Shaping Robot Gestures to Shape Users' Perception: The Effect of Amplitude and Speed on Godspeed Ratings

Amol Deshmukh, Bart Craenen, Alessandro Vinciarelli and Mary Ellen Foster
University of Glasgow, School of Computing Science,
Glasgow, United Kingdom
{Amol.Deshmukh,Bart.Craenen,Alessandro.Vinciarelli,MaryEllen.Foster}@glasgow.ac.uk

ABSTRACT

This work analyses the relationship between the way robots gesture and the way those gestures are perceived by human users. In particular, this work shows how modifying the amplitude and speed of a gesture affect the Godspeed scores given to those gestures, by means of an experiment involving 45 stimuli and 30 observers. The results suggest that shaping gestures aimed at manifesting the inner state of the robot (e.g., cheering or showing disappointment) tends to change the perception of Animacy (the dimension that accounts for how driven by endogenous factors the robot is perceived to be), while shaping gestures aimed at achieving an interaction effect (e.g., engaging and disengaging) tends to change the perception of Anthropomorphism, Likeability and Perceived Safety (the dimensions that account for the social aspects of the perception).

CCS CONCEPTS

• **Human-centered computing → User models**; • **Computing methodologies → Computational control theory**; • **Computer systems organization → Robotic autonomy**;

KEYWORDS

Synthetic Gestures, Perception, Social Signals

ACM Reference Format:
Amol Deshmukh, Bart Craenen, Alessandro Vinciarelli and Mary Ellen Foster . 2018. Shaping Robot Gestures to Shape Users' Perception: The Effect of Amplitude and Speed on Godspeed Ratings. In *6th International Conference on Human-Agent Interaction (HAI '18), December 15–18, 2018, Southampton, United Kingdom.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3284432.3284445

1 INTRODUCTION

Expressiveness is one of the key abilities of social robots because it enables them to stimulate their inner states, personality and other socially relevant information [1]. One approach for robots to express themselves is using non-verbal behaviours such as gestures. In previous work, the primary reason for focusing on gestures is that "gestural expression is intimately involved in acts of spoken linguistic expression" [2], meaning that speech and gestures are processed as a bimodal unit at the neural [3], cognitive [4] and psychological [5] level. In particular, speech and gestures have been shown to mutually enhance one another to make an agent more effective in achieving communicative goals [6]. For this reason it is necessary to develop approaches capable of selecting gestures appropriate for a given situation and shaping them in the same way as a human would do.

Synthetic gestures must be expressed in a way that people can identify and understand [7]. This work investigates the relationship between the gestures that a humanoid robot displays and the perception of the users, i.e., the tendency of the users to attribute to robots certain characteristics over others. The main difference with respect to most previous work in this area is that the approach proposed in this article does not only take into account the selection of gestures displayed by the robot, but also the way in which those gestures are displayed. In particular, the experiments investigate the association between variations of *amplitude* and *speed* — two major parameters that characterise any natural and synthetic gesture — and variations of the users' perception measured with the Godspeed questionnaire [8].

The experiments presented here aim to investigate the interaction between people and robots in public spaces and, more specifically, in environments in which the level of acoustic noise tends to be high enough to make it difficult to hear and understand speech. In such situations, from what we know from the study of biology [9, 10], multiple modalities do not enhance one another, but rather generate redundancy by carrying the same message. In this way, the failure of one modality (e.g., speech that cannot be heard due to high noise) can be compensated by the other modalities (e.g., gestures can be seen irrespectively of acoustic noise). This is the main reason why the experiments presented in this work take into account isolated gestures that do not accompany or interact with spoken messages.

In previous studies the authors published results on the role of personality as a mediation variable between gestures of different energy and spatial extension of a robot [11], the occurrence of a similarity attraction effect for the majority of the observers involved in the experience [12], and the understandability of the gestures displayed [13]. This study shows that changing the way a gesture is displayed is associated, to a statistically significant extend, with changes in the users' perception of those gestures.

However, the results show that this does not happen in the same way for all gestures or for all the dimensions of the Godspeed questionnaire [8]. In particular, no effects were observed for a gesture like Pointing, which, in general, is expected to exchange spatial knowledge and not to achieve interactional goals or to convey the impression of an inner state [14]. Gestures designed to achieve

HAI '18, December 15–18, 2018, Southampton, United Kingdom

ACM ISBN 978-1-4503-5953-5/18/12…$15.00
https://doi.org/10.1145/3284432.3284445

an interactional goal, such as Engaging and Disengaging, are associated with most dimensions; particularly with Likeability, the most socially oriented dimension of the Godspeed questionnaire. Finally, gestures aimed at conveying the impression of an inner state like Cheering and Head-Touching are associated with changes in the perception of Animacy, the dimension that accounts for the perception of inner processes and motivations in the robot.

