# Towards a Real-time Measure of the Perception of Anthropomorphism in Human-robot Interaction

Tsfasman, Maria; Saravanan, Avinash; Viner, Dekel; Goslinga, Daan; De Wolf, Sarah; Raman, Chirag; Jonker, Catholijn M.; Oertel, Catharine

**DOI**
10.1145/3475959.3485394

**Publication date**
2021

**Document Version**
Final published version

**Published in**
MuCAI 2021 - Proceedings of the 2nd ACM Multimedia Workshop on Multimodal Conversational AI, co-located with ACM MM 2021

**Citation (APA)**
Tsfasman, M., Saravanan, A., Viner, D., Goslinga, D., De Wolf, S., Raman, C., Jonker, C. M., & Oertel, C. (2021). Towards a Real-time Measure of the Perception of Anthropomorphism in Human-robot Interaction. In *MuCAI 2021 - Proceedings of the 2nd ACM Multimedia Workshop on Multimodal Conversational AI, co-located with ACM MM 2021* (pp. 13-18). (MuCAI 2021 - Proceedings of the 2nd ACM Multimedia Workshop on Multimodal Conversational AI, co-located with ACM MM 2021). ACM. https://doi.org/10.1145/3475959.3485394

# Towards a Real-time Measure of the Perception of Anthropomorphism in Human-robot Interaction

Maria Tsfasman*
M.Tsfasman@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Avinash Saravanan*
A.Saravanan@student.tudelft.nl
Delft University of Technology
Delft, The Netherlands

Dekel Viner*
D.Viner@student.tudelft.nl
Delft University of Technology
Delft, The Netherlands

Daan Goslinga
D.B.Goslinga@student.tudelft.nl
Delft University of Technology
Delft, The Netherlands

Sarah de Wolf
S.C.M.deWolf@student.tudelft.nl
Delft University of Technology
Delft, The Netherlands

Chirag Raman
C.A.Raman@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Catholijn M. Jonker
C.M.Jonker@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Catharine Oertel
C.R.M.M.Oertel@tudelft.nl
Delft University of Technology
Delft, The Netherlands

## ABSTRACT

How human-like do conversational robots need to look to enable long-term human-robot conversation? One essential aspect of long-term interaction is a human's ability to adapt to the varying degrees of a conversational partner's engagement and emotions. Prosodically, this can be achieved through (dis)entrainment. While speech-synthesis has been a limiting factor for many years, restrictions in this regard are increasingly mitigated. These advancements now emphasise the importance of studying the effect of robot embodiment on human entrainment. In this study, we conducted a between-subjects online human-robot interaction experiment in an educational use-case scenario where a tutor was either embodied through a human or a robot face. 43 English-speaking participants took part in the study for whom we analysed the degree of acoustic-prosodic entrainment to the human or robot face, respectively. We found that the degree of subjective and objective perception of anthropomorphism positively correlates with acoustic-prosodic entrainment.

## CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in collaborative and social computing*; **Empirical studies in interaction design**; • **Computer systems organization** → Robotics.

## KEYWORDS

multi-modal, human-robot interaction, prosody, acoustic-prosodic entrainment

*All three authors contributed equally to this research.



*MuCAI '21, October 24, 2021, Virtual Event, China.*

ACM ISBN 978-1-4503-8679-1/21/10.
https://doi.org/10.1145/3475959.3485394

**ACM Reference Format:**
Maria Tsfasman, Avinash Saravanan, Dekel Viner, Daan Goslinga, Sarah de Wolf, Chirag Raman, Catholijn M. Jonker, and Catharine Oertel. 2021. Towards a Real-time Measure of the Perception of Anthropomorphism in Human-robot Interaction. In *Proceedings of the 2nd ACM Multimedia Workshop on Multimodal Conversational AI (MuCAI '21), October 24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3475959.3485394

## 1 INTRODUCTION

To what degree the anthropomorphism of robots affects human-robot conversation is becoming an increasingly pressing question. Underlying technology such as speech synthesis, speech recognition, and natural language understanding has improved recently to such a degree that long-term human-robot applications are becoming much more feasible. A critical aspect of long-term human-robot interaction is the creation of rapport, a social clicking or bonding between interaction partners [34]. In human-human interaction, rapport has been shown to be related to acoustic-prosodic entrainment [19], an ability of humans to adapt their prosody to each other within the conversation [16]. Acoustic-prosodic entrainment can also be a reliable indicator of conversational involvement and interaction quality [7, 24]. Higher acoustic-prosodic entrainment can indicate more interest in the topic discussed, higher levels of attention [14, 24] and can also be related to the activation of mirror neurons [11, 17].

