done. In this research they concluded that 'The manipulation of voice pitch showed strong effects on how users perceived the robots and the entire interaction' (Niculescu et al, 2011). In this research only very natural voices were used. Outcomes like these are valid and can bring a lot of insights in the understanding of the users' perception of the robot. However, by experimenting with only natural voices the effects can be different or less accurate than if they also were compared with synthesized voices. Therefore, as long as there is no scientific evidence that more natural voices are always preferred over more synthesized voices, we should not base our research only on natural voices. Similar studies using only natural voices are done on a regular basis. To be sure these studies are interesting and valuable, more types of consequences of natural voices need to be analyzed.

Concluding, a lot of research on differences in voices is already done. However, most of these researches focus on more differences than just the voice. Therefore more research is needed to measure the direct consequences of a difference in voice. Measurable consequence values that already showed interesting outcomes within this field are: approach distance, user perception and user feelings.

# 2 Research question

In order to investigate a larger set of consequences for using different robot voices, there is the need to focus on different types of consequences. To be able to investigate them, these consequences need to be measurable. This means that relevant and measurable consequence types are needed. The experiment of Walters et al. (2008) about the approach distance of a human to a robot in HRI is a good experiment to replicate. Because we are interested in more types of consequences, more measures should be implemented. We decided to use the scale brought forth by Bartneck et al. (2009) to measure our users' perception in order to make the variable measurable. The users' perception is a reference to the users' experience and is separated in 5 different components: Anthropomorphism, Animacy, Likeability, Perceived Intelligence and Perceived Safety. In this experiment safety will play no role within the interaction and therefore only the first 4 components: Anthropomorphism, Animacy, Likeability and Perceived Intelligence will be investigated. Investigating the users' perception of a robot can help explain other results and furthermore give an interesting view on the consequences of the experience of a human. Another type of consequence is the quality of the interaction. The interaction quality refers to the customers' perception of the manner in which the service is delivered during service encounters (Lemke et al, 2011). Also, the interaction quality can be measured by focusing on the perception of the participant. However, in this instance it is performed with a different focus. When investigating the interaction quality we can focus on the four components Lemke et al. (2011) introduced to make our values measurable: Content Quality, Interaction Features, Tasks and User Feelings. Each of these components highlight another part of the previous users' perception. In order to rate the service delivered by the robot to the participant a collaborative task is done. Within this task we can measure the task performance at the same time. Besides users' perception, approach distance and interaction quality, the performance of the executed task is measured. The task performance can be defined as the effectiveness with which job incumbents perform activities (Borman & Motowidlo, 1993). A difference in task performance implicates more underlying consequences affecting the interaction between the participant and the robot. In this paper these types of consequences are combined and therefore determine the research question as follows: What is the impact of naturalness in the voice of a robot, based on consequences on approach distance, the perception of the robot, the interaction quality and the task performance?

## 2.1 Hypotheses

This research focuses on the impact of the naturalness of the voice used by a robot on four types of consequences. Previous research work often focused on one particular consequence value only, or used multiple different independent variables. In line with recent research trends this research looks at more different types of consequences. In order to do so this research combines previous research work and only focuses on one specific independent variable, namely the naturalness of the voice of the robot. This should give a wider view in the consequences a more natural voice has when used within a robot. The outcomes of previous work implicate that the outcomes of this research will be the same. However, as mentioned before, not all previous work focused only on differences in voice. Therefore these previous researches do not totally answer the question of what consequences a difference in the voice of robot has. However, considering that these outcomes are not totally based on the right parameters we can still expect that closely related researches have pretty similar outcomes to the outcomes of our own research.

### *2.1.1) Approach distance*

Walters et al. (2008) already concluded that the approach distance to robots with a synthesized voice were significantly bigger. Therefore it is also our expectation for the approach distance to decline when the voice gets more natural.

### *2.1.2) Users' perception*

Trovato et al. (2015) already concluded that a more humanlike robot was preferred over a more mechanical robot. This implies that in the users' perception of the robot would have increased. However in this research also the appearance of the robot was different. Therefore the outcomes of our results are expected to differ a little bit from the outcomes of Trovato et al. (2015). Because the naturalness of the voice still makes the robot more human, it is likely for the consequences for the users' perception to be more moderate. These will be more moderate because more characteristics are relevant in users' perception and this research only focuses on the voice of a robot. Therefore more variating outcomes in the different scales are expected.

### *2.1.3) Interaction quality*

The interaction quality will be measured. In this case we take the assumption of Giger et al. (2019) into account. They state that a key assumption for developers is that humanlike social robots will improve HRI and facilitate their acceptance. In addition, also Trovato et al. (2015) concludes that more humanized robots are preferred. Since the interaction quality refers to the customers' perception of the manner in which the service is delivered during service encounters (Lemke et al, 2011), we can expect that when the customers' perception increases also the interaction quality increases.

### *2.1.4) Task performance*

The task performances will be compared. This measurement monitors the effectiveness of the performed activities. The task performance is based on the collaboration between participant and the robot. In this field little is known about the influence voice has on the task performance. Therefore we can only base our knowledge on related work. All previous work about the impact of natural voices seems to increase quality of the interaction. The task performance is a measure for the interaction. Therefore we can expect that an improvement in the interaction will result in a better task performance of robots with a more natural voice.