The rest of this article is organized as follows: Section 2 surveys previous work in the area, Section 3 describes the experimental approach, Section 4 presents experiments and results, while the final Section 5 draws conclusions.

2 SURVEY OF PREVIOUS WORK

Many of the most popular social robots — such as SoftBank's Nao and Pepper — have few or no moving parts in their faces, and are therefore not equipped to display facial expressions. Also, as mentioned earlier, often the acoustic context of an interaction can make spoken interaction problematic, particularly in noisy public spaces. Thus, the use of gestures, and other bodily displayed cues, plays a critical role in managing social human-robot interaction [15]. Purely emotional body expressions of a social robot–such as raising the hands to show emotions such as joy, anger, or fear–have been successfully used in a range of robot interaction contexts [16–19].

The role of gestures in Human-Robot Interaction has been addressed in various previous work. In most cases, the starting point is the observation that gestures are an essential component of non-verbal communication in Human-Human exchanges [20, 21]. 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 each other [22]. 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 [23] show that people recognise 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 recognise and accept the cooperative gestures of the robot and the ability to recognise human gestures. Similarly, the experiments presented in [24] show that the use of synthetic gestures during robot story-telling 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 [25] 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 [26] show that synthetic gestures can increase the engagement of people involved in an interaction with robots, while the approach proposed in [27] shows that humans can interpret synthetic gestures in terms of emotions.

Regarding deictic gestures, the approach proposed in [28] 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 [29], 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.

A number of previous studies have examined how various parameters can influence the users' reactions to the non-verbal behaviour of a virtually or physically embodied conversational agent. Salem et al. found that a robot is evaluated more positively when non-verbal behaviours, such as hand and arm gestures, are displayed along with speech, even if they do not semantically match the spoken utterance [30]. The model proposed by Amaya et al. [31], for example, transforms neutral animations into emotional animations by using "emotional transforms" which affect the speed and spatial amplitude of the animation. Yamaguchi et al. [32] defined a set of rules for modifying basic motions of a virtual character to express basic emotions, such as joy and sadness, and found that amplitude, position, and speed were the main parameters. The approach described by Kim et al. [18] explored how controlling the size, velocity, and frequency of robot gestures could affect user perception of the robot's personality. It was found that all of these factors had an effect on the perceived robot personality, and that this factor in turn affected users' subjective impressions of the robot.

The model developed by Pelachaud [33] for gesture expressiveness adopts six parameters, including spatial extent, temporal extent, fluidity, power, overall activation, and repetition. In perceptual tests, the six parameters were found to be recognizable and also combine to produce movements with different qualities. The work by Xu et al. [34] proposes a parametrized behaviour model with specific behaviour parameters for bodily mood expression, and applied the model to two concrete behaviours — waving and pointing — of the Nao robot. The most important parameters for creating readable mood expressions were found to be hand height and amplitude, head position, and motion speed [35]. The experiments described in [36] found that various levels of exaggeration in motion of a humanoid robot correlate to human expectations of robot-like, human-like, and cartoon-like motion. Use of exaggerated motion enhanced the interaction through increased levels of engagement and perceived entertainment value.

In a work that is particularly relevant to the current study [37], the authors have recently updated their robot-independent model for upper-body gestures of a social robot [38, 39] to add the ability to modulate functional gestures, such as pointing, to incorporate affective content. In their system, the speed and amplitude of a functional gesture are modified with the goal of projecting a particular affective impression, as expressed by valence and arousal. The choice of those two specific parameters and the definition of their relationship to valence and arousal were based on findings from the literature mentioned above [32, 40]; however, the resulting gestures have not yet been evaluated to determine whether the target affective state was successfully projected.

