The More I Understand it, the Less I Like it: The Relationship Between Understandability and Godspeed Scores for Robotic Gestures

Amol Deshmukh, Bart Craenen, Mary Ellen Foster and Alessandro Vinciarelli

Abstract— **This work investigates the relationship between the perception that people develop about a robot and the understandability of the gestures the latter displays. The experiments have involved 30 human observers that have rated 45 robotic gestures in terms of the Godspeed dimensions. At the same time, the observers have assigned a score to 10 possible interpretations (the same interpretations for all gestures). The results show that there is a statistically significant correlation between the understandability of the gestures — measured through an information theoretic approach — and all Godspeed scores. However, the correlation is positive in some cases (Anthropomorphism, Animacy and Perceived Intelligence), but negative in others (Perceived Safety and Likeability). In other words, higher understandability is not necessarily associated with more positive perceptions.**

I. Introduction

The main question addressed in this work is whether there is a relationship between the perception that people develop about a robot and the understandability of its gestures, where the word "*understandability*" means how unambiguous the meaning of the gestures is for human observers. In addition, the experiments of this work investigate the relationship between two major parameters underlying a gesture — speed and amplitude — and its understandability. The main reason for focusing on such problems is that gestures are the most reliable communication channel in environments in which the level of acoustic noise is high and, hence, the use of speech (or other audio signals) is difficult, if not impossible [1], [2]. In line with this observation, the experiments of this work focus on *symbolic gestures* that "*[...] often are used to communicate when distance or noise renders vocal communication impossible [...] expressing concepts that also are expressed verbally*" [3].

Far from being rare, the noise conditions above are typical of many everyday settings in which robots are likely to play a major role in the future like, e.g., shopping malls, airports, stations and other public spaces. In these contexts, robots should display gestures as understandable as possible because they compete with other stimuli designed to attract and retain attention (advertisement, danger warnings, public announcements, etc.). On the other hand, the literature suggests that human users do not necessarily prefer robots that gesture better: "*[...] participants perceived [the robot] as more likeable [...] this effect was particularly pronounced when the robot's gestures were partly incongruent with speech [...]*" [4].

For the reasons above, the analysis of the relationship between understandability and perception can help to avoid the synthesis of gestures that, while effectively communicating a desired message, might result in negative impressions. Furthermore, the analysis of the effects due to changes in amplitude and speed can help to synthesize gestural stimuli that, while keeping the impressions of the users sufficiently positive, are still effective at conveying their messages. During the experiments, 30 independent observers were asked to watch 45 different gestures performed by *Pepper*, a robotic platform manufactured by Softbank Robotics. The stimuli were obtained by manipulating *speed* and *amplitude* of 5 standard animations available in the library of the robot. For each stimulus, the 30 observers were asked to perform two tasks, namely to complete the Godspeed questionnaire [5] and to rate 10 possible predefined interpretations (higher ratings are attributed to interpretations the observers considered to be more correct).

The understandability has been measured with a function of the *relative entropy* [6], an information theoretic quantity that depends on how uniformly the ratings of the observers distribute across the possible interpretations. The main advantage of using relative entropy is that the quantity does not depend on the interpretation of a gesture, but on whether the different observers tend to give the same interpretation or not; meaning whether the gesture is actually understandable, or not. This is important because the manipulation of amplitude and speed is designed to add noise and, hence, to generate gestures that do not necessarily have a predefined meaning or are difficult to interpret.

The results in this paper show that there is a statistically significant correlation between all Godspeed scores and understandability, thus confirming that the latter contributes to the overall perception that the observers develop about the robot. However, while the correlation is positive in the case of Anthropomorphism, Animacy and Perceived Intelligence, it is negative in the case of Likeability and Perceived Safety. In other words, at least when it comes to certain dimensions of the Godspeed questionnaire, the understandability of a gesture can be achieved only at the expense of having positive impression of the robot. For what concerns the effect of amplitude and speed, the results show that the latter does not change significantly the understandability of gestures, while the amplitude does. Therefore, such a parameter must be tuned more carefully to ensure that a gesture is understandable enough to fulfil its communicative function.

The rest of this paper is organized as follows: Section II

The authors are with the University of Glasgow (School of Computing Science) — University Avenue, G12 8QQ Glasgow (UK), email: <firstname.lastname@glasgow.ac.uk>.

surveys previous work on this subject; Section III describes the process adopted to define and synthesize the 45 stimuli used in the experiments; Section IV introduces the notion of understandability and the approach adopted to measure it; Section V reports on experiments and results and while the final Section VI draws some conclusions.

