The Impact of Robot's Body Language on Customer Experience: An Analysis in a Cafe Setting

Nguyen Tan Viet Tuyen
Department of Engineering, King's College London
London, United Kingdom
tan_viet_tuyen.nguyen@kcl.ac.uk

Mateusz Adamski*
Yuchen Wang*
Department of Informatics, King's College London
London, United Kingdom
{mateusz.adamski,yuchen.3.wang}@kcl.ac.uk

Shintaro Okazaki
King's Business School, King's College London
London, United Kingdom
shintaro.okazaki@kcl.ac.uk

Oya Celiktutan
Department of Engineering, King's College London
London, United Kingdom
oya.celiktutan@kcl.ac.uk

Abstract

Nonverbal communication plays a crucial role in human-robot interaction (HRI) and have been widely used for robots in service environments. While few studies have addressed the understanding customer's acceptance of robots under many different interaction conditions, the impact of robots' nonverbal interaction modalities (i.e., a combination of body language, voice, and touch) on customers' experience has not been investigated truly. To this end, in this paper, we introduce an HRI framework that aims to assist customers in their food and beverage choices in a real-world cafe setting. With this framework, the contribution of this paper are two folds. We introduce a time-synchronised multisensory HRI dataset comprising the interactions between a social robot and customers in a real-world environment. We conduct a user study to evaluate the configuration of multimodal HRI framework, particularly nonverbal gestures, and its contribution to customers' interaction experience in this specific marketing setting.

CCS Concepts

• **Human-centered computing** → *Human computer interaction (HCI)*; **HCI design and evaluation methods**; *Human computer interaction (HCI).*

Keywords

human-robot interaction, nonverbal behaviour generation, multimodal dataset, hospitality environments

ACM Reference Format:
Nguyen Tan Viet Tuyen, Mateusz Adamski, Yuchen Wang, Shintaro Okazaki, and Oya Celiktutan. 2023. The Impact of Robot's Body Language on Customer Experience: An Analysis in a Cafe Setting. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3568294.3580082

*Both authors contributed equally to this research.

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9970-8/23/03…$15.00
https://doi.org/10.1145/3568294.3580082

1 Introduction

From restaurants to shops, social robots are envisioned to have a profound impact on many hospitality sectors. We have already seen many examples of robots making their way into real-world service environments. The use of social robots (e.g., Pepper) for marketing purposes has been investigated in various application domains, from selling Nescafe machines [10] to providing assistance in chocolate stores [2], shopping malls [12], and bakery shops [14]. Most of the works in the literature have focused on developing effective recommendation strategies for robots [5], for instance, combining the information from the physical world (e.g., in-situ customer feedback) with online knowledge databases [3]. A couple of works have developed shopping companion robots, for instance, to assist users in finding shops passed by while navigating a shopping mall [9]. Another work has investigated the location of the robot in a chocolate shop, namely, inside or outside the store [2]. However, to the best of our knowledge, few works have addressed the validation of the impact of robots' nonverbal communication capabilities, particularly robots' communicative gestures, on customer interaction experience.

To address this gap in the literature, this paper introduces an HRI framework enabling a robot to communicate with customers via different interaction modalities in a cafe shop setting. The framework is configured in three different versions, aiming to validate the robot's nonverbal features comprehensively. A total of 171 customers participated in our study and subjective evaluation was carried out to measure the user perception of the robot's behaviors. This paper presents an analysis based on quantitative and qualitative customer data, aiming to shed a light into the impact of robot nonverbal communication skills on customer interaction experience in service environments. The paper also contributes a 7 hours time-synchronized multisensory HRI dataset obtained in a crowded cafe environment, which will benefit researchers in the HRI domain[1].

[1]Anonymised data is shared upon request. Please contact Oya Celiktutan at oya.celiktutan@kcl.ac.uk.

2 Nonverbal Communication Skills in Robots

People tend to use a wide range of nonverbal cues to signal their emotions, intentions, or verbal contents of their speech to their interaction partners. Similarly, communicative gestures encourage robots to better convey verbal contents of their speech to enhance their user's acceptance, trust, and engagement. A considerable effort has gone into the designing of nonverbal interaction skills for social robots. For humanoid robot platforms, nonverbal cues are commonly inspired by human behaviors. One of the main reasons is to ensure the communicative messages, encoded in robots' body movements, are interpretable by humans [13]. Previous works on nonverbal generation can be briefly categorized into two groups: (1) rule-based approach and (2) data driven-approach.