When it comes to the systematic analysis of robot gestures, Table 1 shows the classification into five categories proposed in [41]. In the context of socially intelligent robots in public spaces, as studied here, it is anticipated that the robot may exhibit all five of these

Class	Name	Characteristics
1	Irrelevant/ Manipulative Gestures	- Manipulation of objects, side effects of motor behaviour, body motion - Neither communicative nor socially interactive
2	Side Effect of Expressive Behaviour	- Associated to communication or affective states of human e.g. persons talk excitedly raising and moving their hands in correlation with changes in voice prosody or emphasis of speech.
3	Symbolic Gestures	- Communicative of semantic content, e.g. waving down; use of a conventional hand signals; nodding 'yes'; waving a greeting 'hello' or 'goodbye'
4	Interactional Gestures	- Used to initiate, maintain, regulate, synchronise, organise or terminate various types of interaction e.g. raising the hand toward the partner inviting them or send them away
5	Referential/ Pointing Gestures	- Pointing to all types of effectors: referential, attention-directing e.g. presenting objects, persons, directions or locations by pointing

Table 1: Nehaniv's classification of gestures [41]

gesture classes, with gestures in classes 3–5 particularly relevant to this study, where the goal is to modify the gestures to influence user perception of the robot. While the previous studies listed above considered a range of gesture parameters, all included speed and amplitude parameters in some form. This is not surprising, as these are two dimensions that have been shown to be crucial for controlling gestures for artificial agents [42] — and they are, indeed, the two dimensions that are considered in the this study.

3 EXPERIMENTAL APPROACH

This work is being carried out in context of the MultiModal Mall Entertainment Robot (MuMMER) project, a four-year, EU-funded project[1], with the overall goal of developing a humanoid robot, Pepper, that can interact autonomously and naturally in the dynamic environments of a public shopping mall [43]. The overall concept underlying MuMMER is that for a robot to be successful in such a situation, it must be *entertaining* and *engaging*: that is, it must possess the social intelligence to both understand the needs and interactive behaviour of the users, as well as to produce appropriate behaviour in response. When the robot is able to support such smooth interactions, this should provide a sufficiently engaging experience that will stand up to repeated visits in a long-term deployment context. The goal of this work is to investigate how user perception changes depending on the gestures that a robot displays.

[1]www.mummer-project.eu/

During the experiments, 30 independent human observers were asked to watch 45 different gestures displayed by *Pepper* — a robotic platform manufactured by Softbank Robotics — and to complete, for each gesture, the Godspeed questionnaire [8], a well-known and validated instrument for measuring user impressions of interactive robots. The gestures used in this study are based on 5 animations selected in the standard library available with the robot. The 9 variants of each core gesture have been obtained by manipulating two parameters: *speed*, and *amplitude*. Using this experimental configuration it is possible to investigate whether there is an association between the way the robot displays the gesture and the perception of the users. The motivation behind the choice of speed and amplitude is that they are related to energy and spatial extension, respectively; two characteristics that have been shown to play a crucial role in the expressiveness of artificial agents [42].

The following sections describe the way the gestures adopted in the experiments have been generated (see Section 3.1) and the approach adopted to investigate the relationship between users' perception and gestures (see Section 3.2).

3.1 The stimuli

This section describes the process used to synthesize the 45 gestures — the *stimuli* hereafter — used in the experiments of this work. The first step is the selection of 5 standard gestures — the *core stimuli* hereafter — available in the library accompanying the *Pepper* robot. The joints of the Pepper robot have 17 degrees of freedom (DOF) in total (see Figure 1).

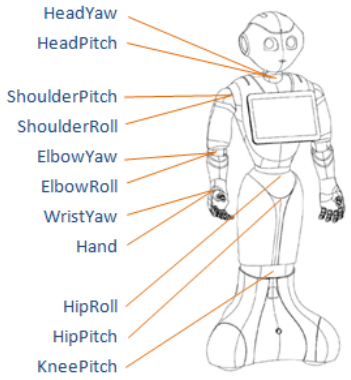

Figure 1: Pepper robot joints, 17 DOF

The selection of the 5 standard gestures targeted gestures that, according to the criteria underlying the taxonomy proposed in [41] (refer Table 1), are relevant to the context addressed in this work, i.e., the interaction between people and robots in public spaces. In this context, the gestures selected include gestures for: attracting attention when the users are not engaged; disengaging when the interaction requires termination or there is overcrowding near the robot; pointing, to give directions; and signalling failure or success in performing a task or interacting with the human. The names that the robot's manufacturer has given to the selected gestures are as follows (see Figure 2) [2]:

[2]The animations associated to the core stimuli are available on the version 1.6B of Pepper in the following directories:

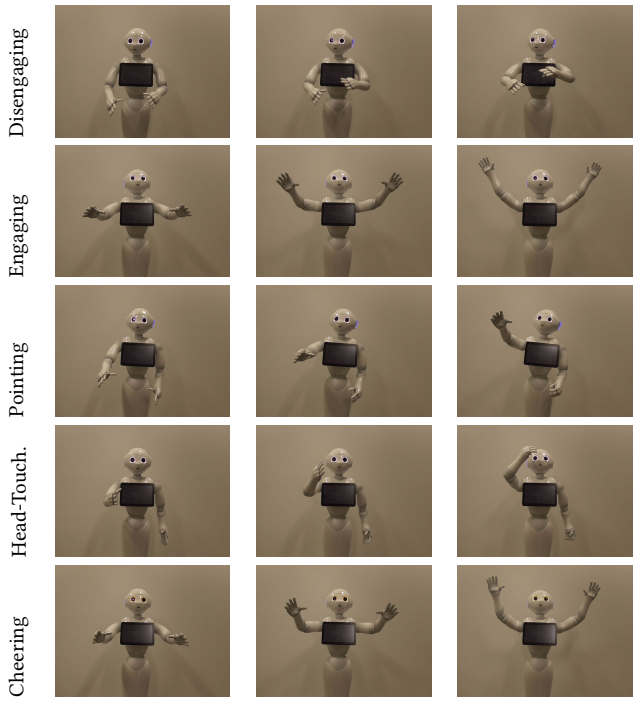

Table 2: The figures show, for each of the five core stimuli, the effect of the parameter α. The rightmost column (α = 1.00) contains the core stimuli.

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

The second step of the process is the synthesis of 9 variants for each of the core stimuli above. Three variants were generated by adopting three different values of the speed λ per core stimulus: 15, 25 and 35 *frames per second* (*fps*), where 25 *fps* is the original speed of the core stimuli. For each of the 15 resulting gestures, another three stimuli can be obtained by modifying the differences $\Delta_i(t) = \theta_i(t) - \theta_i(t-1)$, where $\theta_i(t)$ is the angle between the two mechanical elements connected by joint i at frame t. In particular, the values of the $\Delta_i(t)$ were multiplied, for all values of i and t, by a factor α — the *amplitude* hereafter. Three different values of α were adopted, namely 0.50, 0.75 and 1.00. In the first two cases, the result is a dampened version of the core stimulus, in the last case, the $\Delta_i(t)$ are left unchanged. As a result of the process above, the 9 variants of a given core stimulus correspond to 9 pairs (α, λ), and the pair where $\lambda = 25$ and $\alpha = 1.00$ is the core stimulus itself. The versions of the core stimuli corresponding to the different values of α are portrayed in Table 2.

"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)

3.2 Perception Effects Analysis

The question addressed in this work is whether users perceive robots differently when they display different gestures and, if so, how the perception of the users changes in relation to the characteristics of the gestures. During the experiments, the 30 observers involved in the experiments have watched the 45 stimuli (independent variable for the study) and, for each of them, watched and rated all stimuli using the Godspeed questionnaire [8] (dependent variable for the study). The Godspeed questionnaire is widely accepted as a standard measurement tool for Human Robot Interaction and aims at quantifying the following tendencies underlying users' perception:

- *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.

Completing the questionnaire results in five scores that measure the tendencies above: the higher the score, the more pronounced the tendency (see [8] for full details). For a given stimulus, collating the Godspeed scores leads to a matrix $S = \{s_{ik}\}$, where s_{ik} is the score of observer i (where $i = 1, \ldots, N$) for tendency k (where $k = 1, \ldots, 5$). Thus, the following sum:

$$c_j = \sum_{i=1}^{N} s_{ij} \tag{1}$$

can be interpreted as the total number of points that the observers have accumulated for tendency j. Correspondingly, for tendency j, the total number of points accumulated over all variants of the same core gesture can be calculated as follows:

$$T_j = \sum_{\alpha} \sum_{\lambda} c_j^{\alpha\lambda} \tag{2}$$

where the sums extend over all values of parameters α and λ (see Section 3.1) and $c_j^{\alpha\lambda}$ is the value of c_j obtained for a particular pair (α,λ), i.e., a particular variant of the core stimulus under examination.

The expressions above allow one to define the following χ^2 variable [44]:

$$\chi^2 = \sum_{\alpha} \sum_{\lambda} \frac{(c_j^{\alpha\lambda} - E)^2}{E}, \tag{3}$$

where $E = \frac{1}{9} T_j$. In other words, $c_j^{\alpha\lambda}$ plays the role of the observed number of points for a given variant (α, λ), while the value E plays the role of the expectation that, in this case, corresponds to a uniform distribution of points across the different variants.

The χ^2 variable is then a single value that allows one to test whether the observed distribution of the points over all values of α and λ deviates from the uniform distribution to a statistically significant extent. When this is the case, it is possible to say that the Godspeed tendency associated to column j in S is more or less

pronounced depending on the particular gesture being displayed. More information on the use of χ^2 variables can be found in [44].