In human-robot interaction, acoustic-prosodic entrainment has been implemented as a tool to increase a robot's social presence and a user's rapport towards it [20]. The ability of the robot to entrain to the user has also been shown to increase children's engagement within the interaction [28].

The famous 'uncanny valley' phenomenon [21] implies that if the robot is too human-like, it can decrease the likeability of the robot. Yet, human-like appearance has shown to increase perceived trustworthiness [8, 22] and social presence [13, 29] of conversational agents. Questions related to the link between anthropomorphism of and attitudes towards virtual agents are widely studied [9], yet

questions related to how anthropomorphism influences the user's behaviour are less explored.

Certain qualities within Instructors in educational contexts can potentially impact the educational outcomes as well. In an educational context, the anthropomorphism of the tutor has been shown to improve the understanding and memorisation of information, and human tutors still show better results than a robot or a tablet [36]. Nguyen et. al [33] investigated how acoustic-prosodic entrainment correlates with the quality of information acquisition in a tutoring experiment. They compared knowledge gain and acoustic-prosodic entrainment of students when learning from a tutor with a human in contrast to a synthesized voice. Based on amplitude and pitch features, [33] show that higher acoustic-prosodic entrainment positively correlates with students' knowledge gain. More importantly, they found that students' acoustic-prosodic entrainment was higher in the human voice scenario. In [33]'s study, the virtual tutor was a voice assistant, and the conditions were different in the human-likeness of the virtual tutor's speech. But does visual human-likeness affect the entrainment in the same way?

In this paper, we investigate the effect of a virtual tutor's human versus machine-like appearance on a user's prosodic entrainment. In fact, there have been human-robot interaction studies measuring user entrainment in human-robot interaction. Breazeal [5] showed that humans are able to adapt their turn-taking behaviour to a robot. Strupka et al. [32] investigated how humans adapt their speech to the robots of different genders found that participants exhibited speech divergence (the opposite of entrainment) in both conditions.

If there is a connection between anthropomorphism of robot appearance and acoustic-prosodic entrainment, that could indicate the importance of making virtual agents and social robots as human-like as possible for long-term interaction scenarios such as in a hybrid-intelligence scenario [1].

Real-time assessment of entrainment could act as a non-verbal indication of engagement and rapport towards the robot in the future. This way, no additional questionnaires are needed to access the user perception of the robot. It could be done in an online manner, by processing the user's speech.

Another field that could benefit from understanding the connection between anthropomorphism and acoustic-prosodic entrainment is multi-modal addressee detection. There are many studies focusing on advancing automatic detection of human- versus system-addressed speech [30, 31, 35]. All of the cited algorithms show to benefit from using prosodic features as predictors of addressee tags. However, they have not used acoustic-prosodic entrainment as their feature, and in case our hypothesis is confirmed, it might aid the automatic detection of whether the user is addressing a machine or another human.

## 2 RESEARCH QUESTION

The research question we aim to answer in the present paper is following: Does the type of facial embodiment of a conversational agent influence the level of acoustic-prosodic entrainment of a person interacting with it?

To answer this question in the present work we investigate human acoustic-prosodic entrainment in two conditions:

(1) **Human condition** (Human Face, Human Voice) - A lesson is taught with pre-recorded videos of a human tutor.
(2) **Robot condition** (Robotic Face, Human Voice) - The same lesson is taught by a virtual agent mimicking the exact face movements and using the audio of the human tutor from the first condition.

### 2.1 Hypothesis

In relation to audio cues, it has been tested and confirmed in that the synthesized voice modulates less acoustic-prosodic entrainment than the human voice. However, it hasn't been tested in relation to visual cues independently of the phonetic or non-verbal cues. In this experiment, our main hypothesis is that the human appearance of a virtual tutor will incite more acoustic-prosodic entrainment than the robotic tutor with the same voice and facial expressions. Our second expectation is that the perceived anthropomorphism of the tutor will positively correlate with user acoustic-prosodic entrainment.

