university of groningen / faculty of science and engineering

Using a NAO robot during speech audiometry for hearing-impaired children

Developing a video protocol to detect non-verbal cues during child-robot interactions

Author
Karina Tímea Cozma
S4225368

Master's Thesis
Computational Cognitive Science
Department of Artificial Intelligence
University of Groningen,
The Netherlands

31.08.2022

Internal supervisor
Dr. F. Cnossen
Department of Artificial Intelligence
University of Groningen,
The Netherlands

External supervisors
Dr. G. Araiza-Illan
Department of Otorhinolaryngology
University Medical Center Groningen, The Netherlands

Prof. Dr. ir. D. Baskent
Department of Otorhinolaryngology
University Medical Center Groningen, The Netherlands

Abstract

Engagement is a key factor in audiometry and children with hearing loss undergo regular auditory tests to determine their speech perception abilities and have their hearing devices adjusted. Incorporating the humanoid robot NAO as a support during audiometry could result in an engaging experience for the child. However, in order to determine whether NAO is able to provide a positive experience, children's engagement and acceptance towards the robot must be measured during the auditory sessions.

The goal of this exploratory study was to develop a video protocol of non-verbal cues and evaluate engagement levels of children with hearing loss during a NAO supported speech audiometry. Furthermore, it aimed to explore the non-verbal cues that indicate engagement and no engagement towards the robot, changes in non-verbal behaviors with time and examine these cues in children with lower maximum speech perception levels.

The findings suggest that children display a wide range of non-verbal behaviors during NAO-supported auditory tests including but not limited to “leaning towards the robot”, “actively moving” or “amused”. Furthermore, outcomes of case studies indicate that children who had lower maximal speech perception levels expressed “amused” behavior frequently and for relatively longer periods of time.

Keywords: human-robot interaction, video protocol, non-verbal cues, engagement, speech audiometry

Acknowledgements

My gratitude goes out to everyone who helped me along the way as I wrote this thesis. First of all, I would like to thank Fokie Cnossen for her advice, encouragement, and excellent guidance. In addition, I want to express my gratitude to Gloria Araiza-Illan and Deniz Başkent for providing me the opportunity to work with them at the UMCG and for their ongoing input and valuable insights throughout the course of this project. I also would like to thank them for inviting me to accompany them to De Petteflet and experience how NAO operates in a school setting. Moreover, I would like to thank Luke Meyer and Marleen Schippers for their detailed feedback and inspiring tips. Additionally, I would like to express my sincere thanks to my friends and fellow students for their suggestions and advice. Finally, I would like to thank my partner, Mark, for being my source of motivation and inspiration and my family for always supporting me.

Contents

1 Introduction

1.1 Human-robot interaction

Human-robot interaction (HRI) is an interdisciplinary research field that aims to facilitate the communication, coexistence, and mutual adaptation between robots and humans (Salvini et al., 2011). HRI has a long history that dates back to the 1940s and has grown immensely ever since (Sheridan, 1997). The first autonomous robotic system was “Shakey”, a robot developed in 1958, at the Artificial Intelligent Center at the Stanford Research Institute (SRI) (Speck et al., 2017). It was able to perform advanced tasks such as localization, navigation, object detection, and has contributed to several advancements in artificial intelligence (AI) and robotics (Nilsson, 1984). Since then, with the rapid development of automation and AI, the characteristics and functions of robots have changed, as they are now able to display complex behaviors such as interacting and communicating with humans or other physical agents (such as robots or computers) (Hegel et al., 2009; Carradore, 2022).

For linking the needs of humans with the possibilities that robot technologies are able to provide, HRI plays a critical role by analyzing, designing, modeling, implementing, and evaluating the dynamic and complex interactions between humans and robots (Fong et al., 2003). The field of HRI also aims to examine different communication channels such as vocal, visual communication or displaying gestures to understand whether these are able to facilitate more natural and flexible interactions (Takeda et al., 1997).

According to the Computers are Social Actors (CASA; Reeves & Nass, 1997) paradigm, if computers behave as social actors, people tend to approach them similarly as they approach other humans. Nowadays, robots are able to display rich social behaviors during direct interactions, such as social cues that might impact the way we treat these technologies (Breazeal, 2004). To illustrate, Young et al. (2011) conducted a study to investigate how the social and physical presence of robots can be able to create a more complex and different interaction context compared to interaction experiences with other technologies (e.g. PC, mobile phones or home appliances). The authors argued that HRI is unique since robots are able to elicit salient and emotionally charged interaction experiences due to the social and physical characteristics of these interactions that elicit a strong sense of active agency. Modern robots such as humanoids and social robots are often capable of freely moving, displaying gestures and facial expressions; features which may encourage people to

perceive them as social partners. Next to the physical factors, the experiences of users could be influenced by a multitude of social factors such as affect or social norms (Young et al., 2011). However, the combination of these factors could elicit a phenomenon called "Uncanny Valley" (MacDorman & Ishiguro, 2006). According to Mori (1970), the "Uncanny Valley" is when robots appear more humanlike and familiar, resulting in the positive experience of humans potentially turning to negative as they view them as "eerie" or "unsettling" (see Figure 1). Therefore, Mori (1970) suggested that robot designers should be careful when designing robots that closely resemble humans due to the risk of "falling into the Uncanny Valley". On the other hand, MacDorman & Ishiguro (2006) argued that the experience of "Uncanny Valley" could have a positive aspect for the field of HRI and cognitive science, since it might reflect that our brains are processing some robots as humans; and, therefore, offering novel insights into unique human behaviors.

An interplay of these factors creates a holistic context surrounding the HRI experience, making it a unique interaction compared to other technologies (e.g. computers). This holistic context has important implications for evaluating HRI experiences as well. The following subsection will elaborate further on challenges and possibilities of evaluating and measuring these experiences, focusing on engagement, within the HRI field.

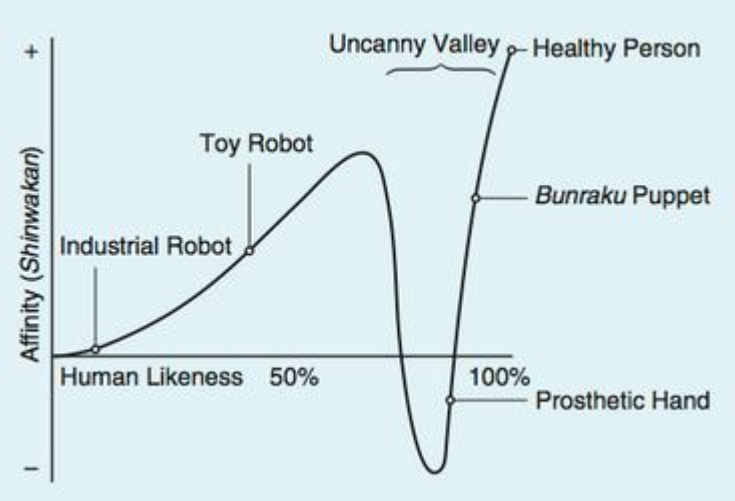


Figure 1. *The Uncanny Valley phenomenon as demonstrated in the original article (Mori, 1970)*

1.1.2 HRI metrics and engagement

The HRI research community has been aiming to establish a set of common metrics to evaluate interactions; however, the wide range of human-robot applications makes it difficult to define metrics that are suitable for all purposes (Steinfeld et al., 2006). Developing a toolset of metrics that fit every scenario may not be feasible or essential. However, adaptable metrics that use well-known scoring techniques might help the HRI area in assessing interactions in a variety of scenarios. Murphy and Schreckenghost (2013) developed a taxonomy of HRI metrics with three parent categories, measuring different aspects: the human, the system, and the robot. The system category is further divided into subsections such as productivity, efficiency, reliability, safety, and coactivity. The human component features metrics such as trust or situation awareness and most of these are inferred from psychophysiological measures, while the robot component includes for example self-awareness or plan state (i.e. robot's progress) (see Figure 2 for the full taxonomy).

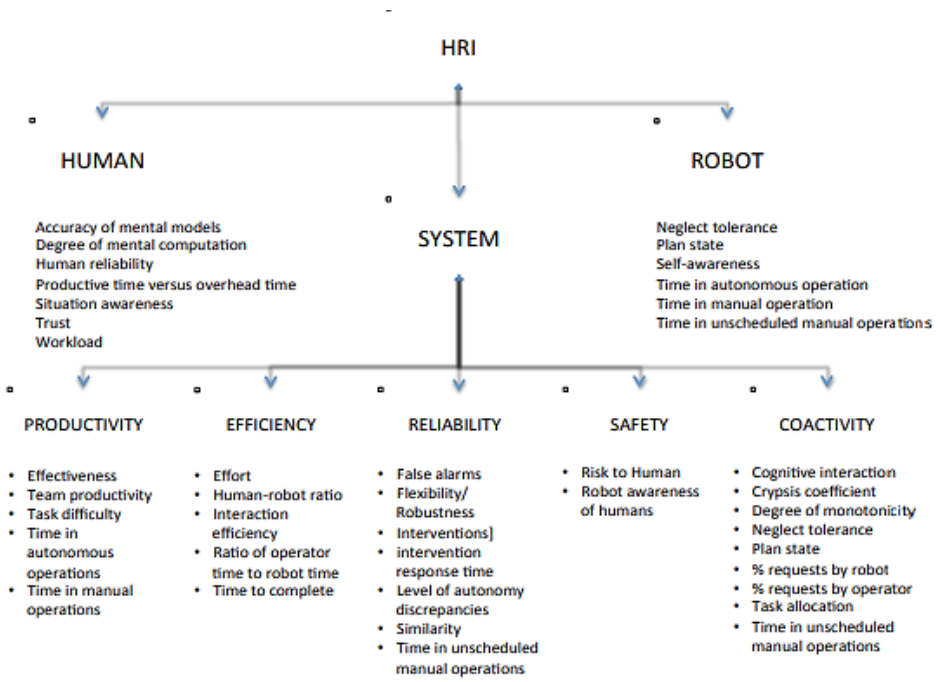


Figure 2. *Taxonomy of the HRI metrics (Murphy & Schreckenghost, 2013)*

Furthermore, HRI metrics can also be organized into task categories such as navigation, perception, social, management or manipulation depending on the primary objectives and tasks to be performed by the robot (Steinfeld et al., 2006). Social metrics can be applicable to robots with

primary tasks that require “social interaction”. In order to evaluate the effectiveness of these interactions, one can assess for example interaction characteristics (e.g. interaction style), trust, persuasiveness, compliance (e.g. adherence of norms) or engagement (Steinfeld et al., 2006). These metrics can be measured using a variety of methods, including questionnaires, psychophysiological measurements, observations, and data directly captured by the robot (Bethel et al., 2007; Rueben et al., 2020).

According to (Sidner et al., 2005) engagement is the process of establishing and maintaining an interaction and is characterized by challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control. Moreover, it can indicate the quality of the interaction and experience of users with the system (Oertel et al., 2020). Since human-robot social interactions are typically complex and multi-modal, analyzing engagement has been a challenge in the HRI research field (Lytridis et al., 2020). In some cases, measuring engagement directly is unfeasible. For example, questionnaires are impractical for children who cannot read yet, and psychophysiological measures are typically sensitive to movements. However, engagement could be inferred by observing social or non-verbal cues that indicate cognitive, effective, and attentional involvement (Lala et al., 2017). The advantages of examining these behaviors are that they are well-identifiable, easily obtainable, and are applicable in various HRI and research settings (Dautenhahn & Werry, 2002). However, context plays an important role when assessing non-verbal cues. For example, in an isolated context with one-on-one interaction, individuals may exhibit different behaviors than in a dynamic setting with more participants, such as auditory testing where children are accompanied by a speech therapist, audiologist and their parents. Since these interactions are more dynamic, the social aspects could play key roles during the interaction (Oertel et al., 2020). Furthermore, when humanoids are involved in these dynamic interactions, social non-verbal cues may be an important factor when assessing engagement. Humanoid robots, for instance NAO, can display social and guiding cues such as nodding, making the robot more salient, which could help users complete the tasks successfully while maintaining a natural interaction flow (Steinfeld et al., 2006).

Evaluating human-robot interactions are often done via observations based on video material (Dautenhahn & Werry, 2002). Observing and recording non-verbal cues could allow researchers to evaluate, for instance, facial expressions, articulated gestures and body posture, as well as direct physical contact between humans and robots such as shaking hands (Breazeal, 2003).

Moreover, these non-verbal cues are beneficial in measuring social characteristics such as capturing attention and holding interest and these are also the key aspects of engagement (Steinfeld et al., 2006).

Successful cooperation between humans and robots might benefit society in a variety of fields (Fong et al., 2003). For instance, healthcare has benefited from social robots in hospitals, clinics, and living-in-space facilities (Breazeal, 2003). In these areas, robots are considered more as collaborators or assistants rather than tools. A socially assistive robot is an interactive robot frequently applied in healthcare as assistants or companions. These robots have the potential to aid healthcare professionals and create a personalized, engaging environment for patients that helps them meet relevant health needs and goals (Breazeal, 2011). The next section will discuss the characteristics and applications of socially assistive robots.

1.2 Socially assistive robots

We distinguish socially assistive robots (SARs) from robots involved in conventional human-robot interactions (Fong et al., 2003). SARs can demonstrate characteristics such as advanced level of communication, being able to learn to recognize virtual agents and humans and display natural cues. These properties are particularly important in fields where robots are involved in peer-to-peer interactions, solving tasks (e.g. navigation) and effective communication (Dautenhahn et al., 2006). SARs could act as assistants or peers and are often applied in the research, medical and educational fields (Fong et al., 2003). The features and goals of these robots are generally specific: they aim to provide assistance by direct and effective interaction while achieving measurable progress in convalescence, rehabilitation, learning, etc. (Feil-Seifer & Mataric, 2005). Moreover, they have been applied in roles such as therapy aid for children dealing with grief or as social mediators for children with autism (Ismail et al., 2012; Kabacińska et al., 2021). In these contexts, evaluating and improving human-robot interactions are essential for creating personalized and pleasant experiences (Feil-Seifer & Mataric, 2005).

Humanoid robots are often used as SARs, as they are designed to resemble physical characteristics of humans, are typically equipped with sensors, actuators, cameras, speakers and could be preprogrammed to perform specific movements (Choudhury et al., 2018). Since these robots support social cues (e.g. nodding), people are able to communicate with them easily without the need of specific training (Breazeal et al., 2004). Additionally, as these robots are designed to

have anthropomorphic attributes such as a head, two arms and two legs; as well as social characteristics including gaze and gestures, we tend to view them as more humanlike (Salem et al., 2011). This could lead to a better natural interaction and a more positive HRI experience (Young et al., 2011). However, both the physical characteristics and non-verbal behaviors of the robot could contribute to the "Uncanny Valley" phenomenon (Thepsoonthorn et al., 2021). Therefore, it is important to examine aspects and features in robots that increase likeability without increasing the "Uncanny Valley" effect in humans.

One of the most popular humanoid robots widely used in research, education, and healthcare is NAO (Choudhury et al., 2018). It was developed by the French company Aldebaran-Robotics, later acquired by SoftBank Robotics (SoftBank Robotics, 2005) and features two 2D cameras, 11 touch sensors (three on each hand, three on the head, two on either foot, and the center chest button), four directional microphones and speakers, as well as the ability to move and adjust to the environment with 25 degrees of freedom (Robaczewski et al., 2021). It also includes vocal recognition and dialogues that are accessible in multiple (20+) distinct languages. Figure 3 shows an image of the NAO robot. These characteristics allow humans to naturally engage with the NAO. Additionally, the appearance and gestures of the NAO robot remain neutral (Thepsoonthorn et al., 2021). Thepsoonthorn et al. (2021) investigated whether the nonverbal behaviors of NAO (while giving a TED talk to the participants) could contribute to the "Uncanny Valley" experience. The results of the study revealed that when the robots expressed combinations of gestures (including hand gestures, gazing, face tracking), human-likeness and affinity were both rated highly (based on the results of the questionnaires). These findings could imply that displaying specific gestures could help overcome the "Uncanny Valley" effect when interacting with a NAO robot. The following section will discuss the practical implications of NAO as an assistant in the healthcare field.