2.1 Rule-based approach

Early studies investigated this approach for building robots' communicative behaviors [1, 8]. The method requires the design of interaction logic manually. It is limited and is not transferable to unforeseen interaction contexts. Regardless of its limitations, most of the existing nonverbal generation frameworks embedded in commercial robots are built upon the rule-based approach due to its simplicity. Indeed, handcrafted gestures ensure the smoothness and human-likeness of robots' motions.

2.2 Data-driven approach

Data-driven approaches provide a solution to transfer human nonverbal communication skills to robots in an end-to-end manner. Using large-scale datasets of human communicative behaviors [16, 17], different learning frameworks have been introduced to capture the relationship between human audio [6, 11], speech text [15] and human co-speech gestures. The learning framework could be constructed in various ways, ranging from auto-regressive [7] to encoder-decoder [6] and generative adversarial networks [11, 15]. Although the data-driven approach is promising, it is still in its infancy due to the lack of universal datasets (i.e., that comprise people with diverse cultural backgrounds, age groups, and personalities) and efficient onboard computation resources for operating in the wild. Compared with the rule-based solution, the data-driven approach has not been widely implemented in social robots operating in the wild. To remedy this problem, in this paper, we propose a hybrid approach that combines these two aforementioned approaches.

3 HRI Framework

Fig. 1 shows the designed HRI framework called "Pepper the Recommender". Pepper verbally interacts with customers through a dialog system to recommend available products at the cafe. The robot also exhibits nonverbal gestures to communicate its verbal messages better. The tablet is applied for showing the verbal contents and for receiving touch-screen responses from customer. In the sequel, the proposed framework is explained in more detail.

3.1 Nonverbal Interaction

In the designed HRI framework, robot's nonverbal gestures are generated using either the rule-based or the hybrid approach.

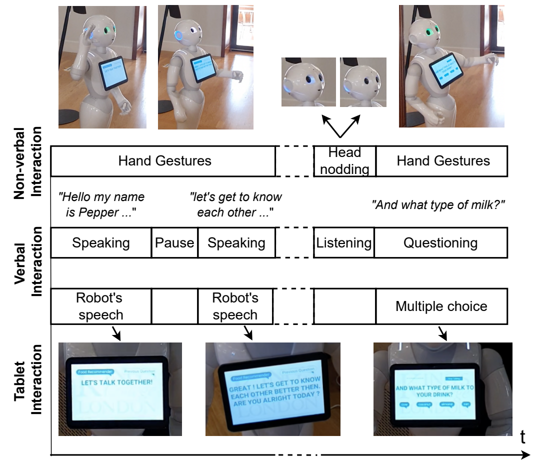

Figure 1: Pepper the Recommender system comprises three interaction modalities, namely, nonverbal, verbal, and tablet interaction.

3.1.1 Rule-based approach Most modern social robots rely on the rule-based strategy to build robots' communication gestures due to its simple and practical properties. In this paper, we apply the robot's off-the-shelf *ALAnimatedSpeech* module, which helps the robot perform upper body gestures to support the verbal contents of its speech. *ALAnimatedSpeech*[2] consists of a set of robot gestures handcrafted to ensure the smoothness and human-likeness of the motions. In our designed framework, *ALAnimatedSpeech* receives raw text delivered by the dialog system, as discussed in Section 3.2, and produces an output robot motion. *ALAnimatedSpeech* selects actions from a pre-defined list of robot gestures and associates them with the robot's speech, being its co-speech gestures.

3.1.2 Hybrid approach We implemented a hybrid approach consisting of a *Gesture Generator* network G and a *Gesture Knowledge Base* K. The approach could be considered as an enhancement of the data-driven solution introduced in [15]. This method has been proven to generate gestures aligned with semantic content as compared to the related works. In our hybrid approach, G is implemented based on the GAN network introduced in [15]. The approach is trained on the large-scale MSR-VTT dataset [16] for capturing the relationship between human motion and language in an end-to-end manner. At the training phase, G tries to produce a co-speech gesture a_f as much similar as to a_r to fool *Gesture Discriminator* network D, while D tries to maximize its ability to differentiate between real action a_r and fake action a_f. To strengthen the data-driven approach, we implemented the *Gesture Knowledge Base* K. K is inspired by a set of gesture animations manually designed for the Pepper robot[3]. We then associate those handcrafted gestures with appropriate keywords.