To sum up, the hypotheses are based on previous outcomes of comparable research. A lot of this related work took more independent variables into account and therefore the expectation is that the outcomes of this research will be more moderate. Important to note is that because naturalness is often found to have a positive impact in our hypothesis, we expect more natural voices to improve positive consequences compared to more synthesized voices.

| | **Approach distance** | **Users' perception** | **Interaction quality** | **Task performance** |
|---|---|---|---|---|
| **Hypotheses** | The approach distance should decline when the voice gets more natural. | More natural voices are expected to be preferred. | When the customers' perception increases also the interaction quality increases. | An improvement in the interaction will result in a better task performance of robots with a more natural voice. |

*Table 1: Overview of the hypotheses of the variating consequence values.*

## 3 Method

The voices we use for this study differ significantly on the a synthesized/naturalness voice scale. We created multiple voices to test whether a specific voice was not just unpleasant or undesirable. In the case a voice is unpleasant or undesirable the general trend of the results will still show how the other voices relate to each other. Within the task it is important that the difference in voice is validated. Therefore participants are asked to confirm what they thought of the voice.

Further, we have four outcome measures. Firstly, the *approach distance* needs to be measured. To measure approach distance, the robot introduces itself with a short monologue. Then, the participant is asked to stand at a preferred distance from the robot. When the participant has chosen a comfortable distance it is measured. Secondly, we measure *interaction quality*. For this, the participant and the robot have an interaction. The interaction between robot and participant is the same for the different voices. In this way we ensure that the voice is the only difference between the separate conditions. In order to be able to rate the interaction quality a task is needed that can be evaluated. Therefore a task where the participant guides the robot to another point in the room is used. When the participant has finished this task the interaction stops. The participant is asked to fill in a questionnaire about the interaction quality and about the *perception of the robot*. Finally, the *task performance* is measured. This can be done by the same measures as the interaction quality. A small interaction leading to an end goal is needed. When this end goal is reached the researcher can check the task performance by measuring the time that the participant used to complete the task.

## 3.1 Materials

In this research multiple materials are needed to make the experiment possible. The following materials will be discussed: the Nao Robot, the Godspeed Questionnaire Series, the interaction quality questionnaire, the used voices, the scenario and the Python code.

### *3.1.1) NAO Robot*

The NAO Robot is a robot created by SoftBank Robotics. It is capable of movement, voice understanding, voice production and is based on a human in appearance. These capabilities make it possible for a human to interact with a NAO Robot. For these reasons it is a suitable robot to use in this research. The NAO robot can be used with help of the RIE cloud robotics platform. This platform enables programable codes to interact with the robot. This has been done by using the Interactive Robotics Cloud Robotics platform to control the robot. The physical appearance of the NAO Robot can be found in Figure 1.

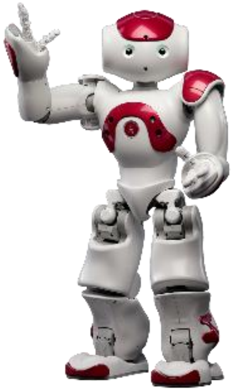

*Figure 1: The NAO Robot*
*(https://d1rkab7tlqy5f1.cloudfront.net/EWI/Actueel/Humans/nao_sayingvvvv.png)*

### *3.1.2) The Godspeed Questionnaire Series*

The Godspeed Questionnaire Series is a series of questionnaires to measure the users' perception of robots. This questionnaire is subdivided in five different questionnaires on the subjects: Anthropomorphism, Animacy, Likeability, Perceived Intelligence and Perceived Safety. In this research we do not use Perceived Safety. Since 2009 this questionnaire is a widespread used tool (Bartneck et al, 2009). By using this tool we can compare the results to the results of other studies. The questionnaire can be found in Appendix A.

### *3.1.3) Interaction quality questionnaire*

Interaction quality is not a widely spread measure in HRI. Therefore the baseline of this questionnaire is based on less than could be done in users' perception. In order to make the questionnaire as valuable as possible it is based on the questionnaire that Niculescu et al. (2011) used. They based their questions on how human-robot interaction should be evaluated by following the principles of Hassenzahl et al. (2003). The questionnaire they made focused on 4 different measures: content quality, interaction features, tasks and user feelings. The questionnaire can be found in Appendix B.

### *3.1.4) Used voices*

In this experiment four different voices for the robot are used. We use the standard Dutch NAO robot voice, this is provided by the text-to-speech engine Nuance. Besides this standard NAO voice this research uses three other voices that were played as audio files through the NAO's speakers. First, the most natural voice was created by recording the voice of a real human being. With the help of an audio recorder we made new audio files. Second, the most synthesized voice was created, by sending the newly recorded audio files into voicechanger.io to create a synthesized robot voice. Third, the voice between the most natural voice and the standard NAO robot voice was created. This was done by using another text to speech generator from ttsmp3.com. The four voices thus ranged from natural to unnatural in the following order: human, humanlike, static, synthesized robot voice.

### *3.1.4) Scenario*

When a participant enters the room the experiment is started. First of all, they are able to see the setting of the experiment. Then they are asked to fill in an informed consent about their participation in the experiment. When this is done the robot can start with its monologue introducing itself and its voice. The actual text said by the robot can be found in Appendix C. Here only the flow of the scenario will be described.