Figure 3. *NAO robot (Courtesy of SoftBank Robotics)*

1.3 NAO in healthcare

NAO has been successfully implemented in several applications within healthcare settings including being a care robot, assistive robot, rehabilitation robot or trainer (Kyrarini et al., 2021). For instance, Qidwai et al. (2020) investigated whether the NAO robot would be a beneficial assistant during teaching activities for children with autism in a short case study conducted at a local school for children with autism spectrum disorder (ASD). The robot was programmed to teach behavioral and academic traits and take part in interactive activities such as storytelling, "Simon Says" game, morning exercise and a song-based game in the role of a teacher-assistant. During the experiment, memory, response time, type and number of trials were assessed. According to the findings, NAO has a strong potential to help autistic children learn more effectively. However, the research demonstrated that children who were afraid of the robot from the start did not perform well. On the other hand, when fear was replaced with more positive feelings as the interaction proceeded, such as curiosity and fun, the performance improved. Children who were intrigued by the robot from the beginning completed the tasks with more ease and fluency (during the teacher plus NAO condition) compared to the condition with the conventional teacher-based style utilizing the same tasks with the same group of children. The authors highlighted the positive outcomes of applying NAO as a teacher assistant but warned against using NAO in a robot-only setting due to

the lack of human-touch and human values that are crucial for social and communicational development.

Belpaeme et al. (2012) published an overview of the ALIZ-E project (Adaptive Strategies for Sustainable Long-Term Social Interaction) which aims to offer supporting robot companions to children diagnosed with metabolic disorders (diabetes and obesity) during their hospital stay via supporting their well-being and teaching them how to manage their health conditions. The ALIZ-E project implemented NAO and the study aimed to evaluate the HRI aspects involved in the project and provide an overview of the developments of the integrated system. The ALIZ-E system includes multiple game-like features (including quiz, dance-game, imitation game) and iteratively developing technologies (natural language competencies, memory structures, user modeling, bodily expression, and emotion). Several exploratory experiments and testing cycles were conducted to evaluate the HRI aspects of ALIZ-E within hospital settings. For instance, diabetic participants were provided with diabetes-based quiz games, and they were playing either with an adaptive or a non-adaptive robot (Blanson Henkemans et al., 2012). The adaptive robot asked personal questions, used this information (the answers from the questions) while interacting with the child and asked questions regarding their experiences at the end of the game. While their understanding of diabetes significantly increased in both conditions, enjoyment ratings (based on questionnaires) were (non-significantly) higher in the adaptive robot condition. Nalin et al. (2012) aimed to examine whether a social bond could emerge in children during their hospital stay when they interact with ALIZ-E. The children were able to play with a robot for three hours during several days and could choose between three games: quiz, dance and imitation and they could stop the game at any point. At the end of sessions, they had to fill out questionnaires that assessed their experiences. The children generally reported positive experiences and feelings of "happiness" even at the third encounter when the novelty effect was not salient anymore. When they were required to choose a label to describe their relationship with the robot, the most commonly used labels were "friend", "brother or sister" and "classmates" and the least often used labels were "acquaintance" and "stranger". These observations could suggest that NAO has the potential to be applied within hospital settings for children and it could be a successful assistant during their hospital stay, however further research is needed that investigate different interfaces (e.g. NAO and PC), technical integrations (e.g. type of gestures), and other experiences (e.g. via examining non-verbal cues during interactions).

Students with intellectual disabilities (ID) may also benefit from NAO, since it can assist them in achieving their unique learning objectives. Hughes-Roberts et al. (2019) conducted a study with adolescents attending therapy centers and schools specializing in multiple disabilities such as ID and autism. Participants had a variety of disabilities, including Down's syndrome, hearing loss, developmental delays, and epilepsy and often suffered from combinations of these disorders. Because the participants were intellectually diverse, they each had their own learning objectives, such as vocal imitation, object identification, or reacting to their own name. Along with a qualified researcher and a teaching assistant, a NAO robot assisted them in achieving these goals. While the results of the study revealed that engagement levels were higher and the percentage of independently obtained goals were larger in the robot sessions compared to the control, these differences were not significant. The researchers argued that since students differed in several aspects, the intervention should be examined on a case-by-case basis to determine the unique factors that contribute to a successful intervention. Furthermore, they concluded that eye gaze may not be a valid indicator of engagement in the studied population since autistic children do not apply eye contact in the same manner as other children do. Therefore, further research is needed on investigating individual cases and disabilities to determine when NAO is most engaging as well as improving the methods for detecting engagement to ensure validity and reliability.

Hughes-Roberts et al. (2019) highlighted that developmental disabilities are often associated with varied degrees of hearing loss. Hearing loss may have a number of implications on students' performance, not only in the cognitive but also in the psychosocial domain. NAO could offer assistance for children who are impacted with hearing loss either due to developmental disability or other external/internal factors. Although prior studies have examined the effectiveness of humanoid robots such as NAO as a companion for children with various health conditions, research investigating whether NAO could be an effective assistant for children with hearing impairment has been limited. The upcoming sections will discuss the impact of hearing loss in children and previous research on applying humanoid robots in auditory fields.

2. Theoretical framework

2.1 The impact of hearing impairment

Hearing impairment has become increasingly prevalent in recent years (World Health Organization [WHO], 2018) and detecting hearing loss early could enhance patients' quality of life. Moreover, it is more than merely a medical condition; it can also have an impact on a person's social and cultural life, making it a multifaceted problem. For instance, it can affect phases on language use, educational experience, and social identification (Hindley, 1997).

The two most common devices to compensate for hearing loss are hearing aids (HAs) and cochlear implants (CIs). HAs aim to compensate for reduced audibility and some suprathreshold distortions (e.g. reduced dynamic range due to loudness recruitment) (Armstrong et al., 2022). HAs are often effective in compensating for reduced audibility, but not all suprathreshold deficiencies can be easily compensated for, such as frequency selectivity (Sanchez-Lopez et al., 2020). CIs process sound and deliver the sound signal directly to the acoustic nerve by electric stimulation and have been shown to provide speech understanding in deaf children and adults (Abdel-Latif & Meister, 2022). CI devices provide speech signal to deaf individuals; however, due to the limitations in electric stimulation of the auditory nerve, the signal delivered is impoverished in spectro-temporal fine details (Zhou et al., 2020). Both situations create difficulties for hearing in users of these devices. Additionally, there are certain differences when comparing CI users to normal hearing individuals in the perception of vocal characteristics of sounds, perceiving speech in noise or perceiving degraded speech due to factors such as: different methods of coding of voice pitch in the CIs, nature of maskers, or the differences in listening effort (Başkent et al., 2016).

Nowadays, CIs (unilateral or bilateral) are very common among children with hearing loss (Vermeulen et al., 2012). Unilateral CI refers to patients fitted with one CI, while bilateral CI users have two CI devices. Bimodal CI+HA is the combined use of a HA and CI. Children with hearing loss undergo regular auditory tests to determine their speech perception abilities and have their HAs or CIs adjusted (Uluer et al., 2021). These tests are generally conducted by the audiologist and/or speech therapist, and the most commonly used tests are pure tone audiometry and speech audiometry (DeBow & Green, 2000). Pure tone audiometry is considered as the “gold-standard” hearing assessment (Maclennan-Smith et al., 2013) and the foundation of this test is detection thresholds of pure tones that are obtained at octave frequencies between 125 Hz and 8 kHz

(Bosman, 1989). Pure-tone audiometry scores can be divided into five categories: normal (<= 20 dB HL), mild (21-40 dB HL), moderate (41-60 dB HL), severe (61-80 dB HL) and profound (> 80 dB HL). Speech audiometry is generally applied complementary to pure-tone audiometry to assess speech perception and communication abilities of the individual, which, depending on the etiology, could differ from what the hearing thresholds indicate (Bosman, 1989). Speech audiometry tests are able to give a rough estimate on how well speech could be perceived in quiet or in noisy environments, depending on the type of the test. NVA (Nederlandse Vereniging voor Audiologie) lists and the DIN (Digit-In-Noise) test are widely used speech audiometry tests in the Netherlands. The NVA lists correspond to a speech-in-quiet test and consist of 60 lists (45 for adults and 15 for children) each containing 12 words, forming a pool of 177 CVC (consonant-vowel-consonant) words (Vanpoucke et al., 2022). The first word of each list is a practice item, and the scoring is performed at the phoneme (basic unit of a speech sound in a language) level. The 11 test items consist of 33 phonemes (three phones per word); therefore, each phoneme corresponds to 3% of the final score for the list. The total percentage score (speech intelligibility score) is the number of correctly heard phonemes multiplied by 3, and 1% is added if the total score is higher than 50% (Veispak et al., 2015). The maximum speech perception score is the total percentage score of the NVA list with speech signal presented at 65 dB sound pressure level (Spirrov et al., 2018). The DIN test is an adaptive speech-in-noise test consisting of a set of 120 unique digit triplet combinations constructed from the digits 0 to 9, separated by silent intervals (Vroegop et al., 2021) and presented as a set of 24 predefined triplets per test. The stimulus is mixed with a constant masking noise. The DIN determines the signal-to-noise ratio (SNR) that is required for an individual to identify 50% of the presented triplets correctly. The final DIN score is the speech reception threshold (SRT) which is the SNR corresponding to 50% correct recognition (SRT is calculated by taking the average SNR of trial 5 to 25) (Van den Borre et al., 2021). For the DIN test, the SRT is assessed by adaptively varying the SNR with a step size of 2dB depending on whether the response is correct or incorrect (one-up one-down procedure) (Vroegop et al., 2021). The standard deviation (SD) of the score determines whether the DIN was reliable. For a reliable DIN test score, SD should be below 3.6 for children with mild to profound hearing loss for the first test and below 3.1 SD for the second test (Vroegop et al., 2021).

The testing equipment (for pure tone audiometry) generally consists of a screening audiometer and headphones throughout a session guided by trained personnel (e.g. speech therapist

or audiologist) (Walker et al., 2013). However, a limited number of studies have been conducted investigating whether robots are able to provide support for children during speech audiometry. The following section will discuss some of these studies that utilized NAO or other socially assistive robots as assistants supporting children with hearing loss and/or during speech audiometry.

2.2 SARs supporting children with hearing loss

Ioannou and Andreeva (2019) conducted a study with hearing impaired children who had HAs or CIs to investigate whether NAO could be an efficient tool for supporting speech therapy. The advantage of NAO over standard speech therapy sessions is that NAO could create a playful, engaging learning environment. Moreover, as children cannot read the lips of the robot, they must rely on their CIs or HAs during the sessions. The results demonstrated that all participating children showed steady progress throughout the six weeks of the intervention period and were able to respond consistently and correctly to sounds towards the last two weeks. Additionally, they displayed positive attitudes towards the robot and found the interventions entertaining. Previous studies demonstrated that children are often highly interested in technological advances such as computers, gadgets or robots (Shamsuddin et al., 2012). Moreover, since NAOs have a unique physical form and novelty, children might find them especially appealing and motivating (Lytridis et al., 2020). Abdul Malik et al. (2014) suggested that since NAO could act like a toy or friend and increase children's attention, the impact of a therapy could be more effective when the humanoid robot is present.

NAO could have the potential to become an engaging tool during speech audiometry since it is able to provide guiding cues such as gaze or gestures, making the sessions more salient (Breazeal, 2003). Ondáš et al. (2019) aimed to develop a research platform to study aspects of robot-supported audiometry. They utilized NAO integrated with an asymmetric multimodal dialogue system, where the robot could use gestures and speech to interact with users. The experimental setting included Conditioned Play Audiometry (CPA), which consists of interactive game-like tasks where the child interacts verbally by performing specific activities such as selecting cards based on the sounds they hear. During the CPA sessions, the children were seen happy and excited, they reacted positively to the robot's invitation to help it learn new words and the sessions were successful. The authors concluded that NAO has a potential in audiometry and in increasing

patient attention and motivation. However, they only included a small sample of children with unaffected hearing, highlighting the need for further research that involves children with hearing disabilities.

Finally, Uluer et al. (2021) investigated whether the humanoid robot Pepper (also developed by SoftBank Robotics) supplemented with emotion recognition could assist children with hearing loss during auditory testing. The RoboRehab project included an affective module that processed facial and physiological data, a gamified tablet interface for the tests, and a social robot assistant that ran the tests and provided feedback. The affective module was used to recognize emotions of children via machine learning and deep learning methods. The experiment was conducted with hearing impaired children and three different setups were assessed: conventional, tablet (tablet-based auditory game with visual feedback), and robot plus tablet (tablet-based game but feedback via gestures). In the conventional setting, the audiologists played pairs of stimuli sequentially and asked the child to answer whether they perceived the pairs to be same or different. The tablet setup featured an auditory perception game with visual feedback on the tablet, and the robot setup included the same tablet-based game, but the feedback was provided via the gestures of the robot. Children were accompanied by an audiologist during both the tablet and robot setups.

In addition to collecting physiological signals and facial expressions, two surveys were conducted to explore the subjective evaluations of children. The results of the affective module indicated that the robot had a stimulating presence during the tests. Furthermore, test metrics showed that while the setup did not impact the hearing test results, the total testing time increased in the robot plus tablet compared to the conventional setup. Video annotations were examined to assess whether the longer testing times were due to increase in engagement levels. Since the conventional study was performed during the COVID-19 epidemic and all the children were wearing facial masks, the behavioral analysis results were inconclusive for the comparison of the conventional setup with the gamified setups. However, the results demonstrated that smiling, mimicking, and talking behaviors appeared in the robot plus tablet condition but not in the tablet only condition, possibly indicating greater engagement towards the robot. Additionally, survey results demonstrated that younger children were more excited to see the robot compared to older ones.

In view of the above-mentioned studies, NAO could provide a pleasant interaction experience for children with hearing loss. The present study aims to further evaluate the potential

of SARs, particularly NAO, in speech audiometry for children with hearing loss. Video annotation could be beneficial in evaluating such interactions especially in settings that involve children with hearing loss, since other methods, such as questionnaires could be less applicable (due to age and cognitive background of the children). The following section will discuss earlier research that assessed behaviors (such as engagement or discomfort) utilizing video schemes or annotation.

2.3 Video schemes in HRI research

To research engagement, social interaction, or other characteristics of HRIs, a wide range of methodologies have been employed including surveys, physiological measurements, and video annotations. Analyzing video recordings with video annotations is a frequent method for evaluating HRI, particularly when verbal reports or questionnaires are not available, optimal, or may bias the findings (e.g. when children are not yet able to read) (Kidd & Breazeal, 2005).

For instance, Koay et al. (2006) investigated the comfort level of seven adult subjects with respect to 12 robot behaviors by a handheld device (where subjects used a continuous scale to judge their current comfort level), questionnaires and analysis of pre-recorded video data to identify instances of discomfort. The participants had to perform two tasks: Negotiated Space Task and an Assistance Task. During the Negotiated Space Task, the robot moved around the room as the subject went through a stack of books, memorizing one title at a time, and writing down each title on the whiteboard. During the Assistance Task, the subject had to copy the book titles from the whiteboard onto paper while seated at a table, underlining particular letters with a red pen. The robot was responsible for bringing the missing pen to the table. The video footage was analyzed by two video coders, using a predefined video annotation scheme. The video scheme helped coders to identify instances of discomfort and included behaviors that indicate discomfort such as jumpy or jerky body movements, surprised facial expressions and the coders also had to indicate details of robot behaviors (robot actions, proximity and motion). To assess the consistency between the coders, Cohen's Kappa was calculated, and the video scheme was assessed applying matching rules (strict or relaxed matching rule depending on whether suggested instances of discomfort of both or only one coder was matched with the data from the handheld device). Findings indicated that when the robot was blocking the path during the experiment, subjects were likely to experience discomfort.

Lala et al. (2017) aimed to detect engagement during conversational HRIs by building a real-time engagement recognition model. Recorded videos were analyzed of 91 conversational human-robot sessions in order to select the most relevant social signals. The videos were annotated by several coders who marked the beginning and end points of each target behavior (verbal backchannels, nodding, laughter, eye gaze). The target behaviors were chosen based on a previous study where behaviors accompanying perceived engagement were identified (Inoue et al., 2016). Six annotators took part in the study by Inoue et al. (2016) and after rating the video sessions, agreements among annotators were assessed and discrepancies were discussed. The most common indicators were determined based on transition relevance places, relationship with multi-modal behaviors and time distribution of annotations. The outcomes of these analyses demonstrated that the most relevant engagement indicators were nodding, laughter, verbal backchannels, and eye gaze, therefore these behaviors were included in the engagement recognition model featured in the research by Lala et al. (2017). After testing the model, the authors noted that, even though the difference was not significant, they noted a drop in performance of the engagement model when using behavior detection compared to manual annotation (Lala et al., 2017). A similar study was conducted by Jang et al., (2014). The authors aimed to develop a classifier to recognize engagement in children during a robot-conducted math quiz game. Videos of seven children were annotated by three different coders. The coders had to indicate different social signals such as posture (straight, stoop), speech (robot/child), gaze direction (e.g., TV, robot), behaviors (e.g., head-nod, headshake, touching-hair, touching-face) or facial expression (smile, laughter, neutral). After coding the social signals, they also had to indicate engagement as two states (engaged = 1, not engaged = 0). The high agreement (Cohen's Kappa above 0.6) behaviors were the following: posture, behavior, gaze direction, game context and speech by robot.