At the generation phase, a raw text, which is released from the dialog system, is split into sentences. If certain predefined keywords are detected from the input sentence, an associated robot nonverbal gesture r_k is selected from the *Gesture Knowledge Base* K. Otherwise, the input sentence is encoded into an embedding vector e and

[2]http://doc.aldebaran.com/2-5/naoqi/audio/alanimatedspeech.html#alanimatedspeech
[3]http://doc.aldebaran.com/2-4/naoqi/motion/alanimationplayer-advanced.html#alanimationplayer-advanced

fed to the G. G produces a human co-speech action a_f represented by joint coordinates defined in human motion space. The generated human action a_f is converted to robot gesture r_f represented by a set of joint angles. Finally, robot gestures released from K and/or G are injected into the time synchronization module with their corresponding speech s before portraying on the robot. Taken together, the hybrid approach benefits from the motion library handcrafted for the Pepper robot as K allows the robot to convey certain keywords of its speech. At the same time, G is trained on a large-scale dataset of human communicative gestures; so the robot can produce human-like gestures in various communication contexts.

3.2 Verbal and Tablet Interaction

In order to control random factors that might affect the user's perception of a robot in the cafe setting (e.g., noisy backgrounds, biased verbal responses, etc.), a basic dialogue is designed, it involves greeting and food recommendations. We construct the dialog system using Qichat[4], which is built in Choregraphe NAOqi[5]. Each session covers a list of dynamic concepts. Here, a concept is defined as a list of keywords and phrases representing a particular idea (e.g., coffee, milk, etc.). Constructing the dialog based on concepts allows the robot to diversify its speech while ensuring the semantic contents remained unchanged.

When the robot is speaking, the tablet mirrors its verbal content. During question-answer interactions, the tablet organizes it as a multiple-choice question. The dialog system accepts verbal and touch-screen responses on a first-come-first-serve basis.

3.3 Robot Conditions

To explicitly verify the impact of the robot's body languages on the customers' experience, the HRI framework was implemented on the Pepper humanoid social robot[6] in three different ways, resulting in three different robot versions.

- **Pepper 1:** The robot is equipped with the dialog system and the tablet interface mentioned in Section 3.2. Nonverbal gestures are not implemented in this version; the robot is almost at a standstill position when communicating with customers through verbal and tablet channels. This configuration serves as a baseline to examine the customers' perception of a robot with/without co-speech gesture capabilities.
- **Pepper 2:** In addition to verbal and tablet interactions used in *Pepper 1*, *Pepper 2* is designed with nonverbal gestures using the rule-based approach discussed in Section 3.1. In *Pepper 2*, the robot's gesture library is handcrafted to display smooth and human-like body gestures. With this configuration, *Pepper 2* tends to produce more beat gestures (rhythmic hand movements) to support its speech.
- **Pepper 3:** In addition to the verbal and tablet communication as in *Pepper 1*, *Pepper 3* is designed with nonverbal communication skills derived from the hybrid approach. Differently from *Pepper 2*, *Pepper 3* tends to perform more iconic gestures to convey the semantic contents of its speech.

[4]http://doc.aldebaran.com/2-5/naoqi/interaction/dialog/dialog.html#dialog-concepts
[5]http://doc.aldebaran.com/2-5/software/choregraphe/index.html
[6]http://doc.aldebaran.com/2-5/home_pepper.html

4 Study Design

4.1 Pepper the Recommender

4.1.1 Sensors and Setup The study was conducted at a cafe shop located on a university campus. The robot was placed at the cafe's entrance to draw customers' attention towards the shop [2]. We used a set of sensors to record customers' behaviors when interacting with the robot: 1) External sensors: A stereo RGB-depth camera (i.e., ZED) was placed behind the robot to record the interactions from a third-person perspective. Additionally, from RGB-depth images recorded by the external camera, human motions are extracted and stored as $3D$ joint coordinates. 2) Onboard sensors: Microphones[7], RGB camera[8], and angle sensors embedded in the robot were implemented to capture the interaction from a first-person perspective. From the robot's angle sensors, robot motions are acquired and stored as a list of joint angles.

4.1.2 Participants and Procedure Participants were customers of the university cafe shop, mostly students, members of staff, and visitors. Upon approaching the entrance, customers, who showed their interest in the study, were provided with the information sheet with a full explanation of the research objectives, procedure, and anonymity of the data collected. If they agreed to participate in the study, after giving their consent, they engaged in an interaction with one of the autonomous robot versions, namely, static Pepper (*Pepper 1*) or dynamic Pepper (*Pepper 2* or *3*). The study was approved by the King's College London Research Ethics Committee.