After the introducing monologue the participant is familiar with the robots voice. The monologue ends with the robot asking the participant to stand at a distance to the robot that has his/her preference. The distance is measured by the researcher. When the researcher has finished measuring, the task continues. The next part of the experiment consists of the robot introducing the collaborative task. When the task is completed the participant is asked to fill in the questionnaire. When the questionnaire is filled in the experiments ends.

### *3.1.5) Task*

The collaborative task is that the participant has to guide the robot to another point in the room. This place is marked with a giant cross and will therefore be referred to as point X. The guiding of the robot by the participant gives the participant a proactive function within this interaction. Furthermore the task is collaborative and interactive. By making the task a collaborative and interactive interaction, where the participant has a leading role, the participant is tested on its estimation of the capabilities of the robot. The participant can lead the robot through the room by giving the robot spoken instructions. The robot always starts at the same place in the room. This point will be referred to as point Y (toes against the undermost line in Figure 2). In the same room is the point X (the big cross in the right top corner in Figure 2) where the robot has to go to. The robot listens to the 4 words: ‘Lopen’, ‘Stop’, ‘Links’ and ‘Rechts’ which respectively mean ‘Walk’, ‘Stop’, ‘Left’ and ‘Right’. With these four words the participant is able to guide the robot to point X. In order to make the interaction more interactive and alive the robot reacts to the instructions given to him. Therefore when the robot is turning or starting to walk it has a chance to say something before he starts to act. The chance of the robot saying something before he starts is 50% for the words: ‘Walk’, ‘Left’ and ‘Right’. Within this 50% chance the robot has two options to say. The chance for any of these 2 sentences is just as big. The sentences that are said can be found in Appendix D.

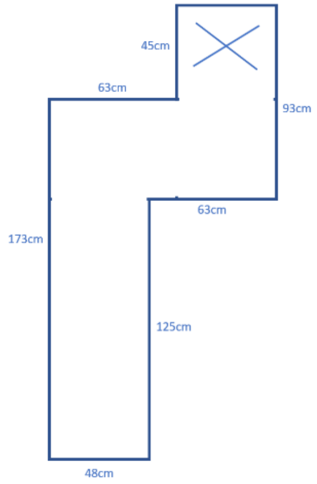

*Figure 2: The dimensions of the track.*

### *3.1.6) Python code*

The last material used was the python code to run the code of the experiment. Python contains libraries that make it possible to connect and run codes on the connected robot. The experiment consisted of two different codes. The first code consisted of the code up to the moment where the approach distance had to be measured. The second code consisted of the code after the measuring was done. This difference had to be made because there is no specific time it takes to measure the distance. Furthermore the hearing can be sensitive to other words that lookalike. All other options were excluded because they would be needed later. Furthermore the splitting of the code was a safe and effective way to structure the experiment.

After the last sentence, of the explanation of the collaborative task, is said. The code falls back into a keyword stream. This stream focuses on the words 'Walk, 'Stop', 'Left' and 'Right'. When the robot hears something it will respond to the word that is most likely to be the right word. A note here is that when the robot talks, the hearing stream can detect these keywords. Therefore we need to avoid the robots to say keywords. This means that all the words the robot says cannot be 'Walk', 'Stop', 'Left' and 'Right'.

### 3.2 Experimental setup / approach

Due to a corona pandemic, governmental restrictions were set to minimize physical contact. This impacted our research. The implications for this research was mainly the recruitment of participants and the experimental setup. Participants were selected from two different student houses. This proved to be the best way to minimize differences in experimental setup and maximize the number of participants. These two student houses were able to deliver 40 different participants. They form a 50/50 gender distribution and ages differ between 17 and 33 years old. Furthermore these participants were randomly assigned to one voice condition.

Within 2 student houses we were able to use sleeping rooms that had enough space to do the experiment. In Figure 2 we can see the space that was needed in a room. To not influence the preferred approach distance both rooms had 230 centimeter space between the robot and the end of the room. This made it fit within the room but also made the environment small and personal for the participant. Personal rooms were needed to minimize outside input to the experiment. Furthermore, an individual experiment could not be of influence to people outside because most of them would participate in the experiment later.

To control for confounding variables, every participant will do the experiment as similar as possible, the setup will be the same, the interaction will be the same, the volume of the voice cannot differ and the gender and age ratio should be as random as possible. Furthermore to avoid confounding variables a lot of choices were made when choosing only two different student houses and therefore only two different rooms. Furthermore by choosing sleeping rooms instead of the living room to minimalize the effects of external factors.

The experiment setup can be seen in Figure 3.

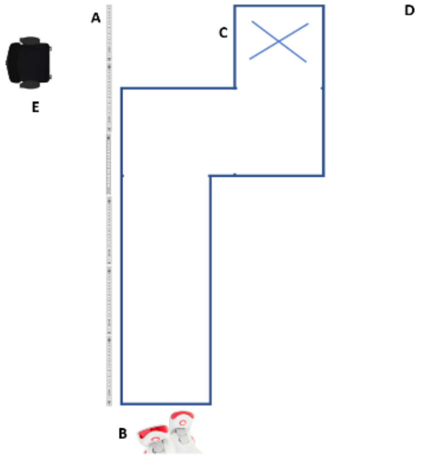

*Figure 3: Setup of the experiment. Segment A was measuring tape, segment B was where the NAO robot was standing, segment C the track the NAO robot had to walk, segment D the place where participants entered the room and segment E is where the researcher supervised the experiment.*