Because the χ^2 variable described above will be compared amongst several Godspeed tendencies, the analysis relies on multiple comparisons. This introduces the multiple comparison problem. To tackle this problem, *False Discovery Rate* (FDR) [45] correction will be applied when several Godspeed tendencies will be compared. FDR correction is a method of conceptualizing the rate of Type I errors in null hypothesis testing when conducting multiple and/or repeated comparisons. FDR-controlling procedures are designed to control the expected proportion of discoveries (rejected null hypotheses) that are false (incorrect rejections). FDR correction was chosen because it provides less stringent control of Type I errors compared to Familywise Error Rate (FWER) controlling procedures (such as the Bonferroni correction), which control the probability of at least one Type I error. As such, FDR-controlling procedures have greater power than FWER-controlling procedures and applying FDR correction ensures that the number of false positives, if any, will be sufficiently low not to change the outcomes of the analysis.

4 EXPERIMENTS AND RESULTS

The experiments of this work involved $N = 30$ observers, asked to watch the 45 stimuli described in Section 3.1 and, for each of them, to fill out the Godspeed questionnaire (see Section 3.2). All observers have performed the tasks above for all stimuli. The stimuli were presented to the observers in random order over three separate sessions (15 stimuli per session), with two stimuli derived from the same base stimulus never presented consecutively. This eliminated any ordering issues, and no ordering effects were later observed from the collected data.

The 30 observers were split into 10 groups of 3 people each, who were asked to participate in the same sessions, while still working independently of each other. The sessions were held over three consecutive days to limit possible tiredness effects due to the repetition of the tasks over extended periods of time. The stimuli were labelled by unique (ID) numbers, and the participants were exposed to those numbers when they were asked to fill out the questionnaire. However, because the ordering of the presented stimuli was randomised, this would not have given participants 'context' about the stimuli, i.e., they would not have been able to pre-classify a stimulus by its number.

Figure 2 shows the experimental setting: the observers filled out the questionnaires while sitting in front of the robot at a distance of roughly 1.5 meters. The questionnaires were filled out using a software interface running on a tablet. The 30 observers were selected randomly from a pool of subjects available at the research institute where the experiments were conducted. In terms of demographics, 20 observers were female and 10 were male; with the age distribution as depicted in Table 3. The participants were of varying ethnic and national origin. Only 3 observers had interacted with a robot before participating in the experiments of this work. The participants received a payment corresponding to the minimum legal hourly wage in the country where the experiments were conducted. The rest of this section presents the results of the analysis performed on the experimental results according to the approach presented in Section 3.

Figure 2: Experimental Setting. The observers sit at a distance of roughly 1.5 meters from the robot and fill out the questionnaires using a tablet.

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

Table 3: Age distributions of the observers involved in the experiments.

4.1 Consistency and Reliability

As a prelude to any analysis of Godspeed questionnaire scores it is common practice to evaluate the internal consistency and the effective reliability of the scores. As advised in [8] Cronbach's alpha ([46]) was used to estimate the internal consistency of the observer's responses. Cronbach's alpha was computed as follows:

$$\alpha_s = \frac{K}{K-1} \cdot \left(1 - \frac{\sum_{i=1}^{K} \sigma^2_{Y_i}}{\sigma^2_X}\right) \tag{4}$$

where K is the number of test-items, $\sigma^2_{Y_i}$ the variance of the scores for item i, and σ^2_X the variance of the observed total test scores. For calculating Cronbach's alpha the number of test-items, K, is equal to the total number of questions in the Godspeed questionnaire, or the number of questions for each of the tendencies assessed by the questionnaire. The Godspeed questionnaire contains 23 questions in total; 5 questions each for the Anthropomorphism, Animacy, Likeability, and Perceived Intelligence tendencies, and 3 questions for the Perceived Safety tendency. The variances $\sigma^2_{Y_i}$ and σ^2_X were calculated over the scores given by the observers per questions, and over the sum of all scores in the questionnaire, either for the total questionnaire, or per tendency. The number of scores per questions equals 1350, the number of stimuli (45) multiplied by the number of observers (30).