Serholt and Barendregt (2016) investigated whether children express signs of social engagement when socially significant events were initiated by a robot. This project was a longitudinal field trial at a primary school lasting for 3.5 months. The interaction sessions took place during standard classroom lessons and included map tasks and treasure hunts guided by a NAO robot. The sessions were recorded and annotated by the two authors. The coders used a predefined video scheme that included both positive and negative indicators based on previous literature by Argyle and Dean (1965), Castellano et al. (2009) and Vacharkulksemsuk and Fredrickson (2012). Examples of social engagement indicators were gazing at the robot, timid or

flushed smiles, reacting to the greeting of the robot, waving, nodding, head shakes and mirroring. Gazing elsewhere other than at the robot, nervous or confused expressions were considered as indicators of no social engagement. The results showed that most frequent facial behaviors were smiling or looking serious, the most frequent verbal responses were indicating understanding or agreement, e.g., “Yes” or “Okay” and the most common gestures were head nods.

These studies demonstrate that a well-defined video scheme could help identify behaviors that might be hidden to other assessments (e.g. questionnaires). Additionally, video annotations provide time stamped data that could be re-processed (Kidd & Breazeal, 2005). However, the observation process is subjective and might be influenced by the personality, attitudes, or other personal factors of the coders. A well-developed and understandable video scheme with clear descriptions (and examples) of the behaviors could help to avoid uncertainties.

The goal of the present study is to focus on non-verbal cues that appear in everyday interactions and to develop a video protocol that is easily adaptable to human-robot interactions in various settings. Furthermore, it seeks to contribute to the preliminary findings of assessing non-verbal cues in HRI studies in the auditory field. The next section will present the details on how these objectives will be examined as well as the relevant research questions and hypotheses.

3. Research questions and relevance of the present study

Even though findings of previous research are promising regarding eliciting engagement during human-robot interactions with NAO (Uluer et al., 2021), there has been limited research on investigating non-verbal cues as engagement indicators in audiometric settings supported by NAO. Engagement is a key factor in audiometry and children with hearing loss undergo several audiometric tests to assess hearing levels and adjust hearing devices. Incorporating NAO as a support for children with hearing disabilities during speech audiometry could result in a pleasant and engaging experience. However, to determine whether NAO is able to provide a positive experience, children's engagement must be measured during the auditory sessions. There is currently no standard approach for evaluating these metrics and detecting them is a challenging task in HRI due to the complexities of these interactions. Observation and annotation of non-verbal cues could be a suitable method in audiometry and has been frequently applied in HRI research (Abdul Malik et al., 2014; Uluer et al., 2021).

The goal of this exploratory study was to create a video protocol with standard scoring metrics of non-verbal cues to evaluate engagement levels in HRI during audiometry with children. The non-verbal cues were selected based on prior studies of child-robot interactions (Dautenhahn & Werry, 2002; Jang et al, 2014; Serholt & Barendregt, 2016). The video scheme included easily detectable, natural non-verbal behaviors that emerge in human-human as well as human-robot interactions. A reliable video protocol could aid researchers and clinicians to share knowledge, compare and evaluate relevant outcomes and findings in the child-robot interaction (CRI) field (Steinfeld et al., 2006).

The research questions addressed in the present study are the following:

1. What are the most and least common non-verbal cues that indicate engagement or no engagement towards the robot during auditory testing with children?

2. How does the frequency of these cues change throughout the auditory sessions; in particular, comparing the beginning and the end of sessions?

3. How do these cues differ for children with lower levels of maximum speech scores compared to the overall sample?

It is hypothesized that the most common non-verbal cues would be engagement indicators while the least frequent ones would not be indicators of engagement (H1). Furthermore, it is expected that more in general, more cues would appear at the beginning of the video and less towards the end as the novelty effect of the robot wears off (H2). Finally, children with lower levels of maximum speech scores would display similar cues compared to other children indicating that the NAO is able to effectively supporting them as well during the interactions (H3).

4. Methods

4.1 Participants

27 hard-of-hearing children were participating in the study conducted at the UMCG. Prior to the testing, all parents provided their informed consent for the sessions and the video recording (as approved by the METc ethical review committee at UMCG: METc 2018/427, ABR NL66549.042.18) and all children took part in the study during their scheduled clinical session at the ENT outpatient clinic in the UMCG. Recorded videos of the testing were analyzed in the present

study. The age range of participants was between 4 and 15. Hearing devices, specifically one or two CIs or one CI and one HA, were used by the participants and neuropathy was not in the exclusion criteria. The video recording of one child was excluded from the study since they only had HA as a hearing device (without CI) and they attended school for normal hearing children; therefore, the total sample size of the current study was n = 26. Participants' relevant demographic characteristics are summarized and presented in Table 1.

Table 1.

Demographic characteristics

Participant characteristics	N (%) / M(SD)
Age	
Mean (SD)	8.97 (2.8)
Hearing device	
Bilateral CI	19 (73%)
Unilateral CI	3 (12%)
Bimodal CI+HA	4 (15%)
Hearing Loss (PTA $_{0.5\text{-}4\ kHz}$ *better ear*, unaided)	
Mild (21-40 dB HL)	1 (4%)
Moderate (41-60 dB HL)	1 (4%)
Severe (61-80 dB HL)	2 (8%)
Profound (>80 dB HL)	22 (84%)
Speech perception in quiet (65 dBSPL,%, aided)	
Mean (SD)	90.8 (13)

Note. M = mean, SD = standard deviation, N = sample size, % = percentage

4.2 Materials

4.2.1 Robot

A NAO robot (SoftBank Robotics) was utilized for this research, which is a 57 cm long humanoid robot with speakers and microphone. The NAO was operated with the Wizard of Oz approach

(WoZ) which refers to a person controlling the robot's movement, navigation, voice, and gestures remotely (Riek, 2012).

4.2.2 Setting

The clinical sessions were in Dutch and lasted approximately 1 hour. The interaction with the robot was also in Dutch and lasted approximately 10 minutes and the children had to sit on a chair in front of a desk, where a NAO robot was placed (Figure 4). Their parents, the speech therapist, audiologist, and researchers were present in the room. The interaction included three different phases: static robot, warble tones and speech audiometry test. The clinician controlled the aspects of the interaction including the time of each stage (except 30s static) and the choice of the speech audiometry test. The speech therapist sat on the other side of the table, assisting the children. Before the interaction began, the therapist had a brief talk with the participants, while the NAO robot sat still on the table, blinking. Afterwards, the speech therapist demonstrated the warble tones. The warble tones (sampled between 150 - 5000 Hz, calibrated at 65 dB) were complex sounds played as triggered events when the tactile sensors on the robot's head, hands and feet were pressed to raise attention in children and offer an opportunity to get to know the robot. Next, depending on the clinician, the participants had to complete either the DIN (speech-in-noise) test and/or NVA (speech-in-quiet) lists. The speech audiometry tests began with the robot getting up and taking up a crouching position facing the participant, while the audiologist or speech therapist explained the task in question. During the tests, the robot provided positive feedback in the form of a nodding head movement. As additional feedback, the audiologist was able to select “fist bump” or “blue eyes”. “Fist bump” (or “boxing”) refers to the robot raising the arm and offering a “fist bump” as positive feedback, while “blue eyes” refer to the robot eyes changing from default white to blue color to increase attention. Before each test, the clinician would select the dB for testing. The levels of the speech materials were calibrated with a Knowles Electronics Mannequin for Acoustic Research (KEMAR, GRAS, Holte, Denmark) located at approximately 80 cm from the NAO and a sound-pressure level meter (Type 2610, Bruël Kjær and Sound & Vibration Analyser, Svan 979 from Svantek). During the NVA lists [45,75 dB] the robot played monosyllabic words that the child had to repeat, while during the DIN test [45,72 dB], the child had to repeat the digits.

Figure 4. *Interaction with the NAO robot*

4.2.3 Video recordings

The video recordings included all the aforementioned child-robot interactions, from the static-robot phase to the speech audiometry test. The parts selected later for analysis were in which children interacted with the robot. The interactions were recorded with two cameras from two angles: lateral and frontal. In the lateral view, the camera captures the entire body, while the frontal view focuses on participants' faces. The parents of the children provided consent for recording the sessions and using the data for further analysis.

4.3 Measurements

4.3.1 Video protocol

The video protocol was created based on earlier research that developed video schemes to assess HRI in various contexts (Dautenhahn & Werry, 2002; Heerink et al., 2012; Henkemans et al., 2017). Table 2 represents the complete video protocol. It includes 18 non-verbal behaviors, 11 of these were related to the body position and movements, and three were directly related to the robot (grabbing, pointing at and waving at the robot). Ten of the behaviors were static (measuring frequency) and one was dynamic (measuring duration, i.e. actively moving). The remaining seven behaviors were focused on the face of the children and were measured as dynamic behaviors.

Table 2.

Complete video protocol

Camera view	Camera focus	Event type
Lateral view	***Position***	***Static***
		Leaning towards the robot
		Leaning away from the robot
		Hide away
		Nodding
		Shaking head
		Covering ears
		Clapping
		Pointing at robot
		Grab robot
		Wave at robot
		Dynamic
		Actively moving
Frontal view	***Face***	***Dynamic***
		Amused
		Bored
		Mocking
		Surprised
		Distressed
		Distracted
		Other

4.3.2 Descriptions of the non-verbal metrics

The static behaviors measured in this research are described as follows:

Leaning towards and leaning away from the robot

Leaning towards the robot indicates that the child is purposefully moving their torso closer to the robot (decreasing the angle formed between the legs and the torso), while leaning away from the

robot implies that the child is purposefully moving the torso away from the robot (increasing the angle formed between the legs and the torso).

Hide away

Hiding away is described as covering the face with hands/other parts of the body in order to be out of sight of the robot/clinicians.

Nodding

Nodding is moving the head in an up-down motion as a sign of agreement or understanding. It has been regarded as a common backchannel in conversations (Lala et al., 2017).

Shaking head

Shaking the head is moving the head in a sideways motion as a sign of disagreement.

Covering ears

Covering the ears is the motion of using hands/other parts of the body for cupping the ears to indicate that the child is not hearing or does not want to hear the sounds surrounding them.

Clapping

Clapping is striking palms together repeatedly (with or without producing sound).

Pointing at robot

This behavior refers to using a finger to direct attention towards the robot.

Grab robot

Grabbing the robot is invading the robot's space by grasping the head, torso, arms and/or legs. It also includes hugging/kissing/lifting the robot.

Wave at robot

Waving at the robot is moving hands in a sideways motion in front of the robot. Waving is a social behavior that can be used to attract the other partner's attention or as a greeting (Rousseau et al., 2013).

The dynamic behaviors are described as follows:

Actively moving

This metric includes behaviors such as dancing, moving, playing with the body, stretching, scratching, or fidgeting.

Amused

This behavior indicates that the child is looking entertained or smiling, laughing, and showing excitement.

Bored

Boredom can be seen when the child is looking weary, being impatient, yawning, rolling eyes.

Mocking

Mocking is when the child is making funny faces to the robot.

Surprised

Being surprised is reflected by gasping when looking at the robot.

Distressed

A child is distressed when for example he/she is crying or throwing a tantrum.

Distracted

Children are distracted when they are not focusing on the robot but looking at the clinician, parent, other people or objects for a longer period of time.

Other

Actions that are as unclassified, or notes from the annotators such as when the child looks confused or puzzled etc.

4.4 Procedure

4.4.1 Video annotation

The video data was manually coded by two independent coders using the predefined video protocol (see Table 2). Depending on the behaviors, occurrences (in case of static/point behaviors) or beginning and endpoints (in case of dynamic/state behaviors) were annotated. The coding of the videos was performed with the BORIS software (Friard & Gamba, 2016). Due to time constraints, one of the coders was not able to fully complete the rating for the frontal view. Figure 5 shows the behavioral coding map of the video scheme.

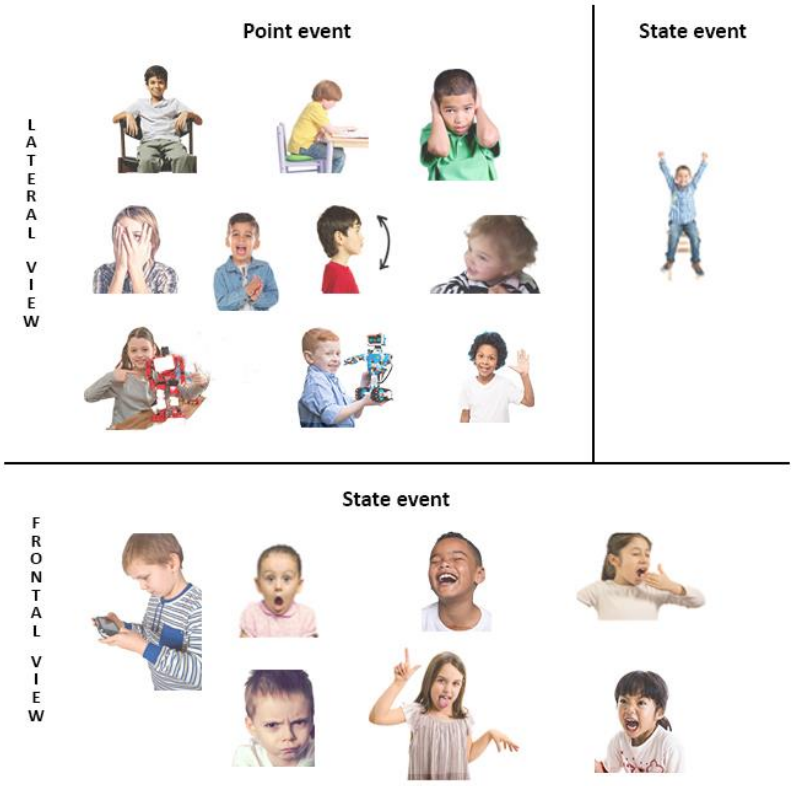


Figure 5. *Behavioral coding map of the video scheme*

4.5. Statistical analysis

4.5.1. Descriptive statistics

The descriptive statistics analysis included total time, mean time, standard deviation, maximum and minimum durations of the behaviors. Moreover, the annotated frontal and lateral view as well as the total count of all non-verbal cues were calculated for the videos that were logged by both coders (processed data) as well as for the raw data.

4.5.2. Intercoder reliability

In order to investigate consistency between coders, percentage agreement was calculated both for the raw data (with missing data regarded as non-agreement) and listwise deleted data (missing data omitted). The listwise deleted data was applied due to two many missing cases for coder 2 which could skew the data. Therefore, the listwise deleted data attempts to provide a better picture on the reliability between the two coders since it only includes the videos logged by both coders. The listwise deletion was performed with the following method: 1. timings of behaviors were matched for both coders (± 3 s), 2. missing data (where one coder data was not available) were deleted. Additionally, times were matched (± 3 s) for the raw data as well. Percentage agreement and Cohen's Kappa (unweighted) was calculated across all behaviors as well as for point and state behaviors separately.

4.5.3. Intercoder comparisons

To examine similarities and differences between the coders for the analyzed behaviors, aggregated counts and percentages of behaviors and lateral and frontal view behaviors were calculated and plotted for as well as total summed durations of state behaviors (measured in seconds). Since the purpose of these analyses was to compare the annotations between the two coders, the counts, percentages, and durations were calculated applying the processed data (i.e. data logged by both coders: frontal view videos n = 4, lateral view videos n = 26).

4.5.4. Intracoder comparisons

Start times of point behaviors were measured and plotted separately for coder 1 and coder 2 and divided into three categories based on the video lengths (short <= 600 s, medium = 600-800 s, long >=800 s). Additionally, aggregated durations of state behaviors were calculated for coder 1 and

coder 2 separately and represented in one plot. The data sample used for the intracoder comparisons was the original (raw) data since the analysis aimed to examine the trends of the non-verbal cues for coder 1 and coder 2 separately.

4.5.5. Case studies

Videos of children who had maximum bilateral speech scores (aided) below 80% were analyzed case by case. Ages of these participants are presented in decimals to maintain consistency with previous studies that included children with hearing loss. Maximum bilateral speech scores were obtained from the scores of the NVA lists. The counts and the durations of the annotated cues were plotted and calculated, and the anecdotal evidence was examined as well. The analyzes for the case studies were performed with the raw data sample (the frontal view coding only includes the annotations of coder 1).

5. Results

5.1. Participant flow and missing data

Due to time and practical constraints, one of the coders could not complete the analysis of the frontal view videos therefore, annotated data of 22 subjects were missing for this viewpoint (total n = 4). Consequently, for the intercoder comparisons and (listwise deleted) intercoder reliability, the final sample for the lateral view behaviors, was n = 26 and for the frontal view behaviors, n = 4. For the intracoder comparisons, the examined sample size for coder 1 was n = 26 both for frontal and lateral view and for coder 2, n = 26 for lateral and n = 4 for the frontal view.