### 3.3 Measures

As an objective measure for distance to the robot, the approach distance is measured. This is done by recording the distance from the participant to the robot in centimeters after the introduction monologue. The researcher recorded this by using measuring tape from the robot to the participant. Secondly, users' perception of the robot is measured. This is done using The Godspeed Questionnaire Series. This questionnaire asked questions about the subjective feeling you get from the interaction with a robot. This questionnaire is divided in four sub questionnaires about: Anthropomorphism, Animacy, Likeability and Perceived Intelligence. Thirdly, we measure interaction quality. The interaction quality is measured by using a questionnaire based on the work of Niculescu et al. (2011). This questionnaire divides the interaction quality into the four components: Content Quality, Interaction Features, Tasks and User Feelings. Both questionnaires ask the participant to rate their agreement with a statement on a 1-5 scale. Lastly the task performance is measured. This is done by measuring the time that is used to complete the task. This time is rounded to seconds and is measures by the researcher that starts and stops the stopwatch when respectively the task begins and ends.

| | **Approach distance** | **Users' perception** | **Interaction quality** | **Task performance** |
|---|---|---|---|---|
| **Method** | Robots holds a monologue to introduce its voice. Afterwards it asks the participant to stand at a preferred distance. | After listening to a monologue and completing a collaborative task the participant needs to fill in a questionnaire with questions about the users' perception. | After listening to a monologue and completing a collaborative task the participant needs to fill in a questionnaire with questions about the interaction quality. | After completing a collaborative task the researcher notates the time it took for the participant to complete the task. |

*Table 2: Overview of the methods for the variating consequence values.*

## 4 Results

The majority of the participants (87,5%) were between the ages of 20 and 25. 2,5% was older, 26-33 years old, and 10% was younger, 17-19 years old. Exactly 50% of the participants was male and therefore also 50% was female. Furthermore 87,5% of the participants had no experience working with similar robots, 12,5% had some experience working with similar robots. No participant worked with similar robots on a regular basis.

Within this section the voices are referred to with numbers. The four different voice can be separated in voice 0, 1, 2 and 3. In this case voice 0 represents the most synthesized voice and voice 3 represents the most human voice. Voice 1 and 2 are respectively to these voices on the synthesized/natural voice scale.

In order test our hypotheses, we first performed a one-way ANOVA with 4 conditions. Our overall analysis (including voice perception on intelligibility, pleasantness, humanity) didn't show a significant overall effect of voice on the combined multivariate outcomes ($F$(33,84) = 1.448, $p$ = .093).

The majority of the effect, as the results in the following sections will show, is due to the effect of the perception of the voice.

*4.1) Voice difference*

Firstly, as mentioned before, in order to make sure there are differences between the voices we used, we asked the participants to rate the voice on a scale on intelligibility, pleasantness, humanity. These measures were combined in one value. This voice perception value was tested. The outcome of this ANOVA shows that there is a significant difference found between voices ($F$(3,36) = 7.183, $p$ = .001 ). The mean rates of these voices are respectively 2,90 (SD = .77), 3,97 (SD = .60), 3,70 (SD = .66) and 4,37 (SD = .68). An increasing line can be found when the naturalness of the voice increases. However still notable is that voice 1 scores better than voice 2. However, this difference is not significant.

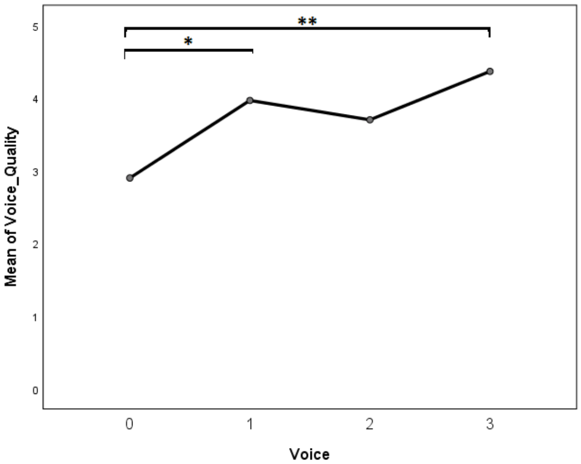

*Figure 4: Differences in voice quality*
*(*. The difference between the mean values is significant at the 0.05 level. **. The difference between the mean values is significant at the 0.001 level, according to post hoc T-tests?)*

### *4.2) Approach distance*

Secondly, the experiment consisted of an approach distance section where differences in voice quality should create differences in the approach distance. The outcomes of this research show that there is no significant difference found between the different voices (*F*(3,36) = 0.326, *p* = .806). The following mean values were found: 138,80 cm for voice 0 (SD = 46,074), 127,10 cm for voice 1 (SD = 61,276), 150,70 cm for voice 2 (SD = 48,491) and 136,50 cm for voice 3 (SD = 59,163). Within these mean values no trend can be found. However, we can see that the originally used NAO voice of the robot found the lowest average approach distance.

### *4.3) Users' perception*

Thirdly, the users' perception was measured. This was done by using the Godspeed questionnaire. This questionnaire subdivides as mentioned the users' perception in 4 different measuring values: Anthropomorphism, Animacy, Likeability, Perceived intelligence. The outcomes of these different values will be presented separately.