II. Survey of Previous Work

Several works in literature address the role of gestures in Human-Robot Interaction. In most cases, the starting point is the observation that gestures are an essential component of non-verbal communication in Human-Human exchanges [7], [8]. Therefore, it should be possible to synthesize gestures aimed at enriching Human-Robot Interactions with layers of socially and psychologically relevant information, in the same way as natural gestures do when people communicate with one another [9]. In other cases, the focus is on *deictic gestures*, i.e., gestures that attract the attention of the users towards objects in the environment. Besides being useful from a practical point of view, these gestures have the advantage of fostering joint attention between robots and their users, a prerequisite necessary for establishing effective interactions.

The experiments proposed in [10] show that people recognize cooperative gestures and that robots displaying them tend to establish more effective collaborations. This happens in particular when the gestures are abrupt and oriented towards the front of the robot. Furthermore, there is a correlation between the tendency to recognize and accept the cooperative gestures of the robot and the ability to recognize human gestures. Similarly, the experiments presented in [11] show that the use of synthetic gestures during robot storytelling is predictive of how well the listeners remember the details of the stories. The use of gestures to improve the performance in a task is the subject of the experiments in [12] as well. In particular, this work shows that the users better understand what a robot says when the latter imitates their gestures, thus showing entrainment. Finally, the experiments described in [13] show that synthetic gestures can increase the engagement of people involved in an interaction with robots, while the approach proposed in [14] shows that humans can interpret synthetic gestures in terms of emotions.

Regarding deictic gestures, the approach proposed in [15], [16] aims at attracting the attention of the users to objects in the environment. The experiments show that the users understand what the targets of the robot's deictic gestures are. In the case of the experiments proposed in [17], it is the robot that recognizes the target of a deictic gesture displayed by a human user through the multimodal analysis of speech and actual gestures.

To the best of our knowledge, the only work on how recognizable the synthetic gestures of a robot are is presented in [18]. There, the experiments revolve around 15 stimuli that are recognized by human observers with an accuracy that ranges between roughly 10% and almost 100%. The main finding of the article is that the limited number of Degrees of Freedom in the robots makes them unable to perfectly imitate human gestures and, hence, the agreement between observers is, on average, around 60%. The main difference between that paper and the work presented here is that here the focus shifts from recognition rate to understandability, i.e., from the ability of human observers to recognize what a gesture is expected to mean, to the tendency of human users to attribute the same meaning to the same gesture. Furthermore, this work analyses the relationship between understandability and users' subjective ratings.

III. The Stimuli

Gestures are "*movements of the body (or some part of it) used to communicate an idea, intention or feeling*" [7]. The process for the definition of the stimuli revolves around *emblems*, the gestures that are "*used intentionally by the sender to communicate a specific message to an individual or group [...] in many cases, to substitute for the spoken word(s)*" [8]. For this reason, the process aimed at the synthesis of the stimuli for the experiments starts with the selection of 5 animations — the *core gestures* hereafter — available in the standard library of the *Pepper*, the robotic platform used in this work. According to the documentation accompanying the robot, these gestures are designed to convey the following messages[1]:

- Disengaging / Send-away;
- Engaging / Gain attention;
- Pointing / Giving Directions;
- Head-Touching / Disappointment;
- Cheering / Success.

The inclusion of two pairs of gestures that convey messages opposite to one another — Engaging vs Disengaging and Cheering vs Disappointment — aims at limiting, as much as possible, ambiguity and confusion between the meaning of the different core stimuli.

The rest of the process aims at adding noise to the core gestures above and, as a consequence, to synthesize gestures of varying understandability. In particular, two major parameters of gestures — speed λ and amplitude α — have been manipulated to generate 9 different variants for each of the 5 core gestures, thus resulting into the 45 stimuli adopted during the experiments. Each core gesture has been synthesized using three different values of λ, namely 15, 25 and 35 *frames per second* (*fps*), where 25 *fps* is the original speed of the core gestures. Furthermore, for each value of λ, the difference $\Delta_i(t) = \theta_i(t) - \theta_i(t-1)$ — where $\theta_i(t)$ is the angle between the two mechanical elements connected by joint i — has been multiplied by three different values of α — 0.50, 0.75 and 1.00 — for all values of i (meaning all joints) and t (meaning all frames).