5.2. Descriptive statistics

The total number of coded behaviors were 1472, of which 729 (49.5%) were state and 743 (50.4%) were coded point behaviors. Coder 1 logged 488 (33.1%) state behaviors and 388 (26.4%) point behaviors, while coder 2 logged 241 (16.4%) state behaviors and 355 (24.1%) point behaviors. The data tables including the original (raw) data can be found in Appendix A. The number of logged events for each behavior for the videos that were analyzed by both coders (lateral n = 26, frontal n = 4) are presented in Table 3**.** The mean duration for the state behaviors for subjects logged by both coders was 13.93 s ± 51.05 s for frontal view and 16.76 s ± 25.48 s for lateral view (Table 4). The total duration the videos was 18342 s, the mean time was 705.46 s ± 189.33 s (Table 5).

Table 3

Counts of coded behaviors

Coded behaviors	N (coder 1)	N (coder 2)	N (total)
Number of state and point behaviors (lateral, n = 26)			
State	100	136	236
Point	388	355	743
Number of state behaviors (frontal, n = 4)			
State	44	79	123

Note. State behavior = actively moving. Number of videos (subjects) both coders analyzed are denoted by "n". Subjects both coders analyzed in the frontal view: 001, 002, 003, 005.

Table 4

Durations for state behaviors

	N (video)	Mean(s)	SD	Median(s)	Min(s)	Max(s)
Frontal	26	13.93	51.05	5.5	0.5	565.95
Lateral	4	16.76	25.48	8	0.74	164.32

Note. s = second. Only the durations of state behaviors of subjects whom were coded by both coders are presented

Table 5

Video lengths

	N (video)	Mean time(s)	SD	Shortest(s)	Longest(s)
Video timings	26	705.46 (11.76 min)	189.33 (3.15 min)	455 (7.6 min)	1156 (19.27 min)

Note: s = second. Min = minute. Only the videos that were coded by both coders are presented in this table.

5.3. Intercoder reliability

5.3.1. Original data

The percentage agreement of the raw data (tolerance = 0) was 18.8% for all behaviors, 20.2% for point behaviors and 13.1% for state behaviors between the two coders. Tolerance is the number of successive rating categories that should be regarded as rater agreement. Tolerance was set to 0 since

the data was nominal (tolerance other than 0 is only applicable to numerical values). Cohen's kappa (κ) was 0.0076 ($z = 0.54$, $p = 0.59$) for all behaviors, 0.0188 ($z = 1.16$, $p = 0.25$) for point behaviors and -0.58 ($z = -9.34$, $p < 0.001$) for state behaviors.

5.3.2 Listwise deleted data

The percentage agreement for the listwise deleted data (tolerance = 0) was 66.7% for all behaviors, 70.3% for point behaviors and 51.2% for state behaviors between the two coders. Cohen's kappa (κ) was 0.602 ($z = 21.6$, $p < 0.001$) for all behaviors, 0.628 ($z = 18.4$, $p < 0.001$) for point behaviors and -0.14 ($z = -1.98$, $p = 0.048$) for state behaviors.

5.4. Intercoder comparisons

The counts of all behaviors can be found in Appendix B. The most frequent behavior for coder 1 was “nodding” (152 occurrences, 13.7% of total behaviors), “actively moving” (100 occurrences, 9% of total behaviors), and “lean towards the robot” (78 occurrences, 7% of total behaviors) while for coder 2, “leaning away from the robot” (120 occurrences, 10.8% of total behaviors), “leaning towards the robot” (119 occurrences, 10.8% of total behaviors) and “actively moving” (93 occurrences, 8.5% of total behaviors). The least frequent behaviors were “mocking” (1 occurrences, 0.1% of total behaviors) and “distressed” (1 occurrences, 0.1% of total behaviors) for coder 1 and “bored” (1 occurrences, 0.1% of total behaviors) and “distressed” (1 occurrences, 0.1% of total behaviors) for coder 2. Behaviors with 0 occurrences were “bored” and “distressed” for coder 1 and “covering ears” and “other” for coder 2. Figure 6 illustrates a bar chart for all behavior counts for coder 1 (red bars) and coder 2 (blue bars). Each bar represents a coded behavior and longer bars indicate higher values. The two different colors highlight differences between the annotated behaviors of the two coders. The graph demonstrates that -with the exception of the behavior „grab robot”- the numbers of the coded behaviors mostly varied between the coders. However, there are similarities between the behaviors with the most and least behavior counts such as „leaning toward the robot” and “actively moving” (both coders annotated as most common) and „distressed” and “bored” (least common).

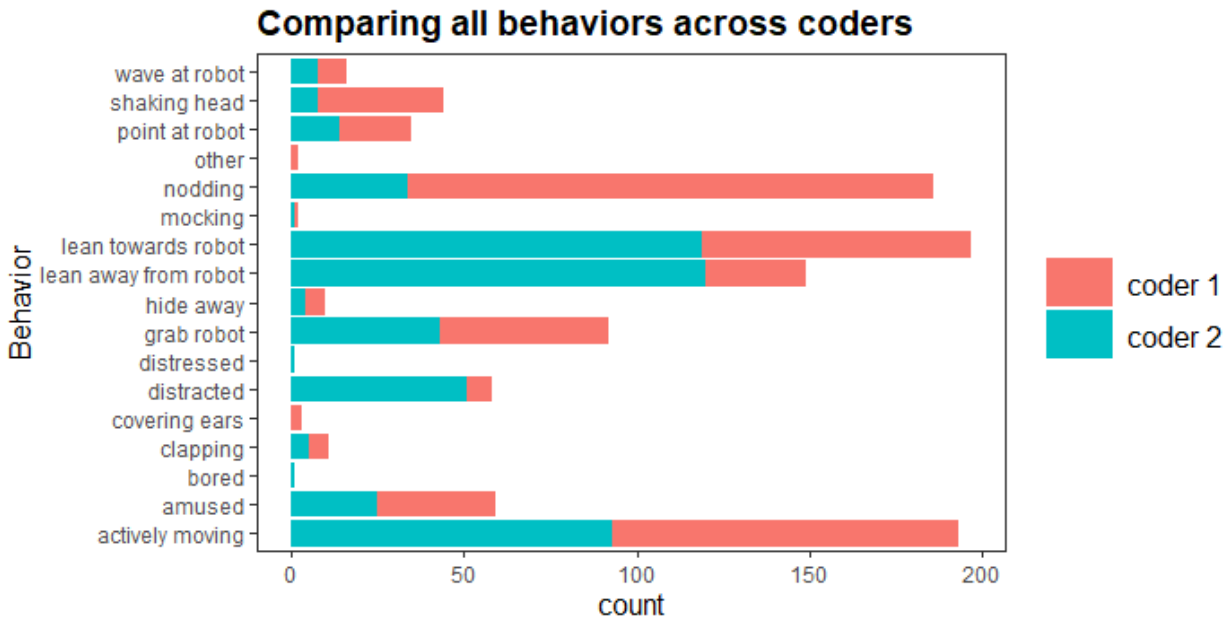


Figure 6. *Number of all behaviors for both coders*

The frontal view behaviors analysis demonstrated that the most frequent behaviors for both coders were “distracted” and “amused” (coder 1 – “distracted”: 7 occurrences [0.6% of total behaviors], “amused”: 34 occurrences [3% of total behaviors]; coder 2 – “distracted”: 51 occurrences [4.6% of total behaviors], “amused”: 25 occurrences [2.2% of total behaviors]). The least frequent behaviors were “bored”, “distressed” (0 occurrence) and “mocking” (1 occurrence) for coder 1 and “bored”, “distressed” and “mocking” for coder 2. Figure 7 illustrates the behavior counts for frontal view videos annotated by the coders. Each bar represents a coded behavior and longer bars indicate higher values. The distinct colors demonstrate the differences between the coders in regards of the annotated behaviors. The chart shows that „distracted” and „amused” had the highest behavior counts for both coders and „bored”, “distressed” and “mocking” had the lowest total counts. The frontal view behavior counts plots demonstrating the raw data (coder 1 frontal view n = 26, coder 2 frontal view n = 4) can be found in Appendix C.

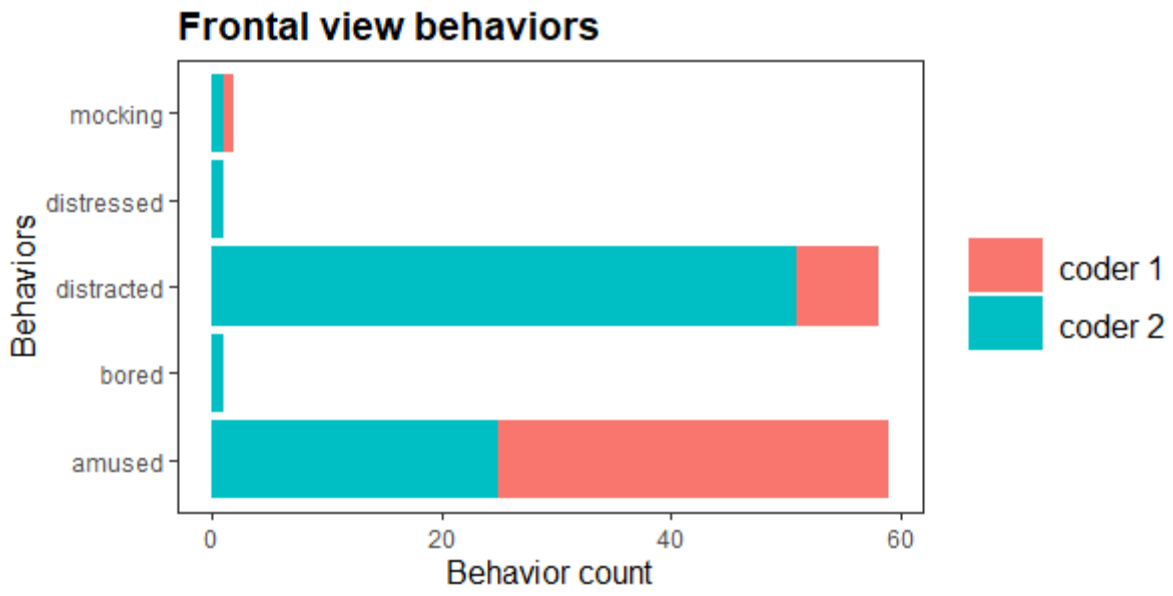


Figure 7. *Behavior counts of frontal view videos*

For the lateral view behaviors, the most frequent behavior was “nodding” (152 occurrences, 13.7% of total behaviors), “actively moving” (100 occurrences, 9% of total behaviors) and “leaning towards the robot” (120 occurrences, 7% of total behaviors) for coder 1 and “leaning away from the robot” (120 occurrences, 7% of total behaviors), “lean towards the robot” (119 occurrences, 7% of total behaviors) and “actively moving” (100 occurrences, 9%) for coder 2. The least frequent behaviors were “other” (3 occurrences, 0.3%), “covering ears” (3 occurrences, 0.3%), “clapping” (6 occurrences, 0.6%) and “hide away” (6 occurrences, 0.6%) for coder 1 and “hide away” (4 occurrences, 0.4%), “clapping” (5 occurrences, 0.45%), “shaking head” (6 occurrences, 0.6%) for coder 2. Behaviors with 0 occurrences were “covering ears” and “other” for coder 2. Figure 8 shows the behavior counts for lateral view videos annotated by the coders. Each bar represents a coded behavior and longer bars indicate higher values. The distinct colors demonstrate the differences between the coders in regards of the annotated behaviors (red: coder1, blue: coder 2). The chart shows commonalities between the two coders regarding highest behavior counts (“leaning towards the robot) and lowest behavior counts (“hide away”). The lateral view behavior counts plots demonstrating the raw data (coder 1 frontal view n = 26, coder 2 frontal view n = 4) can be found in Appendix C.

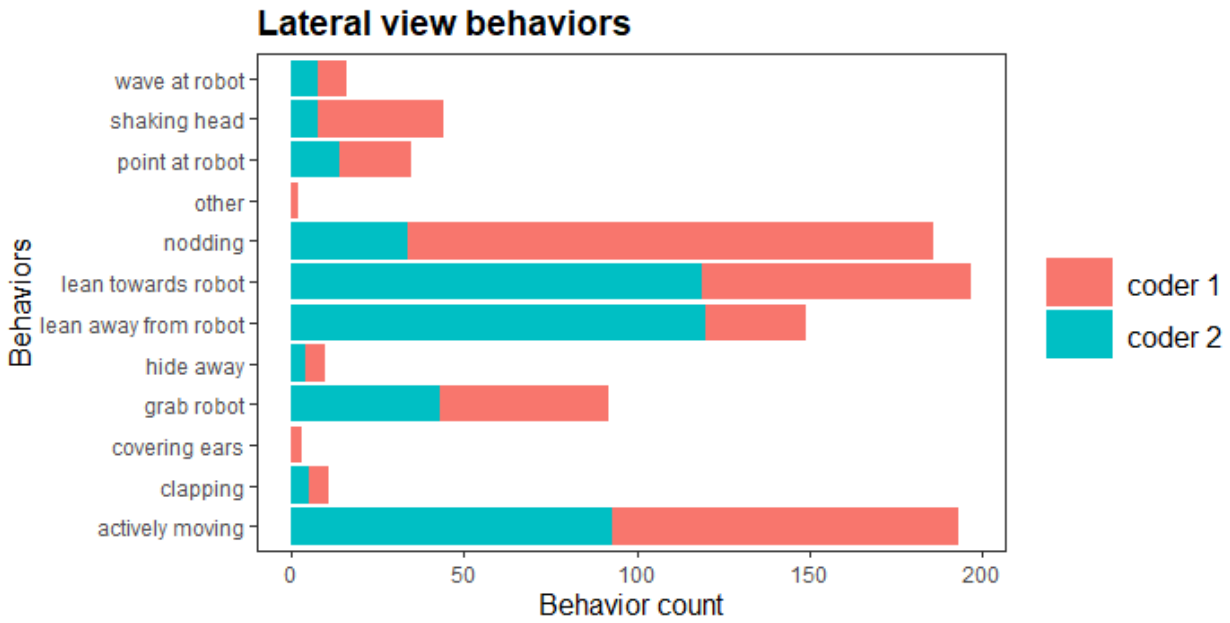


Figure 8. *Total behavior counts of lateral view videos*

The analysis of the durations showed that among state behaviors, “actively moving” (1594.92 s, 26.6% of summed total durations) and “amused” (855.37 s, 14.8% of summed total durations) had the longest total duration for coder 1 and “actively moving” (2468 s, 44.2% of summed total durations) and “distracted” (521 s, 9% of summed total durations) for coder 2. The shortest total durations were “mocking” (0.56 s) and “other” (10 s, 0.2% of summed total durations) for coder 1 and “mocking” (2.5 s, 0.04% of summed total durations) and “bored” (11 s, 0.2% of summed total durations) for coder 2. Figure 9 illustrates the total summed durations of the videos where each bar represents a state behavior and longer bars indicate longer total durations. The colors demonstrate differences between the coders (red: coder 1, blue: coder 2). “Actively moving” had the longest durations, while “mocking” showed the total shortest duration for both coders. The total summed durations of the raw data can be found in Appendix C.

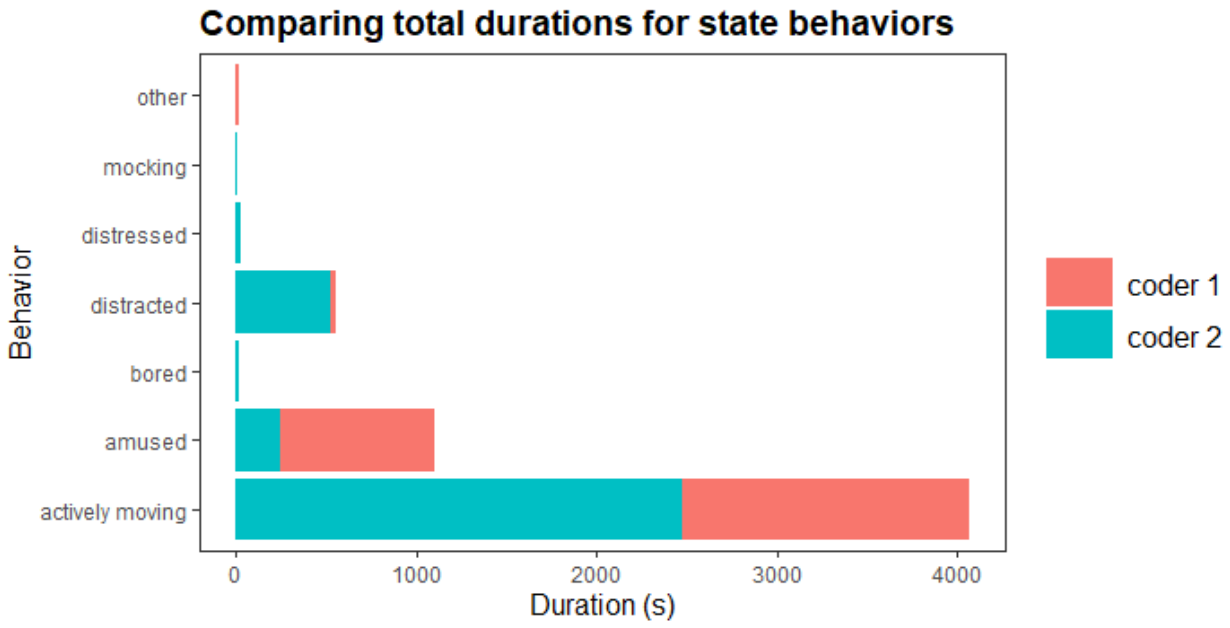


Figure 9. *Durations of state/dynamic behaviors*

5.2. Intracoder comparisons

5.2.1 Time occurrences of point behaviors

Figures 10 and 11 illustrate the time occurrences of point behaviors for coder 1 and coder 2 respectively. The x axes of the scatterplots indicate the (starting) times in seconds, while the y axes represent the point behaviors. The distinct colors demonstrate separate video lengths (red for recordings lasting longer than 800 s, blue for recordings lasting less than 600 s and green for videos lasting between 600 and 800 s).