We hypothesise that :

(1) H1: participants show greater entrainment towards the human than the robot face
(2) H2: the greater the participant's perception of anthropomorphism the greater the degree of entrainment

## 3 METHOD

### 3.1 Stimulus Preparation

*3.1.1 Task.* Because of the COVID-19 pandemic, all participant interactions were carried out via Zoom. The participants followed three lessons on random topics (about meatball production, beer crafting and venomous species). Each lesson lasted around 10 minutes and consisted of multiple pre-recorded videos taught by a virtual tutor (robot or human depending on the condition). Each pre-recorded video lasted for about 2-10 seconds and ended in an open question inviting participants for interaction. The participants were invited to verbally interact with the tutor throughout the whole lesson but especially after the tutor asks a question. After the video, a pre-recorded idle state was played for as long as a participant was replying. The idle states consisted of common back-channels (nodding, smiling, etc.) also pre-recorded with the same virtual tutor. The length of idle states was controlled by the experimenter within each experiment to avoid interrupting participants. After finishing all the lessons participants had to complete a questionnaire on their perception of the interaction and the tutor.

*3.1.2 Conditions.* There were two conditions: (1) **human** and (2) **robot**. In the human condition, the lessons were recorded from a male English-speaking actor. The only difference between the human and the robot condition was the appearance of the tutor. The audio from the human tutor recording was used in both conditions, the content and the post-questionnaire was also the same across conditions. Since the conditions had the same content, the experiment had a between-subject design.

For the robot condition, we used the Furhat SDK [2]. Furhat is "a social robot with human-like expressions and advanced conversational artificial intelligence (AI) capabilities" [27]. For the recording of the robot condition stimuli, we used the Furhat simulation.

To make facial expressions as similar as possible between the two conditions, we used code that resynthesised the human tutor's face movements in Furhat. The importance of this step can be illustrated by Breazeal's [6] findings on the significance of body posture, head tilt and facial expression for human-robot entrainment. The code used OpenCV python library [4] to track mouth movements, smile and head position and rotation in the video of the human tutor. It then proceeded to convert the resulting facial movements and their durations to the Furhat implementation.

Figure 1 shows a screenshot of videos shown to the participants in two different conditions, side-by-side, in the same moment of time:

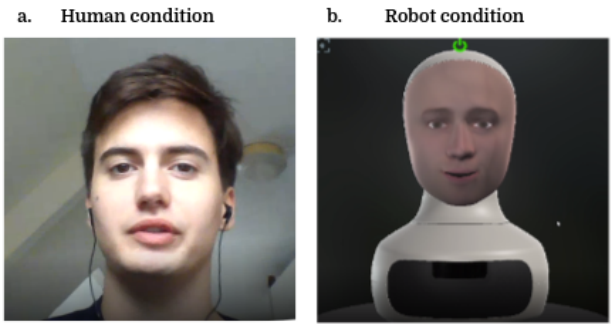


**Figure 1: Screenshot of stimuli videos in human (a) and in robot (b) conditions, at the same point of time**

*3.1.3 Questionnaires.* After the participants have finished the virtual lessons, they answered questions relating to their experience of interacting with the tutor in a Qualtrics survey. These questions included Godspeed questionnaire scales on animacy and anthropomorphism [3] and two questions on perception of whether the videos had smooth transitions between interaction and how interesting were the lessons.

## 3.2 Experimental Set-Up.

*3.2.1 Experimental Study.* We strove to reduce the differences between conditions to the level of human-likeness of tutor appearance to use conditions as an objective measure of anthropomorphism. For the subjective measure of anthropomorphism, we used the ratings collected from the Godspeed anthropomorphism scale. The subjective measure was meant to both validate the distinction between conditions and test our hypothesis.