For computing the effective reliability of the scores Spearman-Brown's prediction formula ([47]) was used. Spearman-Brown's prediction formula, when used for calculating the effective reliability of a test is also called the "standardized Cronbach's alpha", as it is the same as Cronbach's alpha computed using the average item intercorrelation and unit-item variance, rather than the average item covariance and average item variance. Cronbach's alpha is thus related conceptually to the Spearman-Brown prediction formula, as both arise from the basic classical test theory result that the reliability of test scores (ρ_{XX}) can be expressed as the ration of

the true-score (σ_T^2) and the total-score (error plus true score, σ_X^2) variances: $\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2}$. Spearman-Brown's prediction formula for effective reliability was computed as follows:

$$R_{sb} = \frac{n \cdot \overline{r}}{(1 + (n-1) \cdot \overline{r})} \quad (5)$$

where n is the number of observers (30), and $\overline{r}$ is the mean of the $n(n-1)/2$ non-redundant correlation coefficients between the scores of all observers (i.e., the mean of the upper or lower triangular of the correlation matrix). Pearson's correlation coefficient was used to pair-wise calculate the correlation coefficients between the scores produced by all observers.

For both Cronbach's alpha as well as Spearman-Brown's effective reliability holds in general that they are seemed sufficient between 0.7 and 0.8, good between 0.8 and 0.9, and excellent when 0.9 or larger.

Tendency	α_s	R_{sb}
Anthropomorphism	0.904	0.699
Animacy	0.892	0.811
Likeability	0.942	0.862
Perceived Intelligence	0.915	**0.206**
Perceived Safety	**0.242**	0.908
Total	0.934	0.814

Table 4: Cronbach's alpha and Spearman-Brown's reliability for all tendencies, and for the entire Godspeed questionnaire (Total).

Table 4 shows both Cronbach's alpha, as well as Spearman-Brown's effective reliability values for all tendencies assessed by the Godspeed questionnaire, as well as for the entire questionnaire (Total). The values in Table 4 show that for almost all tendencies, as well as the entire Godspeed questionnaire, the internal consistency and the effective reliability of the scores given by the observers is either good or excellent. Only for the Perceived Intelligence and Perceived Safety tendencies is, respectively, the effective reliability and the internal consistency insufficient. The repercussions of this will be discussed in during the analysis of the results described in the next section.

4.2 Gestures and Perception

Table 5 shows the instances in which the distribution of the Godspeed scores across the multiple variants of the same core stimulus deviates, to a statistically significant extent, from the uniform distribution (see Section 3.2 for details about the data analysis approach). When the deviation is statistically significant, the table also shows whether increasing amplitude and speed of a gesture corresponds to higher or lower Godspeed scores. A deviation from the uniform distribution is considered statistically significant when a χ^2 test results in a p-value lower than 0.05. The *False Discovery Rate* (FDR) correction [45] has been applied to all p-values to tackle the repeated comparisons problem. This ensures that the number of false positives, if any, is sufficiently low not to change the conclusions of the experiment.

For the Disengaging gesture, the effects take place in correspondence with Likeability and Perceived Safety. In the case of Likeability, the scores tend to decrease when α and λ increase, while in the case of Perceived Safety the scores tend to increase when α and λ decrease.

	Ant		Ani		Lik		Int		Saf	
Core Stimulus	α	λ	α	λ	α	λ	α	λ	α	λ
Engaging	↑	↑	↑	↑	↑	↑				
Disengaging					↓	↓			↓	↓
Pointing										
Head-Touching			↑	↑						
Cheering			↑	↑						

Table 5: The symbols "↑" and "↓" account for statistically significant effects. The symbol "↑" means that increasing amplitude or speed corresponds to observing higher Godspeed scores. The symbol "↓" means that decreasing amplitude or speed corresponds to observing lower Godspeed scores. Empty cells correspond to cases in which no statistically significant effects were observed.

A possible explanation behind the Likeability effects is that the gesture aims at increasing the physical distance between the robot and its users. Given that physical and social distances were shown to be equivalent (the longer the physical distance, the longer the social distance) [48], increasing the energy of the gesture may look like an attempt by the robot to push people towards distances that, according to proxemic theories [49], correspond to less friendly and more formal relationships.

As for the Perceived Safety effects, any conclusions based on the results of the analysis should be considered carefully, as in the previous section we found that the internal consistency among the scores was found to be insufficient (see Table 4). This indicates that the observers did not agree among themselves on how to score the gestures with regards to Perceived Safety. One possible explanation for this is that the observers may have had different preconceived notions about the Perceived Safety of the robots in general when scoring the gestures. Some may have found robots inherently unsafe, while others may be more trusting towards robots. These differing attitudes would have an effect on scoring the gestures that can not be easily quantified from the Godspeed questionnaire on its own. In any case, the observed effects are still statistically significant. A probable explanation for the effect is that slower movements (lower λ) that do not extend far from the robot's body (lower α) are less likely to harm the users, and may thus be perceived as safer.