Figure 10 demonstrates the timings of point behaviors across all subjects coded by coder 1 separated by video length. “Nodding” (30%) and “leaning towards the robot” (18%) was frequent throughout the sessions. “Shaking head” (7%) and “leaning away from robot” (6%) was also frequent throughout the interaction although these behaviors were less frequent. The least frequent behaviors were “covering ears” (0.6%), “hide away” (1.4%) and clapping (1.4%). “Grab robot” (12%) and “point at robot” (4.5%) were more frequent at the beginning of the video sessions, while “covering ears” (0.6%) and “clapping” (1.4%) were less frequent. Figure 11 demonstrates the timings of point behaviors for coder 2. Both “leaning towards” (27%) and “leaning away from robot” (27%) were frequent throughout the sessions as well as “nodding” (7.5%) and “grab robot” (9.5%). “Pointing” (3%) and “waving at the robot” (1.7%) also appeared throughout the sessions

but these behaviors were more common for “short length” videos. The least frequent behaviors were “hide away” (0.9%) and “clapping” (1.1%).

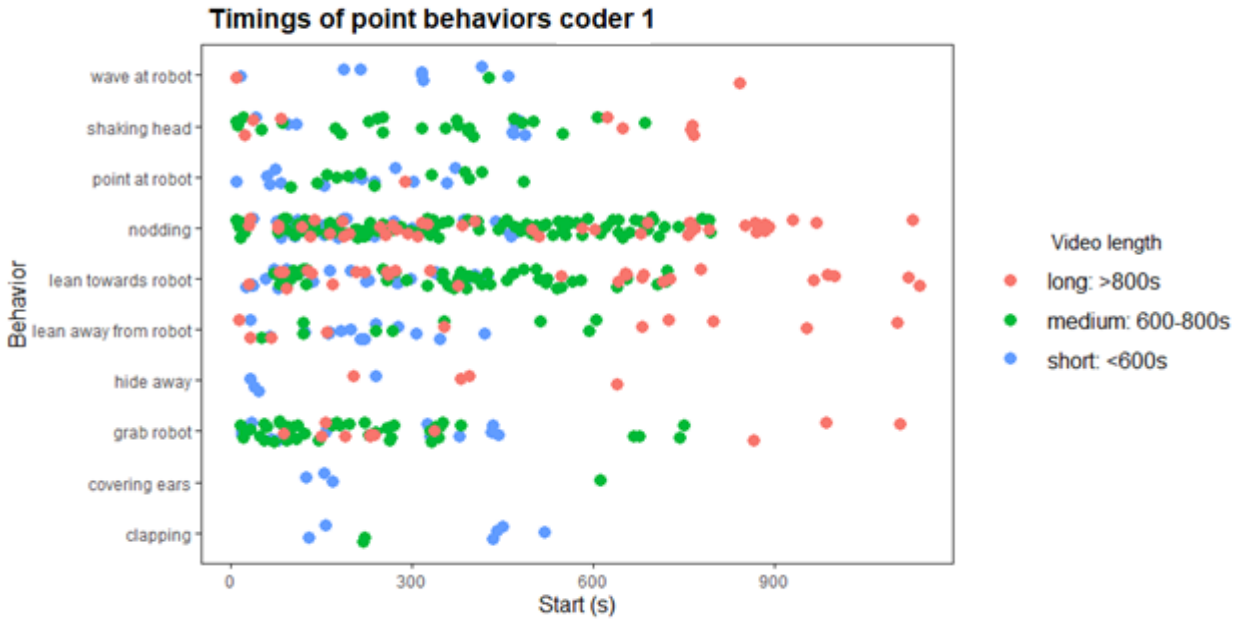


Figure 10. *Time occurrences of point behaviors: coder 1*

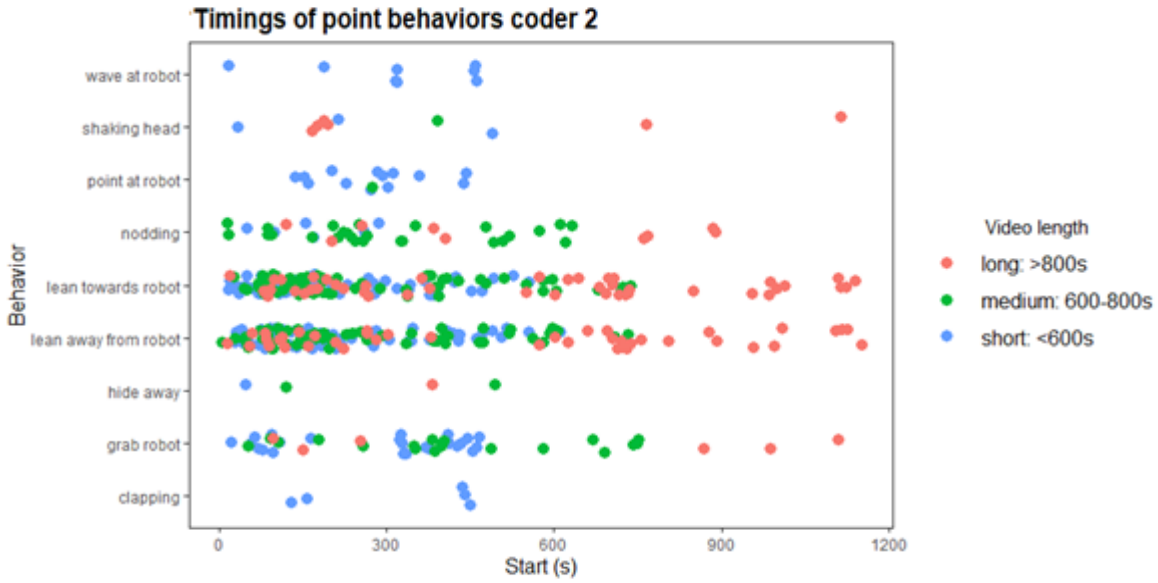


Figure 11. *Time occurrences of point behaviors: coder 2*

5.2.2 Behavior durations and start times of state behaviors

The start and endpoints of behavior annotations of coder 1 indicated that “actively moving” (M = 17.5 s, SD = 27 s) and “amused” (M = 20 s, SD = 48 s) generally lasted longer and occurred dispersed throughout the sessions. At the beginning of the sessions, “amused” had the longest duration and as the sessions proceeded, it gradually disappeared, while “actively moving” was most present during the whole sessions. “Surprised” (M = 6 s, SD = 8 s) and “mocking” (M = 4.5 s, SD = 5.5 s) were present occasionally for shorter periods as well as “other” (M = 9 s, SD = 16 s). “Other” indicate behaviors not presented in the protocol and coder 1 noted these behaviors as “talking to robot”, “catching falling robot” or “uncertain”. The findings of coder 2 indicate that “actively moving” (M = 23 s, SD = 66 s) appeared dispersed throughout the sessions and generally lasted a longer amount of time. “Amused” (M = 12 s, SD = 10 s) was mostly present in the beginning and lasted for shorter durations, while “distracted” (M = 15s, SD = 20 s) was more present in the beginning but lasted for longer durations. “Distressed” (30 s) and “bored” (11 s) were present once, for a short amount of time, at the beginning of sessions. Figure 12 illustrates the time occurrences and durations of the state behaviors. Each row represents one behavior, and the different colors indicate the two coders (red: coder 1, blue: coder 2)

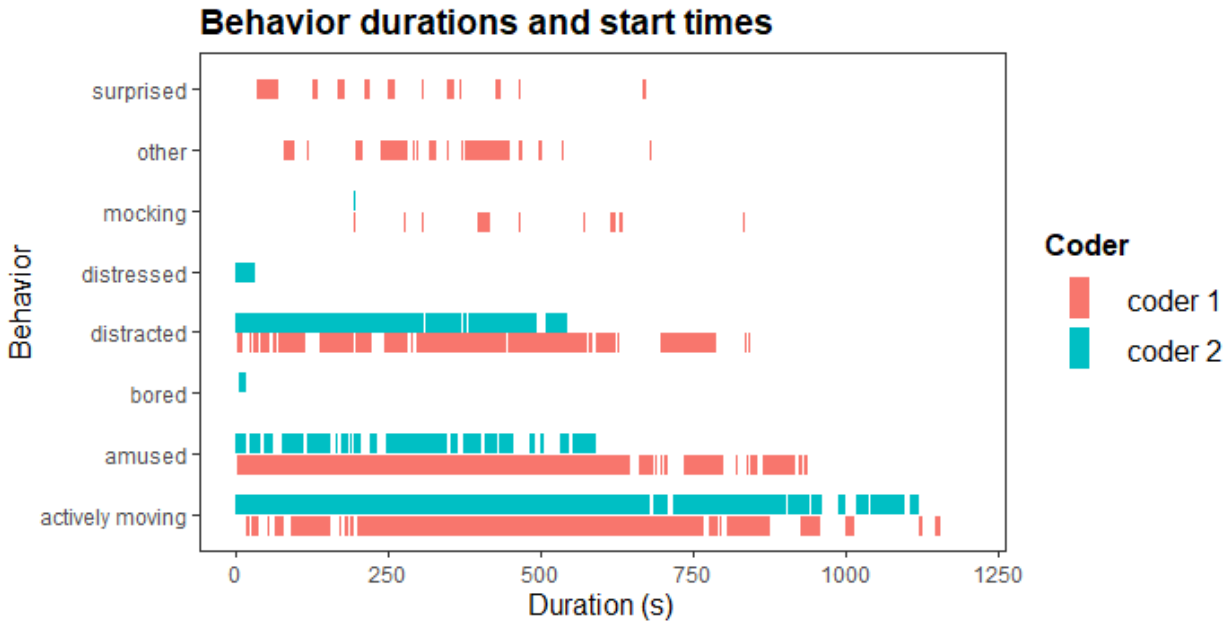


Figure 12. *Durations and start times of behaviors*

5.5. Case studies

5.5.3 Subject 008

Participant 008 is a 6.4-year-old child with unilateral CI, implanted when they were 1 year old. Their maximum speech score (65 dB, aided) was 49%. The anecdotal evidence (notes of the clinicians) indicated that the child is often aggressive, and they often struggle with audiometric tests. Before the beginning of the test, their parent was afraid that they were going to hit or break the robot. However, according to the notes, the child was showing affection towards the robot by kissing and hugging it after the session was over. On the video recording, the child seemed scared and startled when the robot moved for the first time (shouting, hiding away, looking at the parents). After a few seconds of the robot remaining motionless, the child appeared at ease, touched the robot and pointed at it. The participant was excited to take part in the NVA list and gave the robot their complete attention by looking at it and leaning in its direction. They often touched its head and showed emblems towards it (e.g. thumbs up). Moreover, they often imitated the head nodding gesture of the robot. Throughout the second NVA list, they continued to show curiosity and frequently waved, stroked and leaned in to gaze into the robot's eyes. The speech intelligibility score for the first test was 36% (stimuli presented at 60 dB) and 39% for the second test (stimuli presented at 75 dB). Figure 13 illustrates the behavior counts for this subject; each row represents one behavior. The most frequent behaviors were “grabbing the robot” (31 occurrences), “leaning towards the robot” (31 occurrences) and “leaning away from the robot” (27 occurrences) and the least frequent behavior was “mocking” (4 occurrences). Figure 14 illustrates the total behavior durations during the testing, each column represents one behavior. While “amused” behavior had the longest duration (211 s), “mocking” was annotated as the shortest (23 s).

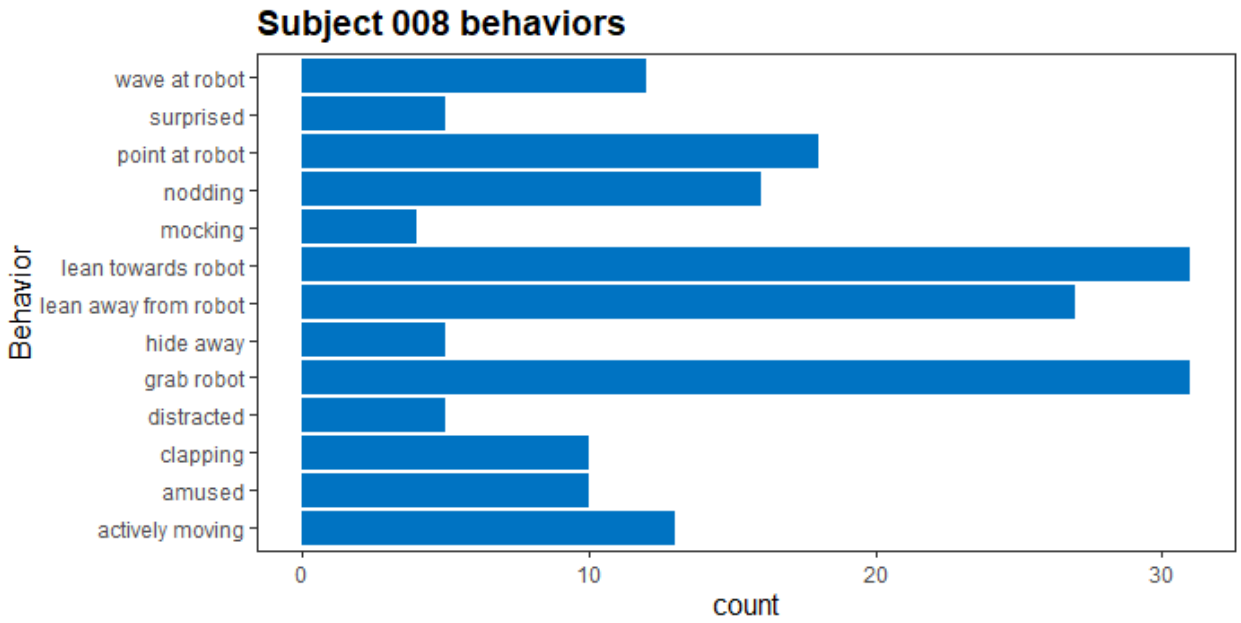


Figure 13. *Behavior counts for subject 008*

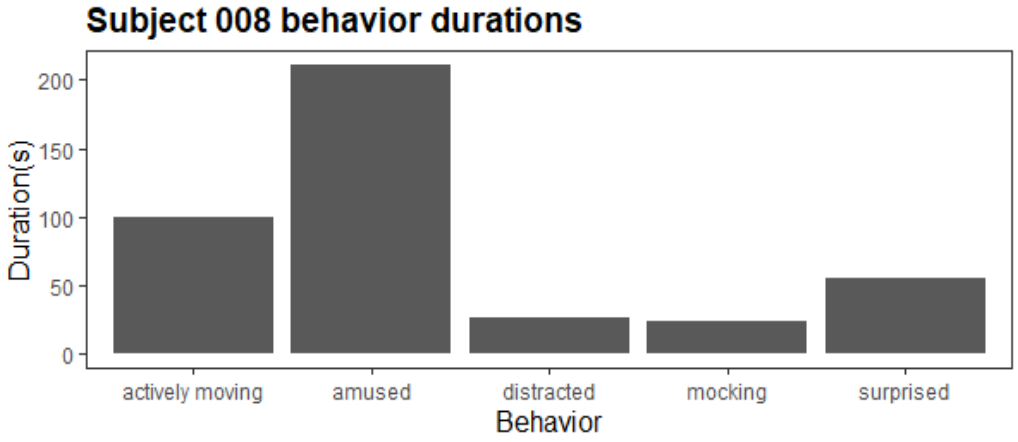


Figure 14. *Behavior durations: subject 008*

5.5.2 Subject 011

Subject 011 is a 12.7-year-old child with bilateral CIs. The child was 5 years old when they got their first CI and 7 when they got their second CI. The maximum speech score (65 dB, aided) of this child was 70%. The anecdotal evidence (notes from the speech therapist) indicated that they do not talk much in general, and the speech therapist was afraid they would not be able to perform the test. The video recording shows that the child was often smiling and pointing at the robot before

the session started. During the warble tones they were often distracted and talked to their parent. They also often looked at their parent. When the robot started to move, they seemed a bit startled and were leaning away from the robot while talking to their parent at the same time. They were often pointing to the robot and referring to it while talking. During the first NVA list, they were often smiling, laughing, and seemed amused and during the second NVA list they paid full attention to the robot by constantly looking at it. Both NVA lists were completed. The results of the first NVA list was: speech intelligibility score = 70% (stimuli presented at 70 dB), and for the second NVA list: speech intelligibility score = 76% (stimuli presented at 60 dB). Figure 15 illustrates the behavior counts for this subject (each row indicating one behavior). Both "pointing at the robot" and "amused" occurred 7 times and were the most common non-verbal cues. "Grab the robot" (1 occurrence) and "nodding" (1 occurrence) were the least frequent cues observed. Figure 16 shows the durations of annotated behaviors with each column indicating one behavior. "Amused" behavior had the longest duration in total (117 s) and "actively moving" had the shortest duration (6.5 s).