The result of the process is a set of 45 stimuli (independent variable for the study) that can be represented as triples

[1] The animations associated to the core stimuli are available on the version 1.6B of Pepper in the following directories: "*animations/Stand/Gestures/No_3*" (Disengaging), "*animations/Stand/Gestures/Hey_2*" (Engaging), "*animations/Stand/Emotions/Negative/Hurt_1*" (Head-Touching), "*animations/Stand/Gestures/Far_3*" (Pointing); and "*animations/Stand/Emotions/Positive/Happy_1*" (Cheering).

Age Range	18-22	23-25	26-30	31-35	36-40	>40
No. of Subjects	11	6	6	3	1	3

TABLE I

AGE DISTRIBUTIONS OF THE SUBJECTS INVOLVED IN THE EXPERIMENTS.

(k, α, λ), where $k \in [1, \dots, 5]$ is an index that accounts for the core gesture the stimulus derives from, α is the amplitude and λ is the speed. The triples in which $\alpha = 1.00$ and $\lambda = 25$ correspond to the core gestures.

A. Annotation

The 30 observers involved in the experiments (see Section V for more details) have watched the 45 stimuli and, for each of them, they have filled the Godspeed questionnaire [5] (all observers have watched and rated all stimuli, first dependent variable for the study). The goal of the Godspeed questionnaire is to measure, in quantitative terms, the perception that people develop about a robot they observe or interact with. In particular, the questionnaire allows one to rate the robot along the following dimensions:

- *Anthropomorphism*: tendency of human users to attribute human characteristics to a robot;
- *Animacy*: tendency of human users to consider the robot alive and to attribute intentions to it;
- *Likeability*: tendency of human users to attribute desirable characteristics to a robot;
- *Perceived Intelligence*: tendency of human users to consider the behaviour of a robot intelligent;
- *Perceived Safety*: tendency of human users to consider the interaction with a robot safe.

The stimuli were administered in random order and, to avoid tiredness effects, they were split in three groups of 15 that were rated in three separate sessions — one hour long each — held over three consecutive days.

The observers were selected randomly among those responding to a call for participation distributed at the research institute where the experiments of this work were carried out. The resulting pool of observers includes 10 women and 20 men of different ethnic and national origin, their age distribution is available in Table I. Only 3 of the $N = 30$ observers had interacted with a robot before having been involved in the experiments of this work. The payment for participation in the experiment was the minimum legal hourly wage in the country where the experiment was carried out.

In addition to filling out the questionnaire, the observers rated $T = 10$ possible meanings that can be attributed to the stimuli (second dependent variable for the study). In particular, the observers assigned a score between 0 and $L = 4$ to each meaning, with higher scores being attributed to meanings considered more correct. The 10 possible interpretations are the same for all stimuli and are as follows: *Getting Distracted*; *Aggressing*, *Flirting*, *Pointing*, *Complaining*, *Cheering*, *Reflecting*, *Teasing*, *Rejecting* and *Welcoming*. Five of these interpretations correspond, according to the documentation provided by the robot's manufacturer, to the actual meaning of the core gestures, while the others were selected to be as different as possible from each other and from the meaning of the core gestures.

IV. UNDERSTANDABILITY AND ITS MEASUREMENT

The rating approach adopted to score the possible meanings of a stimulus (see Section III-A) accounts for the role of emblems — the particular class of gestures the core stimuli belong to — as *codified signals*, i.e., as signals that are "*steadily linked to a meaning, so that the two make a signal-meaning pair [. . .] like it happens, for instance, with the lexical items of a verbal lexicon*" [19]. For this reason, we measure understandability as a function of entropy [6], an information theoretic quantity that accounts for how uniformly the scores of the observers distribute across the possible interpretations of every stimulus.

For each of the 45 stimuli, the result of the meaning rating process (see end of Section III-A) is a matrix $M = \{m_{ij}\}$, where m_{ij} is the score that observer i (where $i = 1, \dots, N$) assigns to interpretation j (where $j = 1, \dots, T$). The following sum can be interpreted as the total number of votes that, for a particular stimulus, interpretation j has received:

$$u_j = \sum_{i=1}^{N} m_{ij}. \tag{1}$$

Following the above, the probability p_k of interpretation k receiving a vote can be estimated as follows:

$$p_k = \frac{u_k}{\sum_{i=1}^{N} \sum_{j=1}^{T} m_{ij}}, \tag{2}$$

where the numerator is the sum over all elements of matrix M. This makes it possible to estimate the *relative entropy* (Kullback-Leibler divergence) of the distribution in the following terms:

$$H_r = \frac{-\sum_{j=1}^{T} p_j \log p_j}{\log T}, \tag{3}$$

where $H_r \in [0, 1]$. The value of H_r is 0 when all votes have been attributed to one particular interpretation k ($p_k = 1$ and $p_j = 0$ for $j \neq k$), while it is 1 when all interpretations have received the same number of votes. In other words, the relative entropy is a measure of uncertainty that ranges between 0 (full certainty) and 1 (full uncertainty). In this way, the following function U can be interpreted as a measure of understandability:

$$U = 1 - H_r. \tag{4}$$

In fact, $U = 1$ corresponds to a situation in which one interpretation attracts scores different from 0 and all the others attract only null scores (the gesture is unambiguously understandable), while $U = 0$ corresponds to a situation in which all interpretations receive the same rating (the gesture is too ambiguous to be understandable).

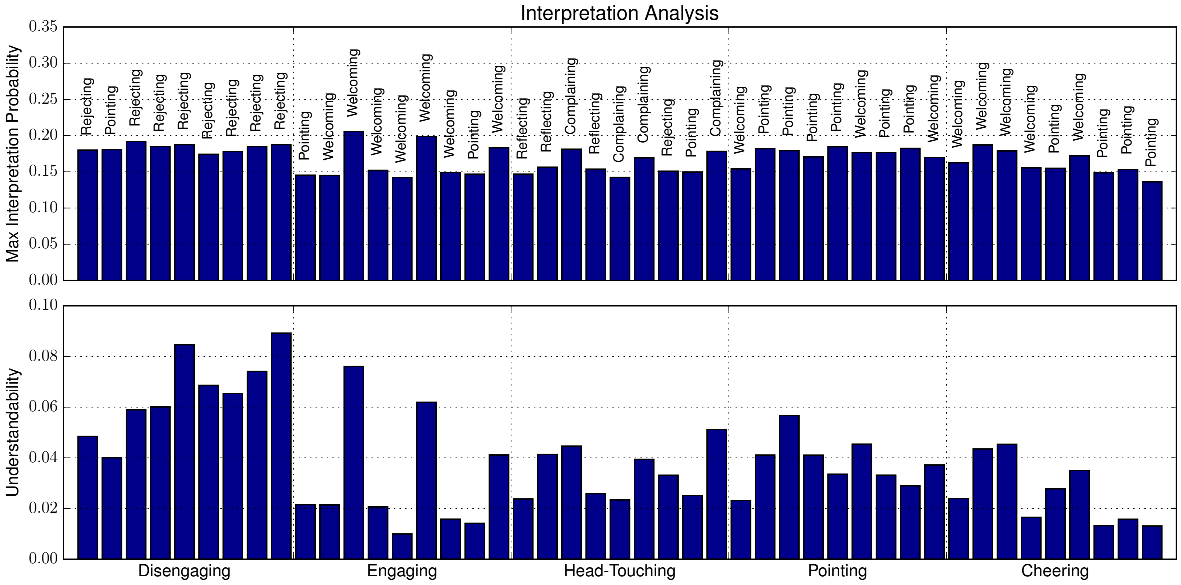

Fig. 1. The upper chart shows the value of $p_{\hat{k}}$ (the probability if the gesture's interpretation that has attracted more votes) and, correspondingly, the most probable interpretation. The lower chart shows the relative entropy associated to the individual stimuli. Each bar corresponds to an individual stimulus; the labels on the bottom group all versions of a given core stimulus.

V. Experiments and Results

Section IV proposes a measure of understandability that is expected to depend not on whether the observers understand the actual meaning of a gesture, but whether they all tend to assign the gesture the same meaning. The upper chart of Figure 1 shows that, for each of the 45 stimuli, it is possible to identify the meaning that has received the largest number of votes or, according to the notation of Section IV, the meaning $\hat{k}$ that satisfies the following equation:

$$\hat{k} = \arg \max_{k \in [1,T]} p_k, \tag{5}$$

where $\hat{k}$ is the index of the most probable meaning according to the ratings of the observers. The chart shows, for each stimulus, the value of $p_{\hat{k}}$ and the corresponding interpretation, while the lower chart shows the understandability value. The interpretations are not always coherent with the meaning of the core gesture a given stimulus derives from. The stimuli that are interpreted more similarly to their respective core gestures are *Disengaging* (8 out of 9 variants are interpreted as *Rejecting*), *Engaging* (7 out of 9 variants are interpreted as *Welcoming*) and *Pointing* (6 out of 9 variants). In the case of *Head-Touching*, the most frequent interpretation is *Complaining* (4 out of 9 variants), and, in the case of *Cheering*, it is *Welcoming* (5 out of 9 variants).