In the case of the Engaging gesture, statistically significant effects have been observed for Anthropomorphism, Animacy and Likeability. In all three cases, increasing amplitude and speed corresponds to higher Godspeed scores.

For Anthropomorphism, one possible explanation is that the human brain has been shown to be more anthropomorphic — meaning that it is more prone to process artificial agents like it processes human ones — when synthetic movements are more similar to those displayed by humans [50]. Lowering α and λ produces gestures that, at least in the case of the Engage core stimulus, are less similar to those displayed by humans.

A possible explanation for the Animacy effects is that higher speed and amplitude results in more energetic gestures and higher motor activation; two factors that play a crucial role when observers consider an agent as alive [8].

The increase of the Likeability scores is likely dependent on the correlation between Anthropomorphism and positive judgements about the robots observed earlier in the literature [51].

Overall, the three effects observed for the Engaging gesture are an advantage in those scenarios in which the robot is expected to proactively start the interaction with the users. The reason is that the effects provide indications on how to make the perception of the users more positive — a prerequisite towards successful interactions with machines that display human-like behaviour (see, e.g., [52]) — at the very moment they enter in contact with the users.

No statistically significant effect was found for the Pointing gesture. A possible explanation for this is that deictic gestures are meant to convey information about spatial knowledge [14] — in particular when it comes to the position of an object of interest in the environment — and not about the social and psychological phenomena underlying the items of the Godspeed questionnaire [8].

Equally, no statistically significant effect was found for the Perceived Intelligence tendency for any gesture included in the study. This may be because the effective reliability of the Godspeed scores was found to be insufficient (see section 4.1); in this case, extending the experiment to include more observers may improve the effective reliability of the scores, which, in turn, may result in a statistically significant effect to be found for Perceived Intelligence. However, the number of observers included in the study did result in sufficient effective reliability of the scores for all the other tendencies. So that it may simply be that gestures of the kind investigated in this study, on their own, without context or other modalities of interaction, simply do not convey any human perceivable information about the intelligence of a robot. To the best of the knowledge of the authors, no prior literature has ever established a link between gestures of this kind, as displayed by a robot in this context, and the perceived intelligence of a robot.

Finally, both the Head-Touching and Cheering gestures show significant effects for Animacy. The main probable reason is that both gestures, when displayed by people, tend to convey information about one's inner state. Head-Touching, in particular, is typically associated with a situation of confusion [20, 21], while Cheering tends to be displayed as a sign of success and satisfaction [6]. This means that a robot displaying these two gestures can elicit the attribution of the same inner states and, ultimately, of Animacy, defined as the very property of being alive [8].

For both Head-Touching and Cheering, the Animacy scores tend to increase when both α and λ increase. Regarding α, the probable reason is that lowering the parameter leads to gestures that have a morphology different from the core stimulus and, hence, fail in conveying the same impression. In the case of λ, the probable reason is that movements have been shown to play a crucial role in the attribution of Animacy, the very difference between animate beings and inanimate objects [8]. Thus, increasing the movement's energy (proportional to speed) tends to attract higher Animacy scores.

5 CONCLUSIONS

This study presented experiments on the relationship between the way a gesture is displayed by a robot and the perception of its users. The results show that, at least in some cases, there is an association between speed and amplitude of a gesture — two parameters that account for energy and spatial extension — and Godspeed scores [8]. Overall, the coherent picture that emerges is that gestures expected to achieve a social goal — Engaging and Disengaging — show effects primarily on the Godspeed dimensions that better account for social aspects of Human Robot Interaction, namely Anthropomorphism (the tendency to attribute human characteristics to the robot) and Likeability (the tendency to attribute desirable characteristics to the robot). Similarly, gestures designed to simulate an 'inner state' — Head-Touching and Cheering — show effects in the area of Animacy, the Godspeed dimension that captures the tendency to consider the robot alive and, hence, capable of experiencing the world. Finally, no effects were found for Pointing, a gesture that, unlike the other stimuli used in the experiments, aims more at sharing knowledge about the environment than at conveying information about the dimensions underlying the Godspeed questionnaire.

The above suggests that the stimuli have been designed correctly and, most importantly, it shows that the Godspeed scores tend to be differ when different values of amplitude and speed are used. The main implication of this observation is that it is not sufficient to just decide which gestures a robot should display during an interaction; but that a decision about how those gestures are displayed is required. In particular, the same gesture should be displayed with different amplitude and speed depending on how much the tendencies underlying the Godspeed scores should be expressed. Whether the robot will still be perceived as a consistent single social agent when displaying gestures with different amplitudes and/or speeds is an avenue for future study.