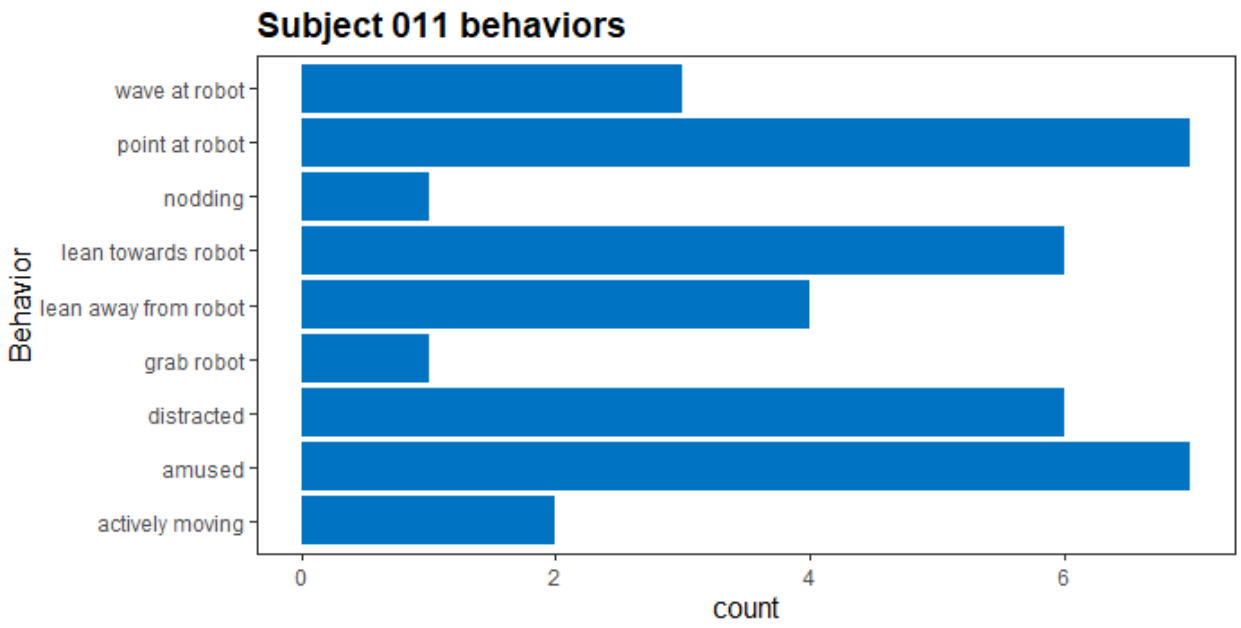


Figure 15. *Behavior counts: subject 011*

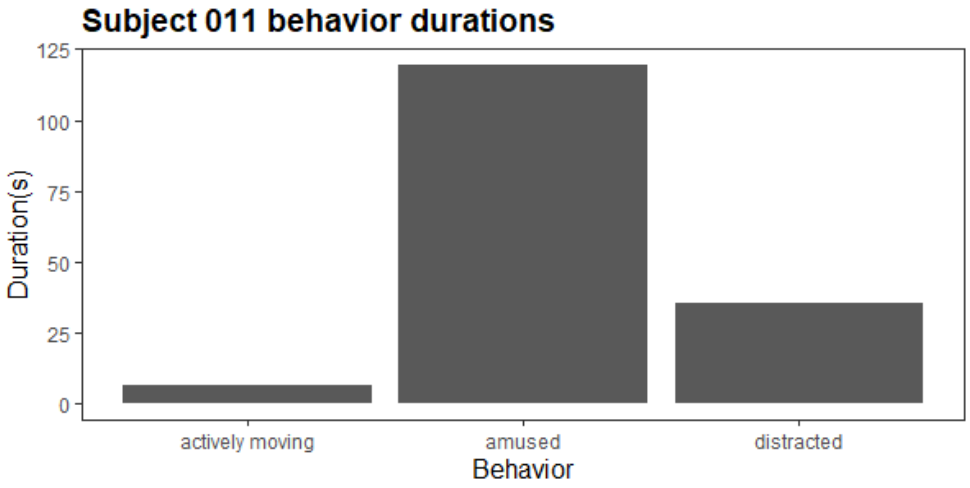


Figure 16. *Behavior durations: subject 011*

5.5.3 Subject 020

Subject 020 is 5.1-years-old child, and they have bilateral CIs. The maximum speech score (65 dB, aided) of this child was 60%. The clinical profile indicates agenesis of corpus callosum. The anecdotal evidence indicates that the child was able to replicate the vowels, but the session in general was not adequately successful. Notes from the previous audiometry sessions (before the child received their CI) indicated "uncontrollable behavior". During pure tone audiometry they were struggling to complete the test, however, four months after receiving the CIs (current session), the child was able to complete it. Figure 17 shows the behavior counts for the present session. The most frequent behavior was "grab robot" (12 occurrences) and "actively moving" (10 occurrences) and the least frequent was "hide away" (1 occurrence), point at robot (1 occurrence) and surprised (1 occurrence). Figure 18 demonstrates the behavior durations for subject 020. "Amused" behavior had the longest (418 s), while "surprised" had the shortest duration (2 s). The child was frequently seen touching the robot's legs and arms at the beginning of the video recording, and often looking to their parents for support (e.g. when the robot moved). They appeared enthusiastic to hear the warble tones and keen to hear all the sounds the robot could play. At the beginning of the NVA list they were often smiling, touching the robot, and pointing at it. However, when they were not able to repeat the words, they lost attention and talked to their parents. The child was not able to complete the NVA list for the first two times, but the third try was successful. When the third test ended, the robot displayed "boxing" motion. The child was initially confused, but when the robot

repeated the action, they became interested and smiled. The speech intelligibility score of the NVA list was 36% (stimuli presented at 75 dB).

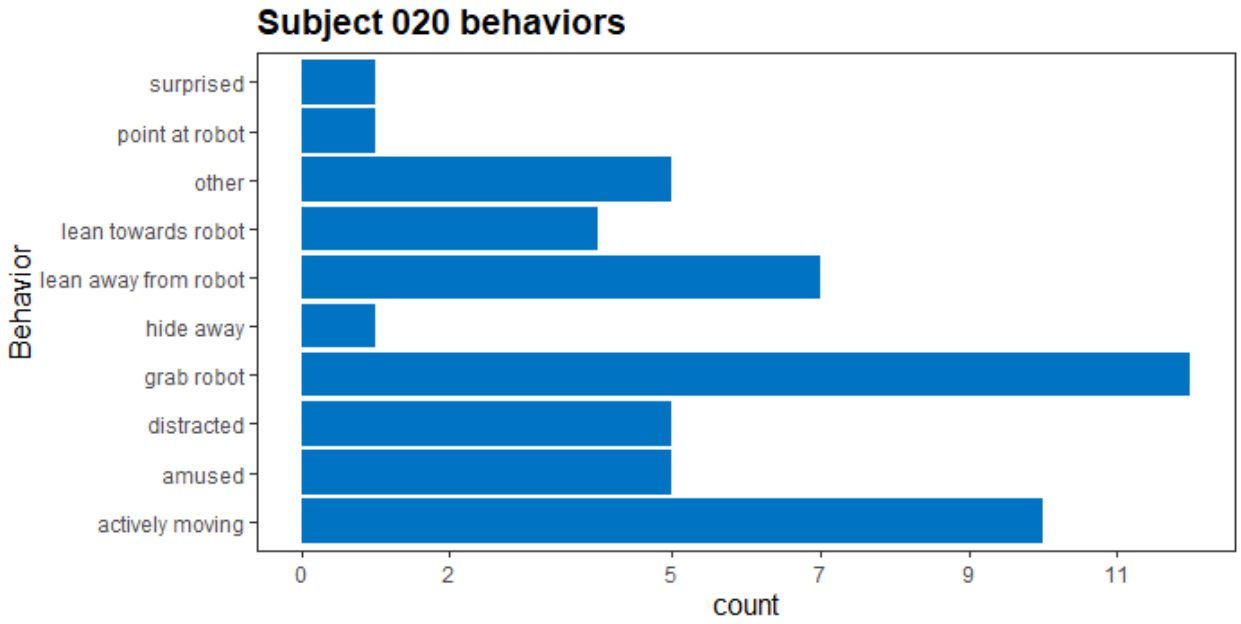


Figure 17. *Behavior counts for subject 020*

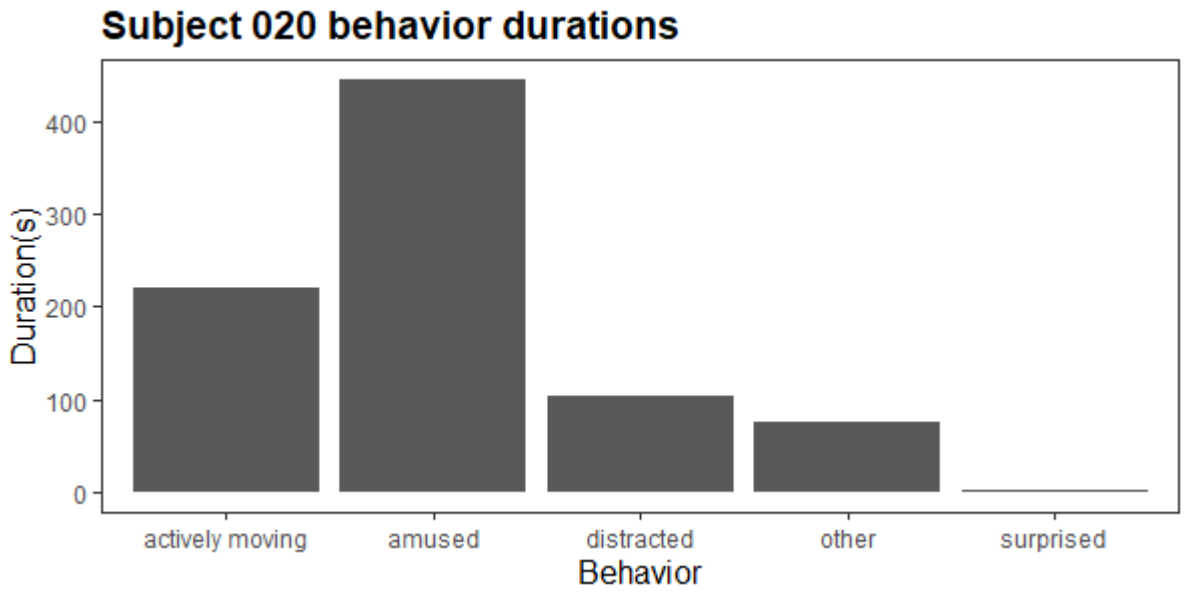


Figure 18. *Behavior durations for subject 020*

6. Discussion

6.1 Overview and the goal of the study

The primary aim of the current study was to develop a video protocol with standard scoring metrics of non-verbal cues to evaluate engagement levels in HRI during audiometry with children. This video protocol attempted to evaluate whether NAO is able to provide a pleasant experience for children with hearing loss by offering a scheme of easily detectable non-verbal cues. The subsequent sections will discuss the main findings according to the research objectives and the limitations.

6.2. Non-verbal cues as indication of engagement

The first research question was the following: What are the most and least common non-verbal cues that indicate engagement or no engagement towards the robot during auditory testing with children? The corresponding hypothesis for this research question was that the most common non-verbal cues would be engagement indicators while the least frequent ones would not be indicators of engagement (H1). Overall, the intercoder reliability for the listwise deleted data indicated moderate agreement (0.41-0.60) for all behaviors and substantial agreement (0.61-0.80) for the point behaviors (Cohen, 1960). State behaviors indicated no agreement (potentially due to the small sample size of frontal videos), therefore the findings regarding these non-verbal cues should be interpreted with caution. According to the annotations and the results of intercoder analysis, "leaning towards the robot" was the most frequent point behavior consistent across both coders (i.e. it had a high frequency count for both coders). “Leaning towards the robot” is when the child is moving their torso closer to the robot and have been linked to inclusion, affirmation, and attentional behaviors in human-human and human-robot interactions (van der Kooij et al., 2006; Johanson et al., 2019). Additionally, it was discovered that sign language communicators also exhibited these characteristics when displaying leaning forward behaviors (Wilbur & Patschke, 1998). Therefore, “leaning towards the robot” could be considered as the most frequent engagement indicator in the current setting. "Actively moving" was the most often annotated state behavior by both coders, and this behavior includes actions like dancing, moving, toying with one's body, stretching, itching, and fidgeting. In their pilot study with a social robot, Albo-Canals et al. (2018) regarded “walking around” and “playing with the body and other objects” as disengagement occurrences. Children were accompanied by the parents during the interaction session with the NAO, and in the videos,

children were sometimes seen walking towards them during the testing. This could provide an explanation for the high counts of “actively moving”. Additionally, the behaviors with the longest durations and relatively high frequencies were “actively moving”, “amused” and “distracted”. “Amused” is when the child is smiling, laughing and showing excitement. Smiling and laughing are considered as universally positive behaviors and indicators of engagement. This is consistent with earlier CRI studies demonstrating that smiling and laughing could reflect engagement and enjoyment (Leite et al., 2015; Young et al., 2011). “Distracted” was described as looking at the clinician, parent, other people or objects for a longer period of time. The parents were sitting next to the children during the testing and the tests were guided by the clinician therefore, the children were often interacting with and looking at other people throughout the session which could be reflected by the long durations of the annotated “distracted” cue.

The least frequent point behaviors appearing in the annotations of both coders were “hide away” and “clapping”. Hiding away is covering the face with hands/other parts of the body and could indicate that the child wishes to be invisible from others out of embarrassment, fear, or other emotions like playfulness. Therefore, “hiding away” is not generally considered to belong under the “engagement” category. “Clapping” is striking palms together repeatedly as a sign of applause or positive valence (Rudovic et al., 2017). The reason why this behavior was not frequent could be due to lack of engagement or due to sign language communicators using a different concept to display applause. This concept is “deaf applause” which is having both hands up in the air and twisting them a couple of times. The least frequent state behaviors were “bored” and “distressed”. “Bored” is when the child looking weary, being impatient, yawning, rolling eyes as an indicator of the lack of engagement. This behavior had the shortest total duration according to the annotations of both coders. “Distressed” was described as crying or throwing a tantrum and is universally regarded as an emotional disengagement behavior.

According to the video annotations, the findings show that positive behaviors and potential engagement indicators that were annotated frequently and/or had the longest total durations were the non-verbal cues "amused" and "leaning towards the robot”. However, it is challenging to draw a conclusion given that other behaviors, including "actively moving" and “distracted” which are not considered as engagement indicators were also rather common.

To provide an answer for the first research question, overall, potential engagement cues such "leaning toward the robot" and "amused" were seen more frequently during the sessions than

“hiding away" or "bored” which are not regarded as engagement indicators. However, the hypothesis was only partially confirmed and further work is certainly required to disentangle inconsistencies (such as the high count of behavior “actively moving” and “distracted”). A matched pairs study design involving children with normal hearing or auditory sessions without the NAO could help to provide a baseline for examining engagement.