First of all, the outcomes for anthropomorphism do not differ significantly between voices (*F*(3,36) = 0.193, *p* = .900). The different mean values found are respectively 2,50 (SD = .67), 2,52 (SD = .59), 2,52 (SD = .98) and 2,32 (SD = .46). These mean values, that represent the anthropomorphism score, do not differ much and high standard deviations lead to little differences between voices. Noticeable is that the only deviating value is the most natural voice and the effect is negative.

Secondly, the outcomes for animacy show not to be significantly different between the voices (*F*(3,36) = 0.958, *p* = .423). The mean values, that represent the animacy score, that we found are respectively 3,02 (SD = .92), 3,20 (SD = .53), 2,86 (SD = .76) and 2,68 (SD = .60). A small downward trend can be noticed however these differences are not significant. Also in this case the originally used NAO voice of the robot contains the highest value.

Thirdly, the outcomes for the likeability show not to be significantly different (*F*(3,36) = 0.356, p = .785). The mean values, that represent the likeability score, that provide this result are respectively 4,14 (SD = .53), 3,90 (SD = .57), 3,94 (SD = .61) and 4,04 (SD = .56). These outcomes show that the different values are very similar to each other with no noticeable trend. Also in this case the mean values do not differ much and high standard deviations lead to non-significant differences between voices.

Lastly, the outcomes for the perceived intelligence are compared. These outcomes show not to be significantly different between voices (*F*(3,36) = 0.196, *p* = .898). The mean values, that represent the perceived intelligence score, found here are respectively 3,18 (SD = .95), 3,32 (SD = .32), 3,12 (SD = .96) and 3,34 (SD = .63). Also in this case the standard deviations are bigger than the differences between the voices and therefore the results are not significantly different. Furthermore no noticeable or remarkable trend can be found within the values that are found.

### *4.4) Interaction quality*

Fourthly, the interaction quality was measured. In this case the questionnaire is based on the questionnaire that Niculescu, van Dijk and Nijholt (2011) used in their research for differences in interaction quality. This questionnaire focused on differences in content quality, interaction features, tasks, user feelings. The results of these different measures will be described separately.

First of all, the outcomes of the content quality is not significant between the different voices (*F*(3,36) = 1.143, *p* = .345). The outcomes are respectively 3,80 (SD = .84), 3,85 (SD = .71), 4,18 (SD = .41) and 3,65 (SD = .58). In this case the mean value, that represents the perceived content quality, seems to increase when the voice gets more natural. However, the most natural voice has the lowest mean value and therefore shows the opposite results.

Secondly, the outcomes for interaction features show not to be significantly different between voices (*F*(3,36) = 0.625, *p* = .603). The mean values, that represent the interaction feature score, are respectively 3,23 (SD = .55), 3,40 (SD = .45), 3,50 (SD = .50) and 3,28 (SD = .40). These outcomes show the same trend as the content quality. Where a trend seems to be increasing when the voice gets more natural. However the most natural voice shows opposite results. Furthermore in this case the differences are smaller.

Thirdly, the outcomes for the tasks are not significantly different between voices (*F*(3,36) = 1.008, *p* = .400). The mean values, that represent the task score, are respectively 4,43 (SD = .69), 4,63 (SD = .37), 4,80 (SD = .23) and 4,63 (SD = .48). Also in this case the results are comparable to the two features measured before. The values seem to increase however the most natural voice has opposite results.

Lastly, the outcomes for the user feelings are not significantly different between voices (*F*(3,36) = 0.787, *p* = .509). The mean values, that represent the user feeling score, are respectively 4.18 (SD = .33), 3,96 (SD = .38), 3,96 (SD = .53), and 3,93 (SD = .42). These outcomes are very divergent form the previous results and show contrasting results. However standard deviations are still too high to find any significantly difference between voices.

*4.5) Task performance*

Lastly, the task performance was measured. This was measured by measuring the time that is used to complete the task. The outcomes of the task performance show not to be significantly different between voices (*F*(3,32) = 0.657, *p* = .585). The mean values are respectively 215,22s (SD = 31.38), 208,78s (SD = 41.11), 204,00s (SD = 31.73) and 225,78s (SD = 34.39). Compared to interaction quality features these results are similar. In this case lower mean values are better because the used time to complete the task is lower. In this case the mean values, in time used to fulfill the task, seem to decrease when the voice gets more natural. However the most natural voice shows opposite results.

| | **Approach distance** | **Users' perception** | **Interaction quality** | **Task Performance** |
|---|---|---|---|---|
| **Voice 0** | 138,80 cm<br>(SD = 46,074) | 2,50 (SD = .67)<br>3,02 (SD = .92)<br>4,14 (SD = .53)<br>3,18 (SD = .95) | 4.18 (SD = .33)<br>4,43 (SD = .69)<br>3,23 (SD = .55)<br>3,80 (SD = .84) | 215,22s<br>(SD = 31.38) |
| **Voice 1** | 127,10 cm<br>(SD = 61,276) | 2,52 (SD = .59)<br>3,20 (SD = .53)<br>3,90 (SD = .57)<br>3,32 (SD = .32) | 3,96 (SD = .38)<br>4,63 (SD = .37)<br>3,40 (SD = .45)<br>3,85 (SD = .71) | 208,78s<br>(SD = 41.11) |
| **Voice 2** | 150,70 cm<br>(SD = 48,491) | 2,52 (SD = .98)<br>2,86 (SD = .76)<br>3,94 (SD = .61)<br>3,12 (SD = .96) | 3,96 (SD = .53)<br>4,80 (SD = .23)<br>3,50 (SD = .50)<br>4,18 (SD = .41) | 204,00s<br>(SD = 31.73) |
| **Voice 3** | 136,50 cm<br>(SD = 59,163) | 2,32 (SD = .46)<br>2,68 (SD = .60)<br>4,04 (SD = .56)<br>3,34 (SD = .63) | 3,93 (SD = .42)<br>4,63 (SD = .48)<br>3,28 (SD = .40)<br>3,65 (SD = .58) | 225,78s<br>(SD = 34.39) |