The lower chart of Figure 1 shows that the highest values of the understandability correspond to the variants of the *Disengaging* core gesture. This seems to suggest that a higher U corresponds to gestures that are recognized more often. However, some of the lowest understandability values can be observed in the case of the other core gesture that is correctly recognized, namely *Engaging*. This confirms that U is independent of the actual meaning of the gesture and it accounts only for how certain the different observers are in assigning meaning to a gesture. In the case of *Engaging*, the observers do rate the actual meaning of the core gesture higher, but they do so with less certainty (meaning that they rate other interpretations high as well). Similarly, the interpretation of some stimuli does not correspond to the underlying core gesture, but the understandability is high (e.g., the *Head-Touching* U values are higher, on average, than the *Engaging* ones).

The observations above therefore confirm that U captures the understandability of a gesture, defined as the property of conveying a message that observers tend to agree upon, rather than the ability of the observers to understand what the core gestures underlying the stimuli mean. This is in line with the goals U has been defined for.

A. *Effect of Speed and Amplitude*

Section III shows that the process for the synthesis of the stimuli includes the manipulation of speed λ and amplitude α with the goal of producing gestures of different understandability U. Figure 2 shows the average of U over all stimuli that share the same values of λ and α. The chart shows that the highest understandability values are observed for the core gestures ($\alpha = 1.00$ and $\lambda = 25\ fps$) and for the stimuli for which $\alpha = 1.00$. This suggests that changing the speed of a gesture does not make it more difficult to understand its meaning. Vice versa, by changing the amplitude, the average understandability values decrease by up to 40% with respect to the core gestures.

Overall, Figure 2 confirms that the approach adopted for the synthesis of the stimuli has been effective in generating gestures of different understandability. Furthermore, the figure suggests that the understandability of a gesture depends

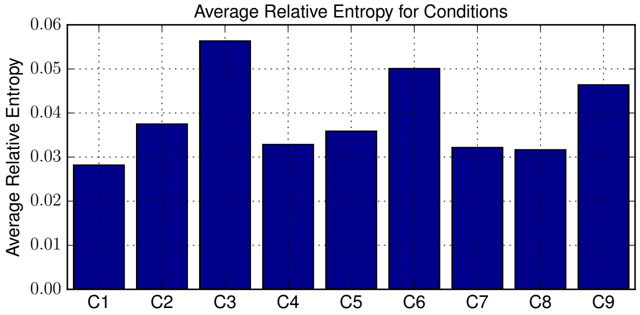

Fig. 2. Average relative entropy across all core stimuli for different (α, λ) pairs: C1 corresponds to $(0.50, 15)$, C2 to $(0.75, 15)$, C3 to $(1.00, 15)$, C4 to $(0.50, 25)$, C5 to $(0.75, 25)$, C6 to $(1.00, 25)$, C7 to $(0.50, 35)$, C8 to $(0.75, 35)$, C9 to $(1.00, 35)$. The C6 bar corresponds to the average over the core stimuli.

	Ant	Ani	Lik	Int	Saf
ρ	0.67**	0.50**	-0.37*	0.41**	-0.34*

TABLE II

CORRELATION BETWEEN UNDERSTANDABILITY AND GODSPEED SCORES. THE DOUBLE STAR MEANS THAT THE CORRELATION IS SIGNIFICANT AT CONFIDENCE LEVEL 0.01, WHILE THE SINGLE ONE MEANS THAT THE CORRELATION IS SIGNIFICANT AT LEVEL 0.05. IN BOTH CASES FDR CORRECTION WAS APPLIED.

on its morphology — represented in this case by the values $\theta_i(t)$ — more than on its speed. This is confirmation that there is no statistically significant difference between the understandability of the gestures in conditions $C3$, $C6$ and $C9$, i.e., those that have different speeds, but do not change the values of the $\theta_i(t)$. Vice versa, that there is a statistically significant difference in terms of understandability between these gestures and the other stimuli. Both findings are statistically significant according to a t-test signed-rank test, after application of the False Discovery Rate correction [20].