*3.2.2 Participants.* The experiment was conducted on 51 English-speaking participants, who have never met the tutor before and never came across Furhat robot. It was important to have participants that have never interacted with the human tutor since prior interactions might bias the way in which people display their entrainment [37]. Of these 51 participants, 43 have been included in the results. The 8 other participants have been excluded because of internet problems and recording errors, which made the data unusable. The resulting set of participants included 22 females and 21 males (mean age 28 +- 10).

*3.2.3 Procedure.* Each experiment lasted for about 40 minutes altogether. Each participant was assigned a random condition. After signing a consent form for audio recording and data storage, participants had to join a Zoom call with the experimenter, and after being fully instructed, watched the lessons through experimenter's screen sharing. The experimenter's video was off throughout the whole call and the audio interaction was reduced to a minimum to avoid participants entraining with the experimenter. The instructions for the participants were to wear a personal headset, listen carefully and verbally interact with the tutor as much as possible.

## 3.3 Analysis

Our experiment contains one independent variable (facial embodiment) and one dependent variable (acoustic-prosodic entrainment). Facial embodiment is a binary variable. Namely, the facial appearance is either a video of a human actor or a robotic face (Furhat robot). This distinction between conditions we use as a measure of objective anthropomorphism. The acoustic-prosodic entrainment is a variable that spans over multiple features and metrics extracted from the audio recording of participants' speech.

*3.3.1 Prosodic feature extraction.* To analyse the audio Parselmouth [12] and Pydub [26] python libraries were used. Parselmouth was used to extract pitch and RMS-intensity.

The prosodic features that we extracted are similar to the ones used in Levitan et al. [15]: mean Intensity, max Intensity, mean Pitch, max Pitch. Because of the online setting of the experiment, we could not control for the quality of the microphone, most participants used wired earphones. This was done to avoid leaking of tutor voice into the audio recording of the participants and to reduce the noise captured in the recording.

*3.3.2 Preprocessing of features.* The prepossessing conducted on the data before computing the entrainment metrics included standardization and KNN regression.

In a normal conversation, there are many moments in which one person is silent while the other is speaking and vice versa. There are also moments in which both speakers are silent. In both of these scenarios, it would make no sense to compute values for the metrics for entrainment. We use KNN regression on the data in order to fill in those gaps [10]. For every feature at each time point take we take the average of the k nearest values (k = 7 in our case). Where the distance per value is computed against the centre time point of the utterance. In other words $\frac{F_{start\_time}+F_{end\_time}}{2}$.

*3.3.3 Acoustic-prosodic entrainment metrics.* We used acoustic-prosodic entrainment metrics introduced by [16] and commonly used for measuring entrainment: proximity, convergence and synchrony.

**Proximity** is computed by taking the negative absolute difference per feature at every time point. In order to make our results comparable between participants of different voice characteristics (such as male and female voices) we standardized values to their z-score.

$$-|f^A(t) - f^B(t)| \quad (1)$$

Important to note that the metrics are computed after the KNN preprocessing so the time points are not referring to the raw data. The closer the metric value is to 0 the higher the assumed entrainment.

**Convergence** measures how proximity changes over time, where $D(t)$ stands for $-|f^A(t) - f^B(t)|$

$$\frac{\int_{t_0}^{t_n}(D(t)-\bar{D})*(t-\bar{t})dt}{\sqrt{\int_{t_0}^{t_n}(D-\bar{D})^2dt\int_{t_0}^{t_n}(t-\bar{t})^2dt}} \quad (2)$$

Convergence applies Pearson correlation and a positive convergence for a feature suggests that over the course of the conversation the proximity between the tutor and the student increases. Meaning the feature values become more similar. Likewise, negative convergence for a given feature means it becomes more dissimilar.

**Synchrony** is here taken simply as Pearson correlation for a given feature between the tutor and student.

$$\frac{\int_{t_0}^{t_n}(f^A(t+\delta)-\bar{f^A})*(f^B(t)-\bar{f^B})dt}{\sqrt{\int_{t_0}^{t_n}(f^A(t+\delta)-\bar{f^A})^2dt*\int_{t_0}^{t_n}(f^B(t)-\bar{f^B})^2dt}} \quad (3)$$

*3.3.4 Statistical Methods.* To investigate the differences between conditions, we utilized Kruskal-Wallis test. The reason for this is due to data not meeting the assumptions of Anova which are normality and homogeneity of variance. These assumptions were tested through Shapiro and Levene tests. Pearson correlation was computed to determine the significance and the direction of the correlations between the entrainment metrics and subjective measures of anthropomorphism, animacy and interest. Finally, power analyses were carried out to determine if the sample size and statistical power were appropriate and we found that the power was adequate (power > 0.8) for all mentioned significant results.

