Does the Appearance of an Agent Affect How We Perceive his/her Voice? Audio-visual Predictive Processes in Human-robot Interaction

Busra Sarigul

Interdisciplinary Social Psychiatry Program, Ankara University, Turkey
Department of Psychology, Nuh Naci Yazgan University, Turkey
busra.srgl@gmail.com

Imge Saltik

Interdisciplinary Neuroscience Program
Bilkent University, Turkey
imge.saltik@bilkent.edu.tr

Batuhan Hokelek

Department of Psychology
Bilkent University, Turkey
batuhan.hokelek@ug.bilkent.edu.tr

Burcu A. Urgan

Department of Psychology & Interdisciplinary Neuroscience Program
National Magnetic Resonance Research Center (UMRAM)
Aysel Sabuncu Brain Research Center
Bilkent University, Turkey
burcu.urgan@bilkent.edu.tr

ABSTRACT

Robots increasingly become part of our lives. How we perceive and predict their behavior has been an important issue in HRI. To address this issue, we adapted a well-established prediction paradigm from cognitive science for HRI. Participants listened a greeting phrase that sounds either human-like or robotic. They indicated whether the voice belongs to a human or a robot as fast as possible with a key press. Each voice was preceded with a human or robot image (a human-like robot or a mechanical robot) to cue the participant about the upcoming voice. The image was either congruent or incongruent with the sound stimulus. Our findings show that people reacted faster to robotic sounds in congruent trials than incongruent trials, suggesting the role of predictive processes in robot perception. In sum, our study provides insights about how robots should be designed, and suggests that designing robots that do not violate our expectations may result in a more efficient interaction between humans and robots.

CCS CONCEPTS

• Human-centered computing • Human-computer interaction (HCI) • HCI design and evaluation methods • User studies

KEYWORDS

Humanoid robots, robot design, audio-visual mismatch, prediction, robotic voice, human perception, cognitive sciences

ACM Reference format:

Busra Sarigul, Imge Saltik, Batuhan Hokelek, Burcu A. Urgan. 2020. Does the Appearance of an Agent Affect how we Perceive his/her Voice? Audio-visual Predictive Processes in Human-robot Interaction. In *Proceedings of ACM HRI conference (HRI'20)*, March 23-26, 2020, Cambridge, UK. ACM, NY, NY, USA, 3 pages. <https://doi.org/10.1145/3371382.3378302>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

HRI '20 Companion, March 23–26, 2020, Cambridge, United Kingdom

© 2020 Copyright is held by the owner/author(s).

ACM ISBN 978-1-4503-7057-8/20/03.

<https://doi.org/10.1145/3371382.3378302>

1. INTRODUCTION

One fundamental cognitive mechanism humans possess is to be able to make predictions about what will happen in their environment [1]. This skill is very important to take the appropriate action based on what we perceive. Literature of predictive processing heavily focuses on visual perception. However, our daily experience is multi-modal in nature [2]. For instance, imagine that a friend of yours sees you across the street and waves to you. Upon seeing his/her gesture, you probably predict that he/she will say “Hello!” to you. Thus, based on what you see, you can have an expectation about what you will hear. As humanoid robots increasingly become participants in our environments such as hospitals, airports or schools, it is likely that we learn from our multimodal experiences with them [3-5], and predict how they will behave in our mutual interaction. For instance, if a robot companion waves to us, we may expect that it immediately says “Hello!”. Moreover, how it appears can give us clues about how it will behave, and accordingly affect how we perceive them [6-7]. Previous research shows that when there is a mismatch between what we expect and what we perceive while interacting with artificial agents, we may find them unacceptable, eerie, or uncanny [8-11]. The aim of the present study is to investigate whether we make predictions about how robots behave based on how they appear, and if so what kind of predictions we make, and whether those predictions are similar to the ones we make for humans.

Figure 1. Visual stimuli in the experiment

2. METHODS

Seventeen healthy adults from Bilkent University (9 females, Mean = 23.47, SD = 2.70) participated in the experiment. All participants had normal vision and no history of neurological disorders.

Figure 2. Results of Condition 1 and Condition 2, and Comparison of Condition 1 and Condition 2.

Before the experiment, all subjects signed a consent form approved by the Ethics Committee of Bilkent University.

Visual stimuli: The visual stimuli consist of static images of three agents. We call them Human, Android, and Robot (Figure 1, and also [9, 10, 12, 13]). Android and Robot are the same machine in different appearances. Android has a more human-like appearance.

Auditory stimuli: The auditory stimuli consist of two sound files which last 2 seconds: the voice of a human saying ‘Good morning’ (human voice), and a modified version of it in which the voice sounds robotic (robotic voice). To make the natural human voice to robotic, the frequency of the original sound file was manipulated. To determine whether the manipulation works and people find the sound really robotic, we did a pilot study. We created 14 different sound files in which the frequency was manipulated between -5 Hz to -20 Hz with a step size of 2 Hz, using the software Audacity 2.3.0. According to ratings, we identified the most robotic voice based on the average of the subjects, so this voice was used as the stimuli in the experiment with the original human voice. We also added white noise to both sound files to make the task harder. The subjects participated in an experiment that included two conditions. Both conditions consisted of 5 blocks, each of which had 80 trials. Each trial started with a fixation cross (1 sec). It was followed by a visual cue (1 sec) which was an image of Human or Robot in Condition 1, and Human and Android in Condition 2. The visual cue was followed by a 2 sec auditory stimulus: human voice or robotic voice. The subjects’ task was to indicate whether the auditory stimulus was a human voice or a robotic voice with a key press. The visual cue informed the subjects about the upcoming auditory stimulus. 80% of the trials, the visual cue was congruent with the auditory stimulus, and in 20% of the trials it was incongruent. The order of the conditions was counter-balanced across subjects.