B. Understandability and Godspeed Scores

Table II shows the correlation between Godspeed scores and understandability. All values are statistically significant according to a t-test signed-rank test, after application of the False Discovery Rate correction [20]. The results show that the correlation is positive in the case of Anthropomorphism, Animacy and Perceived Intelligence, but negative in the case of Likeability and Perceived Safety. In other words, higher understandability is associated with higher ratings along those dimensions that, overall, account for how effective the robot is perceived to be at performing a given task (displaying understandable gestures in this case). Vice versa, higher understandability is associated with lower ratings along those dimensions that account for how effective the robot is perceived to be at establishing interactions acceptable and satisfactory for the human user — its social skills in short. One possible explanation is that such a pattern reproduces the "*task and social-emotional role differentiation*" [21], a widely investigated effect in human-human interactions, especially when it comes to small groups of people expected to achieve a goal. The main trace of such an effect is that people that appear to be more effective at accomplishing tasks tend to be considered less competent in managing social aspects [22]. However, "*This is not to say that the task specialist will actually be disliked, but rather that his task emphasis will tend to arouse some negative feelings [...] Such feelings merely neutralize any strong positive feelings other members may hold toward him*" [23]. In other words, the negative correlations in Table II do not necessarily mean that the users do not like the robot, but simply that the perception of competence prevails.

Similar effects have been observed earlier in the literature on Human-Robot Interaction. For example, experiments aimed at collaborative decision making between people and robots show that "*[...] participants conformed more to the iCub's answers [...] about functional issues than when they were about social issues [...] the few participants conforming to the iCub's answers for social issues also conformed less for functional issues.*" [24], meaning that the subjects either trust the robot from a task point of view or from a social point of view, but not both. Similarly, experiments on collaborative work between people and robots suggest that "*efficiency is not the most important aspect of performance for users [...]*" [25], i.e., users prefer to deal with a socially adept robot than with a fully efficient one, if the two options are alternative to each other. In a similar vein, the users involved in Lego playing "*liked the faulty robot significantly better than the robot that interacted flawlessly*" [26] and, in the case of the interactions between people and companion robots, "*while significantly affecting subjective perceptions of the robot and assessments of its reliability and trustworthiness, the robot's performance does not seem to substantially influence participants' decisions to (not) comply with its requests*" [27].

VI. DISCUSSION AND CONCLUSIONS

This article revolves around the understandability of synthetic gestures displayed by a robot, where the understandability was defined as the property of being attributed the same, or similar, meaning by multiple human observers. In line with this definition of understandability, the article proposed an approach for measuring understandability based on a function of relative entropy, that does not take into account the interpretation people have of a given gesture, but only whether different people observing the same gesture tend to interpret it in the same way.

The experiments presented in the article involved 30 observers that watched and rated 45 different gestural stimuli (all observers observed and rated all stimuli). The results show that the understandability of a gesture depends on its morphology — represented in this work by the angles between mechanical elements at different joints — and not, or only to a limited extent, on its speed. Furthermore, the results show that the understandability correlates positively with certain Godspeed dimensions — Anthropomorphism, Animacy and Perceived Intelligence — and negatively with others — Likeability and Perceived Safety.

One of the main issues observed from the results is that different gestures have different robustness to changes in speed and amplitude. If it is true that 4 core gestures out of 5 attract no more than 2 different interpretations (the only exception is *Head-Touching* which attracts 4), it is true as well that only for two core gestures — *Engaging* and *Disengaging* — there is one interpretation that is attributed at least 7 times out of the 9 variants. In other words, there are three core gestures for which the interpretation changes frequently with the values of α and λ. This is important whenever gestures add variability (or noise) in order to look less mechanic and more realistic. In this respect, one direction of future work is to investigate the limits in the variability of a robotic gesture that need to be respected to avoid undesired changes of meaning attribution.

The role of noise is important, not only for how the gesture is interpreted, but also because of the relationship between understandability and Godspeed scores. In line with other results found in the literature (both in Human-Human Interaction [21]–[23], and Human-Robot Interaction [24]–[27]), improving the performance of the robot leads to lower Likeability and Perceived Safety ratings. In other words, a robot that emphasises the performance of a robot risks failing in the social aspects of an interaction. This is important because the literature shows that an effective interaction between people and robots requires the latter to be appreciated not only for how well they work, but also for how positive the perception they inspire is from a social point of view [27]. In this respect, adding noise to the gestures can be a way to find a good trade-off between the two conflicting (according to the experiments presented in this work, and previous results presented in literature) needs of having a robot that performs well while still being socially acceptable. How to find the optimal point in such a trade-off can be one of the future directions for this work.