6.3 Changes in non-verbal cues with time

The second research question aimed to investigate how the frequency of the non-verbal cues changes throughout the auditory sessions; in particular, comparing the beginning and the end of sessions. It was hypothesized that more engagement indicators would appear at the beginning of the video and less towards the end (H2). However, the findings from the intracoder analysis demonstrated that most annotated behaviors were dispersed throughout the video recordings and were rarely clustered at the beginning or at the end of sessions. Point behaviors “leaning towards the robot” and “nodding” were annotated to regularly occur and were dispersed throughout the videos according to the logs of both coders. “Nodding” is moving the head in an up-down motion as a sign of agreement or understanding and prior research demonstrated that it is related to the level of engagement (Inoue et al., 2016; Inoue et al., 2018). It is also considered as a sign of encouragement and affirmation towards the speaker (McClave, 2000). Given that the robot often nodded during the NVA lists and DIN tests, it is also possible that the children were imitating the robot during the tests. The behavior "grab the robot" regularly appeared at the beginning of the recordings and less frequently as the video progressed. It includes behaviors such as grasping the head, torso, arms and/or legs, or hugging/kissing/lifting the robot. Heerink et al. (2012) demonstrated that showing affection via physical contact during CRI has been regarded as an engaging, interactive activity. Since the auditory tests were preceded with warble tones where children were encouraged to touch the robot, it is not surprising that this behavior was more frequent at the beginning of the videos.

The behaviors “bored” and “distressed” only appeared rarely and mostly at the beginning of the interactions. This could indicate that even though some kids were not as interested in the robot at first, this pattern could have changed as the kids started interacting with it.

In regard to the research question, “grab the robot” behavior fluctuated the most, appearing most frequently towards the beginning of the videos compared to the end. In general, most of

behaviors were relatively consistent throughout the testing and “nodding” and “leaning towards the robot” appeared to be the most persistent behaviors according to the annotations. Therefore, the second hypothesis was only confirmed in the case of the “grab robot” behaviors but not for other engagement indicators.

6.4 Engagement in children with lower maximum bilateral speech scores

The third research question attempted to examine how the non-verbal cues differ for children with lower levels of maximum speech scores compared to the overall sample. The corresponding hypothesis suggested that the cues these children display would be comparable to the cues that are observed in the overall sample. Because three children's maximum bilateral speech scores (65 dB, aided) were below 80% (49%, 60%, 70%), the video recordings and the coders' annotations of these cases were evaluated to explore whether the video protocol could be applied in these videos and to evaluate the experiences with the NAO robot. All three children were successful in completing at least one of the NVA tests. Additionally, all children displayed “amused” behavior frequently and/or for longer periods of time. They were frequently smiling, laughing and actively interacting with the robot by either grabbing it, pointing to it or nodding towards it. Anecdotal evidence suggests that all three kids enjoyed their interactions with the robot.

To provide an answer to the research question, there were no evident patterns that suggest that non-verbal cues considerably differed according to the annotations and the anecdotal evidence. The most frequent cues such as “grab robot” or “amused” were often present in the majority of the videos. Even though the least frequent cues varied greatly between the three subjects, these cues were generally less common when looking at the annotations for all participants (e.g. in case of “hide away” or “mocking”). In summary, it is suggested that children with neurodivergent profile/less intelligibility scores could also benefit from the positive experiences the NAO robot could offer since they display similar (positive) behaviors as the other subjects.

6.5 Limitations

The present study included a relatively small study size of n = 26 for the point and n = 4 for the state behaviors. As a result, the findings of this study should be carefully interpreted and considered as indication to determine whether a video protocol could effectively identify child engagement during NAO-assisted audiometry. Additionally, intercoder reliability indicated no agreement for

the raw data (possibly due to the large number of missing data due to unequal frontal video sample sizes for coder 1 and coder 2), therefore the results should be interpreted with caution and as preliminary findings. Although the intercoder reliability of the listwise deleted data shows moderate reliability for all the behavior and the results from the case studies demonstrated that the non-verbal cues were corresponding to the anecdotal evidence and subjective observations. Another potential limitation is that the current study only included children with hearing loss without providing a baseline to measure engagement levels (i.e. comparison of engagement during speech audiometry testing without the NAO robot). Moreover, the parents actively assisted and engaged with the children during the hearing tests, thus the kids interacted with several individuals in addition to the NAO robot, which might have caused distraction. Finally, due to time restrictions and the additional cognitive load involved in answering questionnaires, the researchers were unable to provide the children with surveys to fill out regarding their interactions with the robot. As a result, this study's central emphasis is entirely exploratory and observational.

7. Conclusion

7.1 Future research

The goal of the current study, which is exploratory in nature, was to create a video protocol for examining nonverbal cues with children who have varying degrees of hearing loss during speech audiometry supported by NAO. All in all, the present video protocol seems to be a promising method to evaluate engagement in this context and the non-verbal cues of the most and least frequent behaviors were comparable across the annotations of the two coders. Additionally, it is a solid foundation for more advanced video schemes that aim to investigate engagement within the fields investigating speech audiometry and HRI. For instance, future research could assess whether the video scheme is sufficiently clear by recruiting three or more coders to examine their agreement. Additionally, a discussion could be implemented to explore disagreements and the reasons behind these disagreements. Furthermore, results showed that the behaviors “mocking”, “hide away” and “actively moving” could be described in a more straightforward way to better distinguish between levels of engagement.

The results of the annotations suggested that NAO could be a promising assistant to clinicians during speech audiometry since children (both with lower and higher maximum speech scores) often showed engagement indicator cues during the interaction. A follow-up study applying

qualitative data triangulation, including quantitative methods as well as qualitative notes, transcripts, and subjective insights from the children and/or patents could provide novel perspectives of the experiences children face during these interactions.

References

Abdel-Latif, K. H. A., & Meister, H. (2022). Speech Recognition and Listening Effort in Cochlear Implant Recipients and Normal-Hearing Listeners. *Frontiers in Neuroscience*, *15*. https://www.frontiersin.org/articles/10.3389/fnins.2021.725412

Abdul Malik, N., Yussof, H., Hanapiah, F. A., & Anne, S. J. (2014). Human Robot Interaction (HRI) between a humanoid robot and children with Cerebral Palsy: Experimental framework and measure of engagement. *2014 IEEE Conference on Biomedical Engineering and Sciences (IECBES)*, 430–435. https://doi.org/10.1109/IECBES.2014.7047536

Albo-Canals, J., Martelo, A. B., Relkin, E., Hannon, D., Heerink, M., Heinemann, M., Leidl, K., & Bers, M. U. (2018). A Pilot Study of the KIBO Robot in Children with Severe ASD. *International Journal of Social Robotics*, *10*(3), 371–383. https://doi.org/10.1007/s12369-018-0479-2

Argyle, M., & Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry*, *28*(3), 289–304. https://doi.org/10.2307/2786027

Armstrong, A. G., Lam, C. C., Sabesan, S., & Lesica, N. A. (2022). Compression and amplification algorithms in hearing aids impair the selectivity of neural responses to speech. *Nature Biomedical Engineering*, *6*(6), 717–730. https://doi.org/10.1038/s41551-021-00707-y

Başkent, D., Gaudrain, E., Tamati, T., & Wagner, A. (2016). *Perception and Psychoacoustics of Speech in Cochlear Implant Users*.

Belpaeme, T., Baxter, P., Read, R., Wood, R., Cuayahuitl, H., Kiefer, B., Racioppa, S., Kruijff-Korbayova, I., Athanasopoulos, G., Enescu, V., Looije, R., Neerincx, M., Demiris, Y., Ros, R., Beck, A., Cañamero, L., Hiolle, A., Lewis, M., Baroni, I., & Humbert, R. (2012). Multimodal

Child-Robot Interaction: Building Social Bonds. *Journal of Human-Robot Interaction*, *1*, 33–53. https://doi.org/10.5898/JHRI.1.2.Belpaeme

Bethel, C. L., Salomon, K., Murphy, R., & Burke, J. (2007). Survey of Psychophysiology Measurements Applied to Human-Robot Interaction. *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*. https://doi.org/10.1109/ROMAN.2007.4415182

Bethel, C., & Murphy, R. (2010). Review of Human Studies Methods in HRI and Recommendations. *I. J. Social Robotics*, *2*, 347–359. https://doi.org/10.1007/s12369-010-0064-9

Blanson Henkemans, O., Bierman, B., Janssen, J., Neerincx, M., Looije, R., Bosch, H., & Giessen, J. (2015). A personalized robot contributing to enjoyment and health knowledge of children with diabetes at the clinic: A pilot study. *Nederlands Tijdschrift Voor Diabetologie*, *10*, 169–169. https://doi.org/10.1007/s12467-012-0132-x

Bosman, A. J. (1989). *Speech perception by the hearing impaired*.

Breazeal, C. (2003). Social Interactions in HRI: The Robot View. *IEEE Trans. on Man, Cybernetics, and Sys – Part C*.

Breazeal, C. (2011). Social robots for health applications. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 5368–5371. https://doi.org/10.1109/IEMBS.2011.6091328

Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., Lieberman, J., Lockerd, A., & Mulanda, D. (2004). Humanoid robots as cooperative partners for people. *International Journal of Humanoid Robots*, *1*.

Carradore, M. (2022). People's Attitudes Towards the Use of Robots in the Social Services: A Multilevel Analysis Using Eurobarometer Data. *International Journal of Social Robotics*, *14*(3), 845–858. https://doi.org/10.1007/s12369-021-00831-4

Castellano, G., Pereira, A., Leite, I., Paiva, A., & McOwan, P. W. (2009). Detecting user engagement with a robot companion using task and social interaction-based features. *Proceedings of the 2009 International Conference on Multimodal Interfaces*, 119–126. https://doi.org/10.1145/1647314.1647336

Choudhury, A., Li, H., M Greene, C., & Perumalla, S. (2018). Humanoid Robot-Application and Influence. *Archives of Clinical and Biomedical Research*, *02*(06). https://doi.org/10.26502/acbr.50170059

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Dautenhahn, K., Bond, A., Cañamero, L., & Edmonds, B. (2006). Socially Intelligent Agents. In *Ai Magazine—AIM* (Vol. 19, pp. 1–20). https://doi.org/10.1007/0-306-47373-9_1

Dautenhahn, K., & Werry, I. (2002a). A quantitative technique for analysing robot-human interactions. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, *2*, 1132–1138 vol.2. https://doi.org/10.1109/IRDS.2002.1043883

Dautenhahn, K., & Werry, I. (2002b). A quantitative technique for analysing robot-human interactions. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, *2*, 1132–1138 vol.2. https://doi.org/10.1109/IRDS.2002.1043883

DeBow, A., & Green, W. B. (2000). *A Survey of Canadian Audiological Practices: Pure Tone and Speech Audiometry*. 9.

Feil-Seifer, D., & Mataric, J. M. (2005). Socially Assistive Robotics. *Proceedings of the 2005 IEEE*, 4.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*(3–4), 143–166. https://doi.org/10.1016/S0921-8890(02)00372-X

Fong, T., Thorpe, C., & Baur, C. (2003). Collaboration, Dialogue, Human-Robot Interaction. In R. A. Jarvis & A. Zelinsky (Eds.), *Robotics Research* (pp. 255–266). Springer. https://doi.org/10.1007/3-540-36460-9_17

Friard, O., & Gamba, M. (2016). BORIS: A free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution*, *7*(11), 1325–1330. https://doi.org/10.1111/2041-210X.12584

Giuliani, M., Mirnig, N., Stollnberger, G., Stadler, S., Buchner, R., & Tscheligi, M. (2015). Systematic analysis of video data from different human–robot interaction studies: A categorization of social signals during error situations. *Frontiers in Psychology*, *6*. https://www.frontiersin.org/article/10.3389/fpsyg.2015.00931

Heerink, M., Boladeras, M., Albo-Canals, J., Angulo, C., Barco, A., Casacuberta, J., & Garriga, C. (2012, September 1). *A field study with primary school children on perception of social presence and interactive behavior with a pet robot*. Proceedings - IEEE International Workshop on Robot and Human Interactive Communication. https://doi.org/10.1109/ROMAN.2012.6343887

Hegel, F., Muhl, C., Wrede, B., Hielscher-Fastabend, M., & Sagerer, G. (2009). Understanding Social Robots. *2009 Second International Conferences on Advances in Computer-Human Interactions*, 169–174. https://doi.org/10.1109/ACHI.2009.51

Henkemans, O. A. B., Bierman, B. P. B., Janssen, J., Looije, R., Neerincx, M. A., van Dooren, M. M. M., de Vries, J. L. E., van der Burg, G. J., & Huisman, S. D. (2017). Design and evaluation of a personal robot playing a self-management education game with children with diabetes type 1. *International Journal of Human-Computer Studies*, *106*, 63–76. https://doi.org/10.1016/j.ijhcs.2017.06.001

Hindley, P. (1997). Psychiatric aspects of hearing impairments. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *38*(1), 101–117. https://doi.org/10.1111/j.1469-7610.1997.tb01507.x

Hughes-Roberts, T., Brown, D., Standen, P., Desideri, L., Negrini, M., Rouame, A., Malavasi, M., Wager, G., & Hasson, C. (2019). Examining engagement and achievement in learners with individual needs through robotic-based teaching sessions. *British Journal of Educational Technology*, *50*(5), 2736–2750. https://doi.org/10.1111/bjet.12722

Inoue, K., Lala, D., Nakamura, S., Takanashi, K., & Kawahara, T. (2016). Annotation and analysis of listener's engagement based on multi-modal behaviors. *Proceedings of the Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, 25–32. https://doi.org/10.1145/3011263.3011271

Inoue, K., Lala, D., Takanashi, K., & Kawahara, T. (2018). Engagement Recognition in Spoken Dialogue via Neural Network by Aggregating Different Annotators' Models. *Interspeech 2018*, 616–620. https://doi.org/10.21437/Interspeech.2018-2067

Ioannou, A., & Andreeva, A. (2019). *Play and Learn with an Intelligent Robot: Enhancing the Therapy of Hearing-Impaired Children* (pp. 436–452). https://doi.org/10.1007/978-3-030-29384-0_27

Ismail, L. I., Shamsudin, S., Yussof, H., Hanapiah, F. A., & Zahari, N. I. (2012). Robot-based Intervention Program for Autistic Children with Humanoid Robot NAO: Initial Response in Stereotyped Behavior. *Procedia Engineering*, *41*, 1441–1447. https://doi.org/10.1016/j.proeng.2012.07.333

Jang, M., Park, C., Yang, H.-S., Kim, J.-H., Cho, Y.-J., Lee, D.-W., Cho, H.-K., Kim, Y.-A., Chae, K., & Ahn, B.-K. (2014). Building an automated engagement recognizer based on video analysis.

Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, 182–183. https://doi.org/10.1145/2559636.2563687

Johanson, D. L., Ahn, H. S., MacDonald, B. A., Ahn, B. K., Lim, J., Hwang, E., Sutherland, C. J., & Broadbent, E. (2019). The Effect of Robot Attentional Behaviors on User Perceptions and Behaviors in a Simulated Health Care Interaction: Randomized Controlled Trial. *Journal of Medical Internet Research*, *21*(10), e13667. https://doi.org/10.2196/13667

Kabacińska, K., Prescott, T. J., & Robillard, J. M. (2021). Socially Assistive Robots as Mental Health Interventions for Children: A Scoping Review. *International Journal of Social Robotics*, *13*(5), 919–935. https://doi.org/10.1007/s12369-020-00679-0

Kidd, C., & Breazeal, C. (2005). *Human-robot interaction experiments: Lessons learned*.

Koay, K., Dautenhahn, K., Woods, S., & Walters, M. (2006). *Empirical results from using a comfort level device in human-robot interaction studies*. *2006*, 194–201. https://doi.org/10.1145/1121241.1121276

Kompatsiari, K., Ciardo, F., De Tommaso, D., & Wykowska, A. (2019). Measuring engagement elicited by eye contact in Human-Robot Interaction. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6979–6985. https://doi.org/10.1109/IROS40897.2019.8967747

Kyrarini, M., Lygerakis, F., Rajavenkatanarayanan, A., Sevastopoulos, C., Nambiappan, H. R., Chaitanya, K. K., Babu, A. R., Mathew, J., & Makedon, F. (2021). A Survey of Robots in Healthcare. *Technologies*, *9*(1), 8. https://doi.org/10.3390/technologies9010008

Lala, D., Inoue, K., Milhorat, P., & Kawahara, T. (2017). Detection of social signals for recognizing engagement in human-robot interaction. *ArXiv:1709.10257 [Cs]*. http://arxiv.org/abs/1709.10257

Leite, I., McCoy, M., Ullman, D., Salomons, N., & Scassellati, B. (2015). Comparing Models of Disengagement in Individual and Group Interactions. *Proceedings of the Tenth Annual*

ACM/IEEE International Conference on Human-Robot Interaction, 99–105. https://doi.org/10.1145/2696454.2696466

Lytridis, C., Bazinas, C., Papakostas, G., & Kaburlasos, V. (2020). *On Measuring Engagement Level During Child-Robot Interaction in Education* (pp. 3–13). https://doi.org/10.1007/978-3-030-26945-6_1

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, *7*(3), 297–337. https://doi.org/10.1075/is.7.3.03mac

Maclennan-Smith, F., Swanepoel, D. W., & Hall, J. W. (2013). Validity of diagnostic pure-tone audiometry without a sound-treated environment in older adults. *International Journal of Audiology*, *52*(2), 66–73. https://doi.org/10.3109/14992027.2012.736692

McClave, E. Z. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, *32*(7), 855–878. https://doi.org/10.1016/S0378-2166(99)00079-X

Mori, M. (1970). The uncanny valley. *Energy*, *7*(4), 33–35.