# 4 RESULTS

## 4.1 Experiment perception

The experiment perception questions at the end of the experiment included animacy and anthropomorphism scores from Godspeed questionnaire in order to confirm the opposition of human vs machine in the conditions. They also included a question on how interesting participants found the lessons and how smooth the transitions between interactions were. The interest score was aimed to control for the fact that participants could be more interested and therefore more engaged in the lesson with Furhat because of its novelty. It also served as a measure of subjective engagement, since entrainment has been linked to engagement before [28]. The question on smoothness of transitions was aimed to control for the differences in conditions connected to the way they were merged together in the experimental stimuli.

Figure 2 illustrates the differences between perception of each condition (human and Furhat). The y-axis shows the score normalised over the maximum score for each type of question, therefore for each perception parameter the maximum score is 1 and the minimum is 0.

The animacy and anthropomorphism scores were significantly higher in the human condition ($p<0.01$). This confirms our assumption of it being perceived more human-like. The animacy scores were also significantly higher in the human condition, as expected ($p<0.01$). The interest and the smoothness scores were insignificantly different between conditions, which confirms that the conditions were similar technically and content-wise.

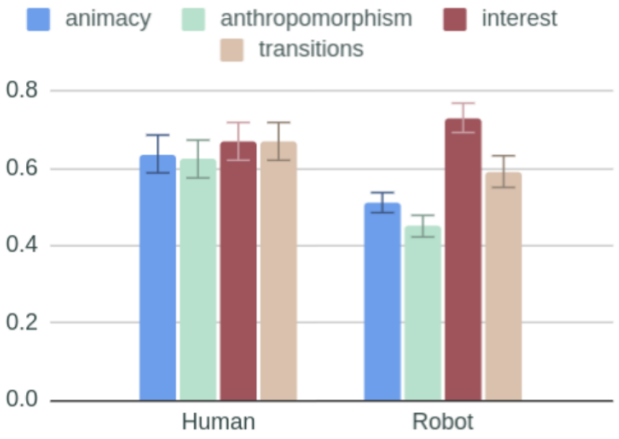


**Figure 2: Average experiment perception scores for different conditions (error bars show standard error).**

## 4.2 Acoustic-prosodic entrainment

Our results indicate that, despite the online setting and artificial nature of the interactions, all participants entrained on the tutor voice: Every participant in both conditions had at least one feature with significantly positive convergence or at least one with significantly positive synchrony.

Although there was entrainment in both conditions, participants' proximity, convergence and synchrony in their intensity (mean and max) and max pitch were insignificantly different between conditions ($p>0.01$). However, for the mean pitch, which is a major predictor of acoustic-prosodic entrainment [10], the convergence was significantly lower in robot condition in comparison to the human condition ($p<0.01$). This said, proximity and synchrony by mean pitch by themselves were insignificantly different between conditions.

Although the variability between participants is quite high (see figure 3), the convergence by mean pitch was significantly positive for 65% of participants in human condition. In robot condition only 39 % of participants displayed significantly positive convergence by mean pitch in robot condition. This means that the acoustic-prosodic entrainment was stronger in the human condition, confirming our hypothesis.

Not only objective human-likeness (i.e. distinction between condition), but also subjective perception of anthropomorphism appeared to positively correlate with convergence by mean pitch (Pearson correlation $p<0.01$). To illustrate the trend, we plotted linear regression over all participants figure 4.

Another significantly positive correlation can be noticed between convergence by mean pitch and the perception of animacy of the tutor (figure 5). The answers on a question on subjective engagement ('Rate how interesting did you find the lesson from 1 to 10') show similar trend: there was a significantly positive correlation of convergence by mean pitch to the interest score (Pearson correlation $p<0.01$).