3. RESULTS

Condition 1: There was a main effect of congruency on reaction times ($F(1,16) = 13.47, p < 0.05$). Subjects responded to auditory targets significantly faster in congruent trials ($M = 1.21$ sec, $SD = 0.04$) than incongruent trials ($M = 1.30$ sec, $SD = 0.03$). We also conducted pair-wise t-tests between congruent and incongruent conditions for each visual cue (Human and Robot) separately. When the visual cue was Robot, subjects performed significantly faster in congruent trials than incongruent trials ($t(16) = -8.02, p < 0.05$). When the visual cue was Human, subjects performed similarly in congruent and incongruent conditions ($t(16) = -0.82, p = 0.94$) (See Figure 2, left).

Condition 2: There was a main effect of congruency on reaction

times. Subjects responded faster in congruent trials ($M = 1.22, SD = 0.04$) than incongruent trials ($M = 1.28, SD = 0.04$) ($F(1,16) = 16.16, p < 0.05$). Furthermore, we performed pair-wise t-tests between congruent and incongruent conditions for each visual cue (Human and Android) separately. When the visual cue was Android, subjects performed significantly faster in congruent trials than incongruent trials ($t(16) = -7.98, p < 0.05$). When the visual cue was Human, subjects performed similarly in congruent and incongruent conditions ($t(16) = 0.58, p = 0.60$) (See Figure 2).

Comparison of Condition 1 and Condition 2: We investigated whether the human-likeness of the visual cue (Human, Android, Robot) affects how people categorize auditory target, and whether it interacts with congruency. So, we compared Condition 1 and Condition 2. There was not a significant difference between the Human part in Condition 1 and the Human part in Condition 2, so, we included only one of the Human parts in the analysis. There was a main effect of congruency on reaction times ($F(1,16) = 31.22, p < 0.05$) but there was no main effect of visual cue ($F(1,16) = 0.01, p = 0.99$). However, interestingly, there was an interaction between congruency and visual cue ($F(1,16) = 28.74, p < 0.05$). A closer look at the pattern of results showed that the difference between the congruent and incongruent conditions was largest for Robot, followed by Android, followed by Human (See Figure 2, right).

4. DISCUSSION

We hypothesized that people would get faster in judging how an agent sounded like if it was preceded by a congruent visual cue than incongruent cue. Our results show that if the visual cue is a robot, people expect that it would sound robotic as demonstrated by shorter reaction times in congruent condition (robot cue and robot voice) than incongruent condition (robot cue and human voice). This was true whether the robot has a more human-like appearance (Android) or less human-like appearance (Robot). An unexpected finding in our study was that people respond to a human voice similarly regardless of the visual cue it precedes. One possible explanation for these results is that the human voice stimulus was not ambiguous enough. People are more likely to use cues (or prior knowledge) when the task at hand is hard (e.g. the stimulus is ambiguous) [14]. One way to resolve this issue is to increase the white noise in the voice stimuli and make the task harder. Nevertheless, the use of a well-established prediction paradigm from cognitive sciences in the present study has opened a new avenue of research in HRI. Appearance and voice are only two features among many, for which we seek a match in agent perception. Future work should investigate what features of artificial agents make us form expectations, how we do that, and under what conditions these expectations are violated.

5. REFERENCES

- [1] A. Clark (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36 (3), 181-204
- [2] O. Doehrmann, & M.J. Naumer (2008). Semantics and the multisensory brain: how meaning modulates processes of audio-visual integration. *Brain research*, 1242, 136-150
- [3] C. Nass (2015). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. MIT Press, Cambridge, MA.
- [4] S.E. Stern, J.W. Mullennix, I. Yaroslavsky (2006). Persuasion and social perception of human vs. synthetic voice across person as source and computer as source conditions. *International Journal of Human-Computer Interaction*, 64, 43-52.
- [5] M.L. Walters, D.D. Dyrda, K.L. Koay, K. Dautenhahn, R. te Boeckhorst (2008). Human approach distances to a mechanical-looking robot with different voice styles. *Proceedings of RO-MAN*, Munich, Germany.
- [6] K. Zibrek, E. Kokkinara, R. McDonnell (2018). The effect of realistic appearance of virtual characters in immersive environments – Does the character's personality play a role? *IEEE Transactions on Visualization and Computer Graphics*, 24 (4), 1681-1690.
- [7] C. Mousas, D. Anastasiou, O. Spantidi (2018). The effects of appearance and motion of virtual characters on emotional reactivity. *Computers in Human Behavior*, 86, 99-108.
- [8] Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2), 98-100.
- [9] B.A. Urgan, M. Kutas, A.P. Saygin (2018). Uncanny valley as a window into predictive processing in the social brain. *Neuropsychologia*, 114, 181-185
- [10] W.J. Mitchell, K.A. Szerszen, A.S. Lu, P.W. Schermerhorn, M. Scheutz, K.F. Macdorman (2011). A mismatch in the human realism of face and voice produces an uncanny valley. *Iperception*, 2 (1), 10-12.
- [11] A.P. Saygin, T. Chaminade, H. Ishiguro, J. Driver, C. Frith (2012). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive Affective Neuroscience*, 7, 413-422.
- [12] B.A. Urgan, M. Plank, H. Ishiguro, H. Poizner, A.P. Saygin (2013). EEG theta and mu oscillations during perception of human and robot actions. *Frontiers in Neurobotics*, 7:19.
- [13] B.A. Urgan, S. Pehlivan, A.P. Saygin (2019). Distinct representations in occipito-temporal, parietal, and premotor cortex during action perception revealed by fMRI and computational modeling. *Neuropsychologia*, 127, 35-47.
- [14] F.P. de Lange, M. Heilbron, P. Kok (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, 22 (9), 764-779.