*Table 3: Overview of the results*
*Users' perception is separated in respectively: Anthropomorphism, Animacy, Likeability, Perceived intelligence.*
*Interaction quality is separated in respectively: content quality, interaction features, tasks, user feeling.*

### 4.6 Discussion

The main focus of this study was to find the impact of naturalness in the voice of a robot, based on consequences on approach distance, the perception of the robot, the interaction quality and the task performance. In this section we discuss the meaning of these outcomes.

First of all, the differences between the quality of the voices are measured. This section was included to check whether the intended differences within the voices would be noted by the participants. We found a significantly difference in the naturalness of the voices used for the collaborative interaction. Between voice 1 and 2 there is a minor decrease, however the outcomes showed that the difference between these two voices is not significant and therefore the trend still shows an increase in quality of the voice for more natural voices. This implicates that there are significant differences between the voices used and in the way as we expected.

Secondly, the differences between the approach distances are measured. The outcomes of this measure showed no significant difference between voices. This is in contrast with the work of Walters et al. (2008). Therefore the outcome of this measure is remarkable. To explain the differences between the outcomes of the two experiments the differences between the experiments are shown. The number of participants in this research was smaller. This implicates that the differences within groups is hard to be significant. Another difference between the two experiments is that other voices were used. Therefore it is possible that the voices used by Walters et al. (2008) would have implicated the results between their voices but do not implicate the direct differences between voices on a synthesized/naturalness scale. To explain this, a lot of voices that tend to natural voices also have other characteristics like: friendly/mean/man/woman/low/high. Therefore there can be differences between natural voices. Furthermore it is unlikely that the differences between the voices of Walters et al. (2008) were the same level as in this experiment. In any case, we could not exactly replicate their research. This means that there is need for more research to find out what affects the approach distance.

Thirdly, we found no differences between the users' perceptions. The highest mean values and lowest mean values fluctuate between the different measures and because the standard deviation is too high we cannot say much about differences between the voices. Interpreting these outcomes we can state that no significant difference is found due to a low number of participants. Unfortunately this may be true. However the fluctuating highest and lowest mean values also implicate that the outcomes may be more complicated than straightforward differences in preferences between the voices. In the results we find that the most natural voice scores the highest value for one measure and the lowest value for another. This could implicate that naturalness in voice can have conflicting consequences. Therefore more research is needed to be able to state clear differences, in consequences for users' perception, between the different voices.

Fourthly, we found no differences between interaction quality. The reason for the outcomes not to be significant can come from the small number of participants, variating characteristics that impact the interaction quality or just the fact that the differences between voices is not big. However the results that are found show interesting outcomes. The fact that the most natural voice scores lower is remarkable. A possible explanation for this difference is that the participant starts to overestimate the qualities of the robot. When the participant overestimates the robot it's quality he/she is less likely to pronounce better or speak louder. When this happens the interaction quality decreases. No significant result however still implicates that there are no major differences or that more research is needed to accept or reject this assumption.

Lastly, we found no significant differences between task performances. It is noticeable that the most natural voice scores the worst in time used to complete the task. Also in this case overestimation by the participant might play a role. When the interaction quality decreases the time to fulfill the task increases. Furthermore the number of participants and other complications that played a role within the task would have had a great impact on the final results in this measure. Therefore also here more research is needed to find the direct implications for the task performance.

Evaluating these outcomes, we have to take into account that participants does not only rate the robot by its voice. The appearance and qualities of the robot are also part of the evaluation of the participant. Furthermore, the voice is a subconscious variable that may or may not have a big influence on the way the participant experiences the robot. Therefore when the voice variable doesn't have a big influence the results could be limited. Furthermore the number of participants that could participate in this experiment was limited. Therefore the mean values are less accurate and therefore it is harder to find significant differences. Lastly the quality and fluency of the executed task has an impact on how the robot is perceived. Assuming that the varying quality and fluency of the task is normalized by the number of participants this should not have to much impact on the outcomes. However when the number of participants was not high enough, to normalize the variation in quality and fluency of the robot, it could have impact on the quality of our outcomes. For example the coincidence that the robot listened better to the instructions of the participant when it had one voice due to a more clear voice of the participant. Considering these counter arguments we can discuss the value of our outcomes.

Overall we found a significant difference between the different voices we used. However the different measured consequences did not show any significant difference. This would implicate that a more natural voice is not always preferred over a less natural voice. However, the small number of participants makes the outcomes less reliable. Furthermore other characteristics of the robot may play a big role, in the experience of the participant, and therefore a difference in voice cannot implicate big differences in consequences. In addition, differences in the quality and fluency of the robot within the task affects the experience of the participant. For all these reasons it is acceptable that very few and small differences in consequences were found. However small differences and interesting trends make research in this field attractive for further research.