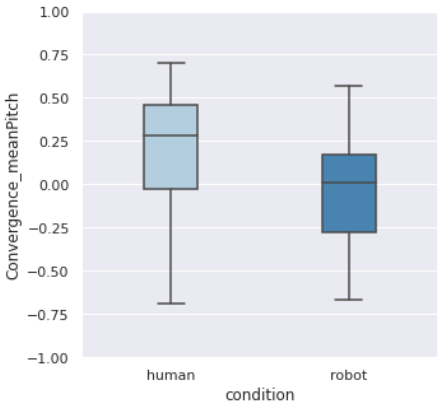


**Figure 3: Differences in mean pitch convergence between conditions.**


**Figure 4: Significant positive correlation between convergence by mean pitch and perceived anthropomorphism.**


**Figure 5: Significant positive correlation between convergence by mean pitch and perceived animacy.**

## 5 DISCUSSION

We found that mean pitch convergence was significantly higher in the human than in the robot condition. This confirms H1. The fact that humans entrain more to other humans than to a robot is in line with findings by Strupka et al. [32].

We found that the acoustic-prosodic entrainment in mean pitch was positively correlated with participant's perception of human-likeness. In fact this was true for both the subjective (perceived anthropomorphism) and objective (the distinction between conditions) anthropomorphism measures. This confirms H2.

The fact that we did not find prosodic entrainment in other prosodic features is in line with previous research on human-human entrainment [10]. This might be related to the variability in participants' mother-tongue - although they all were English-speaking, they had different degree's of English proficiency and English accents; they also had different cultural backgrounds and varied in gender and age. Since all those can affect acoustic-prosodic entrainment [16, 18]. A bigger sample study will be carried out in the future.

Our results expand the previous study by [33] to the visual domain: in [33]'s experiment the less human-like voice triggered less acoustic-prosodic entrainment than a human voice. In our experiment, the robotic appearance of a tutor triggered less entrainment than a human tutor. This also is in line with the distinction between human-addressed and machine-addressed speech found in [31, 35]. Adding entrainment as one of the features could prove to be beneficial for those algorithms.

The fact that we discovered the effect of anthropomorphism on entrainment brings various implications, since there are many perceptual factors that are entailed by anthropomorphism in human-robot interaction and entrainment in human-human interaction. More human-like conversational agents seem more trust-worthy [8, 22] and socially present [13, 29]. Therefore, it may suggest that higher prosodic entrainment could correlate with conversational agent's perceived trustworthiness and social presence in human-robot interaction, as well as speakers' rapport [19] and engagement [16, 24].

All in all acoustic-prosodic entrainment appears to be a promising behavioural measure of access to a person's perception of animacy and anthropomorphism in his/her conversational partner. A real-time measure based on prosodic entrainment could widely benefit fields such as social robotics and hybrid intelligence [1]. It might further be a relevant measure to further investigate in relation to human-human/machine addressee detection [30, 31, 35].

On a cognitive level, there might be a link between anthropomorphism and entrainment via mirror neurons, which is activated when a human interacts or observes another human [11, 25]. The mirror neurons are thought to activate embodied experiences and therefore aid imitation learning [23]. Since entrainment is imitation in itself, mirror neurons can also be viewed as crucial mechanism in social entrainment [17]. This might explain why anthropomorphism of the tutor triggered more acoustic-prosodic entrainment in our experiment, and why interest and animacy scores also correlated with higher entrainment.

## 6 CONCLUSION AND FUTURE WORK

In this work, we investigated the effect of a human versus robot face on prosodic entrainment in an educational use-case scenario. We could show that humans converged to a higher degree in mean pitch to another human face than a robot face. Maybe more importantly, though, we could show that the greater the perception of animacy and anthropomorphism, the greater the degree of prosodic entrainment. In future research, we plan to add the variable of age and gender to our experimental setup. Using purposeful manipulations of prosodic convergence, we aim to explore their effect on participants' recollection of the conversation.

## ACKNOWLEDGMENTS

We would like to thank Navin Laxminarayanan Raj Prabhu for providing us with his code for replicating facial features in Furhat.