4.1.3 Data Statistics The study was conducted in 3 weeks with a total of 171 participants (74 males, 92 females, and 5 non-binary), resulting in 57 participants for each robot version. 85% of participants were undergraduate and postgraduate students. 78% participants were in the age group $18 - 24$. We collected a time-synchronized multimodal dataset of HRIs during approximately 7 hours.

4.2 Subjective Evaluation

After the interaction took place, a study of subjective evaluation was conducted to investigate the user interaction experience with the three robot conditions (i.e., Pepper 1-3) in light of two constructs, including time consistency of gestures and semantic content of gestures. Each construct includes two question items. All items were measured with a 5-point Likert scale from 1 being "strongly disagree" to 5 being "strongly agree". We derived those scale items from previous works [4, 6]. Time consistency is used to verify the time synchronization between robots' gestures and speech. On the other hand, semantic content is applied to validate the effectiveness of robot body language in conveying verbal contents of its speech [6].

4.3 Hypothesis

With the customer behavior data collected during interaction, we counted the number of touch-screen responses to check whether they rely on tablet to interact with the robot, especially *Pepper 1*, a standstill robot. In other words, we examined whether customers tend to perceive *Pepper 1* as a tablet kiosk, so they interact with this

[7]http://doc.aldebaran.com/2-5/family/pepper_technical/microphone_pep.html
[8]http://doc.aldebaran.com/2-5/family/pepper_technical/video_2D_pep.html

robot through the tablet more frequently than in the case of *Pepper 2* and *3*. On this basis, we contemplate:

- **H1:** The number of touching events measured from the tablet of *Pepper 1* will be greater than those in *Pepper 2* and *Pepper 3*.

In terms of the robot's nonverbal communication, *Pepper 2* and *Pepper 3* are equipped with different gesturing skills. While *Pepper 2* tends to produce rhythmic movement of hands to support its speech, *Pepper 3* emphasizes semantic hand gestures to convey the verbal content. Regardless of the differences in gesture styles, the gesture timing and the robot's speech are expected to be well-matched in both *Pepper 2* and *Pepper 3*. More formally:

- **H2:** Temporal consistency between gestures and speech in *Pepper 2* and *Pepper 3* will not be different.

Pepper 3 is designed with the hybrid approach. This configuration is expected to provide *Pepper 3* a better ability to convey semantic contents of its speech via generated nonverbal gestures. On this basis, we contemplate the following:

- **H3:** The semantic content encoded in the gestures of *Pepper 3* will be greater than that in *Pepper 1* and *Pepper 2*.

5 Results and Discussions

5.1 Objective Evaluation

As a preliminary objective evaluation, we analysed participants' movements in terms of velocity and their distance to the robot. We first calculated the average velocity of upper-body movements and then calculated the frequency of velocity values over time, when they were interacting with three different robot conditions. Concerning velocity, the distributions of human movements were not different among participants interacting with a static robot (*Pepper 1*) versus dynamic robots (*Pepper 2*, *Pepper 3*). On another note, it was observed that there were two distinct groups of participants, which can be categorized with respect to their proxemics behaviors. Experimenters observed that some participants tended to interact with the robot at a distance, while the others held a personal distance, even sometimes an intimate one. One of the possible reasons could be the differences in their attitude, familiarity, and comfort towards interacting with a humanoid robot in the service environment.

5.2 Subjective Evaluation

Overall, customers tend to communicate with robots through its tablet more than one time during their interaction. Although the mean values are slightly higher than in *Pepper 1* and *Pepper 3* as compared to *Pepper 2*, the post-hoc test analysis in Table 1 indicates that there are no statistically significant differences regarding the number of touch-screen events concerning different robot conditions. Thus, *H1* was not supported by our data.

There are statistically significant differences in both time consistency and semantic content of robot gestures. The post-hoc test indicates significant differences ($p < 0.05$) in time consistency between *Pepper 1* and *Pepper 2*, *Pepper 1* and *Pepper 3*, but no statistical difference between *Pepper 2* and *Pepper 3*. Table 1 reveals that customers observed that the semantic content encoded in co-speech gestures of *Pepper 3* are greater than in *Pepper 1*'s gestures. Similarly, *Pepper 2* performed gestures to express semantic content of its speech greater than *Pepper 1*. However, there are no statistical differences between *Pepper 2* and *Pepper 3* in terms of semantic content. Thus, our data provided support for *H2* but not for *H3*.

Table 1: Preliminary analysis results (P1 = Pepper 1; P2 = Pepper 2; P3 = Pepper 3). * is based on p < .05.