# 5 Conclusion

The main focus of this study was to find and state the differences in consequences between the use of different robot voices on a synthesized/naturalness scale. Doing so we had to answer the following research question: What is the impact of naturalness in the voice of a robot, based on consequences on approach distance, the perception of the robot, the interaction quality and the task performance? Studying our test results we find no significant difference between the voices in the fields of approach distance, users' perception, interaction quality and task performance.

Explanations of these insignificant results can be that a low number of participants has influence on a lower likeliness to find significant results. Furthermore voice naturalness is not the only characteristic of the robot. Therefore we could not expect a lot of significant values. However finding no significant values shows the other side of the research. When no significance is found between the voices we cannot conclude that a more natural voice is in any way 'better' than a more synthesized voice.

Furthermore the results of this research show that it is hard to replicate earlier research work. The hypotheses based on earlier research work were that approach distance shows a significant difference and the users' perception shows an increase in rated value when voices got more natural. However our results do not confirm these hypotheses.

Lastly, interesting trends between mean values are found. These trends have no significant difference yet. However, the exploratory outcomes that we found could, with a little more investigation, add value within this research field.

## 5.1 Further research

By finding no significantly different consequences between more natural voices we can conclude that naturalness in voices is not preferred over synthesized voices. However this research had limitations that moderate this conclusion. Therefore more research is needed to be sure naturalness is voices is not preferred over a synthesized voice. Research fields that show immediate interest are within approach distance and users' perception. These fields showed contrasting results compared to previous researches.

Furthermore the increasing trend often shows negative outcomes when a voice gets too natural. Reasons for this can be overestimating the capabilities of the robot or just negative reflexes when a voice gets to natural. Other results are again contrasting to these and show a minor increase in all more natural voices.

To sum up, this research shows contrasting outcomes to previous work and is contradicting to the design goal of creating robots that act and think like humans. This means that we are not done with research on the effects of natural versus synthesized voices. Often too minor differences are found to state that one is 'better' than the other. However small trends can be detected that are interesting to investigate in more dept.

# References

1. Niculescu, A., van Dijk, B., Nijholt, A. et al. (2013) Making Social Robots More Attractive: The Effects of Voice Pitch, Humor and Empathy. *Int J of Soc Robotics* 5, 171–191, DOI: 10.1007/s12369-012-0171-x
2. Robert, L. P. (2017) The Growing Problem of Humanizing Robots, *International Robotics & Automation Journal*, 3(1), DOI: 10.15406/iratj.2017.03.00043.
3. M. L. Walters, D. S. Syrdal, K. L. Koay, K. et al. (2008) Human approach distances to a mechanical-looking robot with different robot voice styles. *The 17th IEEE International Symposium on Robot and Human Interactive Communication,* 707-712, DOI: 10.1109/ROMAN.2008.4600750.
4. Trovato, G., Ramos, J.G., Azevedo, H. et al. (2015) Designing a receptionist robot: Effect of voice and appearance on anthropomorphism, *24th IEEE International Symposium on Robot and Human Interactive Communication* , 235-240, DOI: 10.1109/ROMAN.2015.7333573.
5. Giger, J.C., Alves- Oliveira, N.P.P., Arriaga, R.O.P., et al. (2019) Humanization of robots: Is it really such a good idea*? Human Behavior and Emerging Technologies*, 1(2), 111-123, DOI: 10.1002/hbe2.147
6. Tamagawa, R., Watson, C.I., Kuo, I.H. et al. (2011) The Effects of Synthesized Voice Accents on User Perceptions of Robots. *Int J of Soc Robotics* 3, 253–262, DOI: 10.1007/s12369-011-0100-4
7. Shen, J., Pang, R., Weiss, R.J. et al. (2018) Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 4779-4783, DOI: 10.1109/ICASSP.2018.8461368.
8. Bartneck, C., Croft, E., Kulic, D. et al (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1), 71-81. DOI: 10.1007/s12369-008-0001-3
9. Hassenzahl M., Burmester M., Koller F. (2003) AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. *In: Szwillus G., Ziegler J. (eds) Mensch & Computer 2003*, 57, 187-196, DOI: 10.1007/978-3-322-80058-9_19
10. Niculescu, A., van Dijk, B., Nijholt, A. et al. (2011) The influence of voice pitch on the evaluation of a social robot receptionist, *2011 International Conference on User Science and Engineering (i-USEr ),* 18-23, DOI: 10.1109/iUSEr.2011.6150529.
11. Lemke, F., Clark, M. and Wilsom, J. (2011) Customer experience quality: an exploration in business and consumer contexts using repertory grid technique, *Journal of the Academy of Marketing Science*, 39, 846- 869, DOI: 10.1007/s11747-010-0219-0
12. Borman, W.C., & Motowidlo, S.J. (1993) Task Performance and Contextual Performance: The Meaning for Personnel Selection Research. *Human performance* 10, 99-109. DOI: 10.1207/s15327043hup1002_3