## REFERENCES

[1] Zeynep Akata, Dan Balliet, Maarten de Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, Hayley Hung, Catholijn Jonker, Christof Monz, Mark Neerincx, Frans Oliehoek, Henry Prakken, Stefan Schlobach, Linda van der Gaag, Frank van Harmelen, Herke van Hoof, Birna van Riemsdijk, Aimee van Wynsberghe, Rineke Verbrugge, Bart Verheij, Piek Vossen, and Max Welling. 2020. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer* 53, 8 (2020), 18–28. https://doi.org/10.1109/MC.2020.2996587

[2] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction. In *Cognitive Behavioural Systems*, Anna Esposito, Antonietta M. Esposito, Alessandro Vinciarelli, Rüdiger Hoffmann, and Vincent C. Müller (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 114–130.

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

[4] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[5] Cynthia Breazeal. 2001. Regulation and Entrainment in Human-Robot Interaction. In *Experimental Robotics VII*, Daniela Rus and Sanjiv Singh (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 61–70.

[6] Cynthia Breazeal. 2002. Regulation and Entrainment in Human—Robot Interaction. *The International Journal of Robotics Research* 21, 10-11 (2002), 883–902. https://doi.org/10.1177/027836490202101009 6 arXiv:https://doi.org/10.1177/027836490202101009 6

[7] Céline De Looze, Catharine Oertel, Stéphane Rauzy, and Nick Campbell. 2011. Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In *ICPhS 2011*.

[8] E.J. de Visser, S.S. Monfort, R. McKendrick, M.A.B. Smith, P.E. McKnight, F. Krueger, and R. Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22, 3 (2016), 331–349. https://doi.org/10.1037/xap0000092 cited By 111.

[9] Julia Fink. 2012. Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction. In *Social Robotics*, Shuzhi Sam Ge, Oussama Khatib, John-John Cabibihan, Reid Simmons, and Mary-Anne Williams (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 199–208.

[10] Ramiro H. Gálvez, Lara Gauder, Jordi Luque, and Agustín Gravano. 2020. A unifying framework for modeling acoustic/prosodic entrainment: definition and evaluation on two large corpora. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1st virtual meeting, 215–224. https://aclanthology.org/2020.sigdial-1.27

[11] Tamara Gog, Fred Paas, Nadine Marcus, Paul Ayres, and John Sweller. 2008. The Mirror Neuron System and Observational Learning: Implications for the Effectiveness of Dynamic Visualizations. *Educational Psychology Review* 21 (03 2008), 21–30. https://doi.org/10.1007/s10648-008-9094-3

[12] Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* 71 (2018), 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

[13] Ran Hee Kim, Yeop Moon, Jung Ju Choi, and Sonya S. Kwak. 2014. The Effect of Robot Appearance Types on Motivating Donation. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 210–211.

[14] Rivka Levitan, Stefan Benus, Ramiro Gálvez, Agustín Gravano, Florencia Savoretti, Marián Trnka, Andreas Weise, and Julia Hirschberg. 2016. Implementing Acoustic-Prosodic Entrainment in a Conversational Avatar. 1166–1170. https://doi.org/10.21437/Interspeech.2016-985

[15] Rivka Levitan, Agustín Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. 2012. Acoustic-prosodic entrainment and social behavior. 11–19.

[16] Rivka Levitan, Agustín Gravano, Laura Willson, SÌ†tefan BenÌ†usÌ†, Julia Hirschberg, and Ani Nenkova. 2012. Acoustic-Prosodic Entrainment and Social Behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, 11–19. https://www.aclweb.org/anthology/N12-1002

[17] J. Levy, A. Goldstein, O. Zagoory-Sharon, O. Weisman, I. Schneiderman, M. Eidelman-Rothman, and R. Feldman. 2016. Oxytocin selectively modulates brain response to stimuli probing social synchrony. *NeuroImage* 124 (2016), 923–930. https://doi.org/10.1016/j.neuroimage.2015.09.066 cited By 29.