	P1	P2	P3	F	p	Post-hoc test*
Touch-screen events	1.75	1.46	1.80	0.103	0.901	-
Time consistency of gestures	3.39	3.94	3.72	5.724	.004	P1<P2 P1<P3
Semantic contents of gestures	3.12	3.77	3.57	8.635	<.001	P1<P2 P1<P3

5.3 Discussion

The frequency of customers' touch-screen responses provides a clue for understanding how frequently customers rely on this channel to communicate with the robot rather than through verbal messages. While *H1* was not supported, it implies the way how users perceive Pepper humanoid robot in service environments. Even when customers interact with a standstill robot (*Pepper 1*), without nonverbal gesture skills, they are still curious to use verbal feedback to actively interact with this robot rather than treating it as a tablet kiosk and relying on touch-screen communication only.

On the other hand, *H2* was supported while *H3* was not supported. This result implies that customers could not observe the differences in body language skills between *Pepper 2* and *Pepper 3*. It could be explained by considering a crowded environment at the cafe shop with many random factors and distractions (e.g., noise, passersby, etc.). Consequently, customers might not be provided enough space to comprehensively observe distinct behaviors exhibited by the robot, particularly the semantic content encoded in the robot's communication gestures. The finding suggests that, in crowded service interaction settings, the power of the hybrid approach is not fully acknowledged by customers, while the rule-based strategy seems to be simple but practical.

6 Conclusion

In this paper, we have presented a preliminary analysis of the relationship between robot's nonverbal communication and customer experience in service environments. We have demonstrated a case study where an HRI framework was developed to operate in a cafe shop setting. Finally, we have introduced a 7 hours time-synchronized multisensory HRI dataset collected in a crowded environment. As a future work, we will extend our objective and subjective evaluations using the data collected.

Acknowledgment

This work was supported by The Royal Society project CL-HRI (Grant Ref.: RGS\R2\212084), the EPSRC project LISI (Grant Ref.: EP/V010875/1), and King's Undergraduate Research Fellowship scheme. The authors thank the cafe managers and personnel for their support to their research.

References

[1] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*. Springer, 163–185.

[2] Laurens De Gauquier, Malaika Brengman, Kim Willems, Hoang-Long Cao, and Bram Vanderborght. 2021. In or out? A field observational study on the placement of entertaining robots in retailing. *International Journal of Retail & Distribution Management* 49, 7 (2021), 846–874.

[3] Roland Graef, Mathias Klier, Andreas Obermeier, and Jan Felix Zolitschka. 2022. What to buy, pepper?–Bridging the physical and the digital world with recommendations from humanoid robots. *Journal of Decision Systems* (2022), 1–27.

[4] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 79–86.

[5] Koji Kamei, Kazuhiko Shinozawa, Tetsushi Ikeda, Akira Utsumi, Takahiro Miyashita, and Norihiro Hagita. 2010. Recommendation from robots in a real-world retail shop. In *International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction*. 1–8.

[6] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 97–104.

[7] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 242–250.

[8] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics symposium on computer animation*. 25–35.

[9] Takahiro Matsumoto, Satoru Satake, Takayuki Kanda, Michita Imai, and Norihiro Hagita. 2012. Do you remember that shop? computational model of spatial memory for shopping companion robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 447–454.

[10] Nestlé. 2014. *Nestlé to use humanoid robot to sell Nescafé in Japan.* Retrieved Oct 1, 2022 from https://www.nestle.com/media/news/nestle-humanoid-robot-nescafe-japan

[11] Tan Viet Tuyen Nguyen and Oya Celiktutan. 2022. Agree or Disagree? Generating Body Gestures from Affective Contextual Cues during Dyadic Interactions. In *International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE.

[12] Marketta Niemelä, Päivi Heikkilä, and Hanna Lammi. 2017. A social service robot in a shopping mall: expectations of the management, retailers and consumers. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 227–228.

[13] Shane Saunderson and Goldie Nejat. 2019. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics* 11, 4 (2019), 575–608.

[14] Sichao Song, Baba Jun, Junya Nakanishi, Yuichiro Yoshikawa, and Hiroshi Ishiguro. 2022. Service Robots in a Bakery Shop: A Field Study. *arXiv preprint arXiv:2208.09260* (2022).

[15] Nguyen Tan Viet Tuyen, Armagan Elibol, and Nak Young Chong. 2020. Conditional generative adversarial network for generating communicative robot gestures. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 201–207.

[16] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.

[17] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4303–4309.