## Appendix A: The Godspeed Questionnaire (in Dutch)

Beoordeel de interactie met de robot aan de hand van onderstaande schalen:

| | | |
|---|---|---|
| Onecht | 1 2 3 4 5 | Natuurlijk |
| Lijkend op een machine | 1 2 3 4 5 | Lijkend op een mens |
| Onbewust | 1 2 3 4 5 | Heeft een bewustzijn |
| Kunstmatig | 1 2 3 4 5 | Levensecht |
| Houterige bewegingen | 1 2 3 4 5 | Vloeiende bewegingen |
| Dood | 1 2 3 4 5 | Levend |
| Stilstaand | 1 2 3 4 5 | Levendig |
| Mechanisch | 1 2 3 4 5 | Organisch |
| Kunstmatig | 1 2 3 4 5 | Levensecht |
| Passief | 1 2 3 4 5 | Interactief |
| Apatisch | 1 2 3 4 5 | Responsief |
| Afkeer | 1 2 3 4 5 | Geliefd |
| Onvriendelijk | 1 2 3 4 5 | Vriendelijk |
| Niet lief | 1 2 3 4 5 | Lief |
| Onplezierig | 1 2 3 4 5 | Plezierig |
| Afschuwelijk | 1 2 3 4 5 | Mooi |
| Onbekwaam | 1 2 3 4 5 | Bekwaam |
| Onwetend | 1 2 3 4 5 | Veel wetend |
| Onverantwoordelijk | 1 2 3 4 5 | Verantwoordelijk |
| Onintelligent | 1 2 3 4 5 | Intelligent |
| Dwaas | 1 2 3 4 5 | Gevoelig |

## Appendix B: Interaction quality questionnaire (in Dutch)

Beoordeel de interactie met de robot aan de hand van onderstaande schalen:

De kwaliteit van de content is:

| | | |
|---|---|---|
| Ongeloofwaardig | 1 2 3 4 5 | Geloofwaardig |
| Inhoudsloos | 1 2 3 4 5 | Informatief |
| Irrelevant | 1 2 3 4 5 | Relevantie |
| Onduidelijk | 1 2 3 4 5 | Duidelijk |

De robot was:

| | | |
|---|---|---|
| Slechte spreker | 1 2 3 4 5 | Goede spreker |
| Slechte luisteraar | 1 2 3 4 5 | Goede Luisteraar |
| Gesloten | 1 2 3 4 5 | Transparant |
| Ingewikkeld | 1 2 3 4 5 | Makkelijk |
| Traag | 1 2 3 4 5 | Snel |
| Statisch | 1 2 3 4 5 | Flexibel |

De taken waren:

| | | |
|---|---|---|
| Apart | 1 2 3 4 5 | Gewoon |
| Onbegrijpelijk | 1 2 3 4 5 | Begrijpelijk |
| Ingewikkeld | 1 2 3 4 5 | Simpel |

Jouw gevoel bij dit experiment was:

| | | |
|---|---|---|
| Ongemotiveerd | 1 2 3 4 5 | Gemotiveerd |
| Bedroeft | 1 2 3 4 5 | Verheugd |
| Druk | 1 2 3 4 5 | Kalm |
| Onzeker | 1 2 3 4 5 | Zelfverzekerd |
| Oncomfortabel | 1 2 3 4 5 | Comfortabel |
| Onvoldaan | 1 2 3 4 5 | Voldaan |
| Chaotisch | 1 2 3 4 5 | In controle |

## Appendix C: Text said in the scenario of the experiment (in Dutch)

**First monologue introducing the robot:**

*"Welkom bij dit experiment. Tijdens dit experiment zal er gekeken worden naar hoe jij mij ervaart. Als je dit niet prettig vind of om een andere reden niet mee wilt doen met dit onderzoek kan je dat nu zeggen."*

Robot is waiting for an answer for 2 seconds..

*"Oké, fijn dat je mee wilt doen met dit onderzoek. Het onderzoek bestaat uit twee verschillende delen. Het eerste deel bestaat uit een simpele vraag waar wij graag antwoord op willen. Het tweede deel werkt iets anders. Hierbij gaan wij samen een opdracht uitvoeren. Bij deze opdracht zal jij mij naar het aangegeven kruis in de kamer moeten begeleiden. Maar voor we dit gaan doen wil ik je eerst een vraag stellen. Kan je op een prettige afstand van mij gaan staan?"*

**Second monologue thanking the participant and explaining the collaborative task:**

*"Bedankt, dit was het eerste deel van het experiment. We gaan nu verder met het tweede deel. Dit deel bestaat uit de opdracht voor jou om mij te laten begeleiden naar het kruis in de kamer. Dit kan je doen door 'links' of 'rechts' te zeggen. In dat geval zal ik de gewenste kant op draaien. Wanneer je 'lopen' zegt zal ik rechtdoor blijven lopen totdat je 'stop' zegt. Is dit duidelijk?"*

Robot is waiting for an answer 2 seconds..

*"Oké, dan kunnen we nu beginnen met het tweede deel van het onderzoek."*

## Appendix D: Text robot says as a response in the task (in Dutch)

**Reaction when 'walk' is said:**

*"Zo, ik ga maar weer eens aan de wandel"*

Or

*"Prima, Ik kreeg al zin om weer te wandelen"*

**Reaction when 'right' is said:**

*"Pas op, ik ga nu de linkerkant op draaien"*

Or

*"Zo, ik ga nu even linksom draaien"*

**Reaction when 'left' is said:**

*"Pas op, ik ga nu de rechterkant op draaien"*

Or

*"Zo, ik ga nu even rechtsom draaien"*