[18] Eva M. Lewandowski and Lynne C. Nygaard. 2018. Vocal alignment to native and non-native speakers of English. *The Journal of the Acoustical Society of America* 144, 2 (2018), 620–633. https://doi.org/10.1121/1.5038567 arXiv:https://doi.org/10.1121/1.5038567

[19] Nichola Lubold and Heather Pon-Barry. 2014. Acoustic-Prosodic Entrainment and Rapport in Collaborative Learning Dialogues. In *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (Istanbul, Turkey) *(MLA '14)*. Association for Computing Machinery, New York, NY, USA, 5–12. https://doi.org/10.1145/2666633.2666635

[20] Nichola Lubold, Erin Walker, and Heather Pon-Barry. 2016. Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 255–262. https://doi.org/10.1109/HRI.2016.7451760

[21] M. Mori. 1970. Uncanny Valley. *Energy* 7, (4) (1970).

[22] M. Natarajan and M. Gombolay. 2020. Effects of anthropomorphism and accountability on trust in human robot interaction. *ACM/IEEE International Conference on Human-Robot Interaction* (2020), 33–42. https://doi.org/10.1145/3319502.3374839 behavior and anthropomorphism of the agent are the most significant factors in predicting the trust and compliance with the robot.

[23] R. Ramsey, D.M. Kaplan, and E.S. Cross. 2021. Watch and Learn: The Cognitive Neuroscience of Learning from Others' Actions. *Trends in Neurosciences* 44, 6 (2021), 478–491. https://doi.org/10.1016/j.tins.2021.01.007 cited By 0.

[24] Uwe D Reichel, Katalin Mády, and Jennifer Cole. 2018. Prosodic entrainment in dialog acts. *arXiv preprint arXiv:1810.12646* (2018).

[25] G. Rizzolatti and L. Craighero. 2004. The mirror-neuron system. *Annual Review of Neuroscience* 27 (2004), 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230 cited By 5010.

[26] James Robert, Marc Webbie, et al. 2018. Pydub. http://pydub.com/

[27] Furhat Robotics. 2020. *Furhat robotics about page*. Accessed: 2020-12-13.

[28] Najmeh Sadouohi, André Pereira, Rishub Jain, Lolanda Leite, and Jill Fain Lehman. 2017. Creating Prosodic Synchrony for a Robot Co-Player in a Speech-Controlled Game for Children. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. 91–99.

[29] Ryan M. Schuetzler, G. Mark Grimes, and Justin Scott Giboney. 2020. The impact of chatbot conversational skill on engagement and perceived humanness. *Journal of Management Information Systems* 37, 3 (2020), 875–900. https://doi.org/10.1080/07421222.2020.1790204 arXiv:https://doi.org/10.1080/07421222.2020.1790204

[30] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tur, and Larry Heck. 2012. Learning When to Listen: Detecting System-Addressed Speech in Human-Human-Computer Dialog. 1 (01 2012).

[31] Elizabeth Shriberg, Andreas Stolcke, and Suman Ravuri. 2013. Addressee Detection for Dialog Systems Using Temporal and Spectral Dimensions of Speaking Style. In *Proc. Interspeech* (proc. interspeech ed.). ISCA - International Speech Communication Association, 2559–2563.

[32] Eszter Strupka, Oliver Niebuhr, and Kerstin Fischer. 2016. Influence of Robot Gender and Speaker Gender on Prosodic Entrainment in HRI.

[33] Jesse Thomason, Huy Nguyen, and Diane Litman. 2013. Prosodic Entrainment and Tutoring Dialogue Success. 7926 (06 2013). https://doi.org/10.1007/978-3-642-39112-5-104

[34] Linda Tickle-Degnen and Robert Rosenthal. 1990. The Nature of Rapport and Its Nonverbal Correlates. *Psychological Inquiry* 1, 4 (Oct. 1990), 285–293. https://doi.org/10.1207/s15327965pli0104_1

[35] T. J. Tsai, Andreas Stolcke, and Malcolm Slaney. 2015. A Study of Multimodal Addressee Detection in Human-Human-Computer Interaction. *IEEE Transactions on Multimedia* 17, 9 (2015), 1550–1561. https://doi.org/10.1109/TMM.2015.2454332

[36] J. K. Westlund, L. Dickens, Sooyeon Jeong, P. Harris, D. DeSteno, and C. Breazeal. 2015. A Comparison of Children Learning New Words from Robots , Tablets , People.

[37] Jiahong Yuan, Mark Liberman, and Christopher Cieri. 2007. Towards an integrated understanding of speech overlaps in conversation. (01 2007).