Acknowledgements

This research has been partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 688147 (MuMMER, `http://mummer-project.eu/`).

References

[1] S. Partan and P. Marler, "Issues in the classification of multimodal communication signals," *The American Naturalist*, vol. 166, no. 2, pp. 231–245, 2005.

[2] ——, "Communication goes multimodal," *Science*, vol. 283, no. 5406, pp. 1272–1273, 1999.

[3] R. Krauss, Y. Chen, and R. Gottesman, "Lexical gestures and lexical access: a process model," in *Language and Gesture*, D. McNeill, Ed. Cambridge University Press, 2000.

[4] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 313–323, 2013.

[5] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[6] R. Gray, *Entropy and information theory*. Springer Verlag, 2011.

[7] M. Knapp and J. Hall, *Nonverbal Communication in Human Interaction*. Harcourt Brace College Publishers, 1972.

[8] V. Richmond, J. McCroskey, and S. Payne, *Nonverbal behavior in interpersonal relations*. Prentice Hall, 1991.

[9] T. Wharton, *Pragmatics and non-verbal communication*. Cambridge University Press, 2009.

[10] L. Riek, T.-C. Rabinowitch, P. Bremner, A. Pipe, M. Fraser, and P. Robinson, "Cooperative gestures: Effective signaling for humanoid robots," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2010, pp. 61–68.

[11] C.-M. Huang and B. Mutlu, "Modeling and evaluating narrative gestures for humanlike robots." in *Robotics: Science and Systems*, 2013, pp. 57–64.

[12] T. Ono, T. Kanda, M. Imai, and H. Ishiguro, "Embodied communications between humans and robots emerging from entrained gestures," in *Proceedings of the IEEE International Symposium on Computational Intelligence in Robotics and Automation*, vol. 2, 2003, pp. 558–563.

[13] C. Sidner, C. Lee, C. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1-2, pp. 140–164, 2005.

[14] H. Narahara and T. Maeno, "Factors of gestures of robots for smooth communication with humans," in *Proceedings of the International Conference on Robot Communication and Coordination*, 2007.

[15] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "Three-layered draw-attention model for humanoid robots with gestures and verbal cues," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 2423–2428.

[16] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, N. Hagita, and Y. Anzai, "Humanlike conversation with gestures and verbal cues based on a three-layer attention-drawing model," *Connection science*, vol. 18, no. 4, pp. 379–402, 2006.

[17] A. Brooks and C. Breazeal, "Working with robots and objects: Revisiting deictic reference for achieving spatial common ground," in *Proceedings of the ACM SIGCHI/SIGART conference on Human-Robot Interaction*, 2006, pp. 297–304.

[18] J.-J. Cabibihan, W.-C. So, and S. Pramanik, "Human-recognizable robotic gestures," *IEEE Transactions on Autonomous Mental Development*, vol. 4, no. 4, pp. 305–314, 2012.

[19] I. Poggi, *Mind, hands, face and body. A goal and belief view of multimodal communication*. Weidler, 2007.

[20] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B*, pp. 289–300, 1995.

[21] P. Burke, "The development of task and social-emotional role differentiation," *Sociometry*, pp. 379–392, 1967.

[22] R. Bales, "Task roles and social roles in problem-solving groups," *Readings in Social Psychology*, vol. 3, no. 43, 1958.

[23] P. Slater, "Role differentiation in small groups," *American Sociological Review*, vol. 20, no. 3, pp. 300–310, 1955.

[24] I. Gaudiello, E. Zibetti, S. Lefort, M. Chetouani, and S. Ivaldi, "Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to icub answers," *Computers in Human Behavior*, vol. 61, pp. 633–655, 2016.

[25] A. Hamacher, N. Bianchi-Berthouze, A. Pipe, and K. Eder, "Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical human-robot interaction," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, 2016, pp. 493–500.

[26] N. Mirnig, G. Stollnberger, M. Miksch, S. Stadler, M. Giuliani, and M. Tscheligi, "To err is robot: How humans assess and act toward an erroneous social robot," *Frontiers in Robotics and AI*, vol. 4, p. 21, 2017.

[27] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 141–148.