Murphy, R. R., & Schreckenghost, D. (2013). Survey of metrics for human-robot interaction. *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 197–198. https://doi.org/10.1109/HRI.2013.6483569

Nalin, M., Baroni, I., Kruijff-Korbayová, I., Cañamero, L., Lewis, M., Beck, A., Cuayáhuitl, H., & Sanna, A. (2012). *Children's Adaptation in Multi-session Interaction with a Humanoid Robot*. IEEE. https://doi.org/10.1109/ROMAN.2012.6343778

Nilsson, N. J. (1984). Shakey The Robot. *SRI International*, 149.

Oertel, C., Castellano, G., Chetouani, M., Nasir, J., Obaid, M., Pelachaud, C., & Peters, C. (2020). Engagement in Human-Agent Interaction: An Overview. *Frontiers in Robotics and AI*, *7*. https://www.frontiersin.org/article/10.3389/frobt.2020.00092

Qidwai, U., Kashem, S. B. A., & Conor, O. (2020). Humanoid Robot as a Teacher's Assistant: Helping Children with Autism to Learn Social and Academic Skills. *Journal of Intelligent & Robotic Systems, 98*(3–4), 759–770. https://doi.org/10.1007/s10846-019-01075-1

Read, J., MacFarlane, S., & Casey, C. (2009). Endurability, Engagement and Expectations: Measuring Children's Fun. *Interaction Design and Children*.

Reeves, B., & Nass, C. (1997). The media equation: How people treat computers, television, and new media like real people and places. *Choice Reviews Online*, *34*(07), 34-3702-34–3702. https://doi.org/10.5860/CHOICE.34-3702

Riek, L. (2012). Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, 119–136. https://doi.org/10.5898/JHRI.1.1.Riek

Robaczewski, A., Bouchard, J., Bouchard, K., & Gaboury, S. (2021). Socially Assistive Robots: The Specific Case of the NAO. *International Journal of Social Robotics, 13*(4), 795–831. https://doi.org/10.1007/s12369-020-00664-7

Rousseau, V., Ferland, F., Létourneau, D., & Michaud, F. (2013). Sorry to Interrupt, But May I Have Your Attention? Preliminary Design and Evaluation of Autonomous Engagement in HRI. *Journal of Human-Robot Interaction*, *2*(3), 41–61. https://doi.org/10.5898/JHRI.2.3.Rousseau

Rudovic, O., Lee, J., Mascarell-Maricic, L., Schuller, B. W., & Picard, R. W. (2017). Measuring Engagement in Robot-Assisted Autism Therapy: A Cross-Cultural Study. *Frontiers in Robotics and AI*, *4*, 36. https://doi.org/10.3389/frobt.2017.00036

Rueben, M., Elprama, S. A., Chrysostomou, D., & Jacobs, A. (2020). Introduction to (Re)Using Questionnaires in Human-Robot Interaction Research. In C. Jost, B. Le Pévédic, T. Belpaeme, C. Bethel, D. Chrysostomou, N. Crook, M. Grandgeorge, & N. Mirnig (Eds.), *Human-Robot Interaction: Evaluation Methods and Their Standardization* (pp. 125–144). Springer International Publishing. https://doi.org/10.1007/978-3-030-42307-0_5

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2011). *Effects of Gesture on the Perception of Psychological Anthropomorphism: A Case Study with a Humanoid Robot*. *7072*, 31–41. https://doi.org/10.1007/978-3-642-25504-5_4

Salvini, P., Nicolescu, M., & Ishiguro, H. (2011). Benefits of Human—Robot Interaction [TC Spotlight]. *IEEE Robotics Automation Magazine*, *18*(4), 98–99. https://doi.org/10.1109/MRA.2011.943237

Sanchez-Lopez, R., Dau, T., & Jepsen, M. (2020, April 21). *Hearing-aid settings in connection to supra-threshold auditory processing deficits*. https://doi.org/10.13140/RG.2.2.23585.35681

Serholt, S., & Barendregt, W. (2016, October 27). *Robots Tutoring Children: Longitudinal Evaluation of Social Engagement in Child-Robot Interaction*. https://doi.org/10.1145/2971485.2971536

Shamsuddin, S., Yusoff, H., Ismail, L., Mohamed, S., Hanapiah, F., & Zahari, N. (2012). Initial Response in HRI- a Case Study on Evaluation of Child with Autism Spectrum Disorders Interacting with a Humanoid Robot NAO. *Procedia Engineering 2012;41:1448-1455*, *41*, 1448–1455. https://doi.org/10.1016/j.proeng.2012.07.334

Sheridan, T. (1997). *Eight ultimate challenges of human-robot communication*. 9–14. https://doi.org/10.1109/ROMAN.1997.646944

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., & Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, *166*(1), 140–164. https://doi.org/10.1016/j.artint.2005.03.005

Smits, C., Kapteyn, T. S., & Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International Journal of Audiology*, *43*(1), 15–28. https://doi.org/10.1080/14992020400050004

Smits, C., Theo Goverts, S., & Festen, J. M. (2013). The digits-in-noise test: Assessing auditory speech recognition abilities in noise. *The Journal of the Acoustical Society of America*, *133*(3), 1693–1706. https://doi.org/10.1121/1.4789933

Softbank, Robotics. (2015). *NAO the humanoid and programmable robot*.

Speck, D., Dornhege, C., & Burgard, W. (2017). Shakey 2016—How Much Does it Take to Redo Shakey the Robot? *IEEE Robotics and Automation Letters*, 1–1. https://doi.org/10.1109/LRA.2017.2665694

Spirrov, D., Van Eeckhoutte, M., Van Deun, L., & Francart, T. (2018). Real-time loudness normalisation with combined cochlear implant and hearing aid stimulation. PLoS ONE, 13(4), e0195412. https://doi.org/10.1371/journal.pone.0195412

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., & Goodrich, M. (2006). Common metrics for human-robot interaction. *Proceeding of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction - HRI '06*, 33. https://doi.org/10.1145/1121241.1121249

Takeda, H., Kobayashi, N., Matsubara, Y., & Nishida, T. (1997). *Towards Ubiquitous Human-Robot Interaction*.

Thepsoonthorn, C., Ogawa, K., & Miyake, Y. (2021). The Exploration of the Uncanny Valley from the Viewpoint of the Robot's Nonverbal Behaviour. *International Journal of Social Robotics*, *13*, 1–13. https://doi.org/10.1007/s12369-020-00726-w

Uluer, P., Kose, H., Gumuslu, E., & Barkana, D. E. (2021). Experience with an Affective Robot Assistant for Children with Hearing Disabilities. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-021-00830-5

Vacharkulksemsuk, T., & Fredrickson, B. L. (2012). Strangers in sync: Achieving embodied rapport through shared movements. *Journal of Experimental Social Psychology*, *48*(1), 399–402. https://doi.org/10.1016/j.jesp.2011.07.015

Van den Borre, E., Denys, S., van Wieringen, A., & Wouters, J. (2021). The digit triplet test: A scoping review. *International Journal of Audiology*, *60*(12), 946–963. https://doi.org/10.1080/14992027.2021.1902579

van der Kooij, E., Crasborn, O., & Emmerik, W. (2006). Explaining prosodic body leans in Sign Language of the Netherlands: Pragmatics required. *Journal of Pragmatics*, *38*(10), 1598–1614. https://doi.org/10.1016/j.pragma.2005.07.006

Vanpoucke, F., De Sloovere, M., & Plasmans, A. (2022). The Thomas More Lists: A Phonemically Balanced Dutch Monosyllabic Speech Audiometry Test. *Audiology Research*, *12*(4), 404–413. https://doi.org/10.3390/audiolres12040041

Veispak, A., Jansen, S., Ghesquière, P., & Wouters, J. (2015). Speech audiometry in Estonia: Estonian words in noise (EWIN) test. *International Journal of Audiology*, *54*(8), 573–578. https://doi.org/10.3109/14992027.2015.1015688

Vermeulen, A., De Raeve, L., Langereis, M., & Snik, A. (2012). Changing Realities in the Classroom for Hearing-Impaired Children with Cochlear Implant. *Deafness & Education International*, *14*(1), 36–47. https://doi.org/10.1179/1557069X12Y.0000000004

Vroegop, J., Rodenburg-Vlot, M., Goedegebure, A., Doorduin, A., Homans, N., & van der Schroeff, M. (2021). The Feasibility and Reliability of a Digits-in-Noise Test in the Clinical Follow-Up of Children With Mild to Profound Hearing Loss. *Ear and Hearing*, *42*(4), 973–981. https://doi.org/10.1097/AUD.0000000000000989

Walker, J. J., Cleveland, L. M., Davis, J. L., & Seales, J. S. (2013). Audiometry Screening and Interpretation. *American Family Physician*, *87*(1), 41–47.

Wilbur, R. B., & Patschke, C. G. (1998). Body leans and the marking of contrast in American sign language. *Journal of Pragmatics*, *30*(3), 275–303. https://doi.org/10.1016/S0378-2166(98)00003-4

World Health Organization. (2018). *Ear and hearing care: Indicators for monitoring provision of services*. World Health Organization. https://apps.who.int/iris/handle/10665/324936

Young, J., Sung, J.-Y., Voida, A., Sharlin, E., Igarashi, T., Christensen, H., & Grinter, R. (2011). Evaluating Human-Robot Interaction: Focusing on the Holistic Interaction Experience. *I. J. Social Robotics*, *3*, 53–67. https://doi.org/10.1007/s12369-010-0081-8

Zhou, N., Dixon, S., Zhu, Z., Dong, L., & Weiner, M. (2020). Spectrotemporal Modulation Sensitivity in Cochlear-Implant and Normal-Hearing Listeners: Is the Performance Driven by Temporal or Spectral Modulation Sensitivity? *Trends in Hearing*, *24*, 2331216520948385. https://doi.org/10.1177/2331216520948385

Appendix A – Data tables of the raw data

Total number of coded behaviors

Coded behaviors	N (coder 1)	N(coder 2)	N (total)
Total number of state and point behaviors			
State (coder 1: n = 4, coder 2: n = 26)	488	241	729
Point (coder 1 & 2: n = 26)	388	355	743

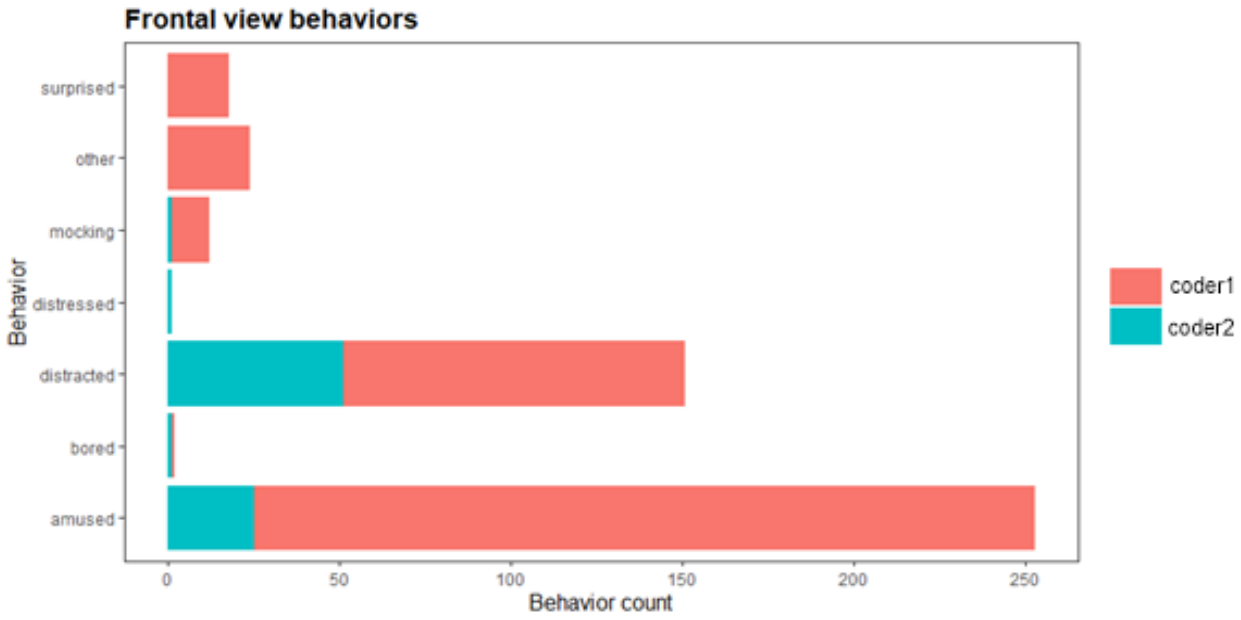
Appendix B – Counts of non-verbal cues

Counts of coded non-verbal cues (frontal view only includes videos coded by both coders)

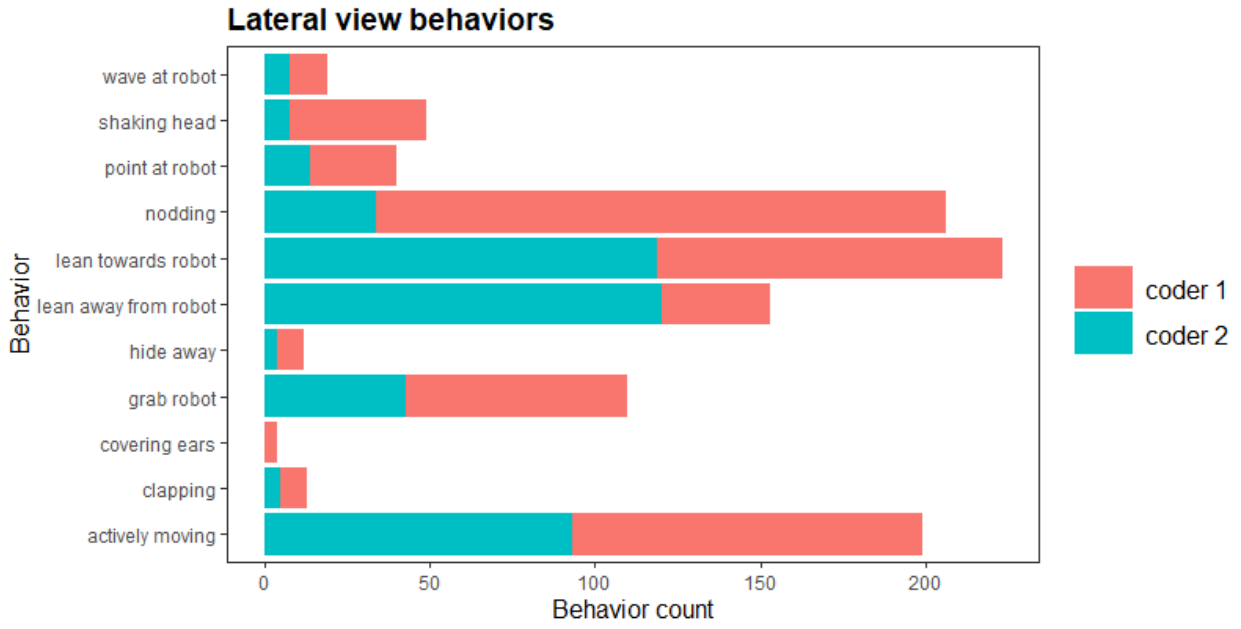
Coded behaviors	N (coder 1)	N (coder 2)	N (total)
Lateral view			
Leaning towards the robot	78	119	197
Leaning away from the robot	29	120	149
Hide away	6	4	10
Nodding	152	34	186
Shaking head	36	8	44
Covering ears	3	0	3
Clapping	6	5	11
Pointing at the robot	21	14	35
Grab robot	49	43	92
Wave at robot	8	8	16
Actively moving	100	136	236
Frontal view			
Amused	34	25	59
Bored	0	1	1
Mocking	1	1	2
Surprised	0	0	0
Distressed	0	1	1
Distracted	7	51	58
Other	2	0	2

Appendix C – Graphs demonstrating behavior counts of the raw data

Behavior counts for the frontal view videos of the raw data



Behavior counts for the lateral view videos of the raw data



Behavior durations for the state behaviors of the raw data