The experiments described in this study involved displaying isolated gestures without support of other modalities or context, such as spoken messages. In addition, great care was taken to avoid any ordering effects, primarily by randomising the order in which the gestures were displayed to the observer. No ordering effects were observed in the collected data. The experiments were designed in this way because the scenarios in which the gestures will be used involve public spaces where the likely level of acoustic noise will be high. In this context, the gestures will be expected to compensate for difficulties in hearing and understanding spoken messages in line with biological studies about the use of multiple modalities in noisy environments [9, 10]. Such noisy condition as described 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, airport, stations, and other public spaces. In these contexts robots should display gestures as understandable as possible, because they will be competing with other stimuli designed to attract and retain attention (e.g., advertisements, danger warnings, public announcements, etc.). However, future work will aim at investigating how the findings of this work may change when the gestures are accompanied by speech, the most frequent case in everyday human-human, and human-robot interactions [2, 6].

ACKNOWLEDGMENTS

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] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166, 2003.

[2] A. Kendon. Language and gesture: unity or duality? In D. McNeill, editor, *Language and Gesture*, pages 47–63. Cambridge University Press, 2000.

[3] S.D. Kelly, C. Kravitz, and M. Hopkins. Neural correlates of bimodal speech and gesture comprehension. *Brain and language*, 89(1):253–260, 2004.

[4] J.P. de Ruiter. The production of gesture and speech. In D. McNeill, editor, *Language and Gesture*, pages 284–311. Cambridge University Press, 2000.

[5] S.D.D Kelly, A. Özyürek, and E. Maris. Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science*, 21(2):260–267, 2010.

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

[7] C.L. Breazeal. *Designing sociable robots.* MIT press, 2004.

[8] 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*, 1(1):71–81, 2009.

[9] S.R. Partan and P. Marler. Issues in the classification of multimodal communication signals. *The American Naturalist*, 166(2):231–245, 2005.

[10] S.R. Partan and P. Marler. Communication goes multimodal. *Science*, 283(5406):1272–1273, 1999.

[11] B.G.W. Craenen, A. Deshmukh, M.E. Foster, and A. Vinciarelli. Do we really like robots that match our personality? The case of Big-Five traits, Godspeed scores and robotic gestures. In *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Nanjing, China, 2018.

[12] B.G.W. Craenen, A. Deshmukh, M.E. Foster, and A. Vinciarelli. Shaping gestures to shape personalities: The relationship between gesture parameters, attributed personality traits and Godspeed scores. In *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Nanjing, China, 2018.

[13] A. Deshmukh, B.G.W. Craenen, M.E. Foster, and A. Vinciarelli. The more I understand it, the less I like it: The relationship between understandability and Godspeed scores for robotic gestures. In *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Nanjing, China, 2018.

[14] J. Haviland. Pointing, gesture spaces, and mental maps. In D. McNeill, editor, *Language and Gesture*, pages 13–46. Cambridge University Press, 2000.

[15] C. Breazeal, C.D. Kidd, A.L. Thomaz, G. Hoffman, and M. Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 708–713. IEEE, 2005.

[16] M. Zecca, Y. Mizoguchi, K. Endo, F. Iida, Y. Kawabata, N. Endo, K. Itoh, and A. Takanishi. Whole body emotion expressions for kobian humanoid robot-preliminary experiments with different emotional patterns. In *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, pages 381–386. IEEE, 2009.

[17] M. Häring, N. Bee, and E. André. Creation and evaluation of emotion expression with body movement, sound and eye color for humanoid robots. In *Ro-Man, 2011 Ieee*, pages 204–209. IEEE, 2011.

[18] H. Kim, S.S. Kwak, and M. Kim. Personality design of sociable robots by control of gesture design factors. In *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, pages 494–499. IEEE, 2008.

[19] Myrthe Tielman, Mark Neerincx, John-Jules Meyer, and Rosemarijn Looije. Adaptive emotional expression in robot-child interaction. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 407–414. ACM, 2014.

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

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

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

[23] L.D. Riek, T.-C. Rabinowitch, P. Bremner, A.G. 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*, pages 61–68, 2010.

[24] C.-M. Huang and B. Mutlu. Modeling and evaluating narrative gestures for humanlike robots. In *Robotics: Science and Systems*, pages 57–64, 2013.

[25] 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*, volume 2, pages 558–563, 2003.

[26] C.L. Sidner, C. Lee, C.D. Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164, 2005.

[27] 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.

[28] 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*, pages 2423–2428, 2005.

[29] A.G. 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*, pages 297–304, 2006.

[30] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4(2):201–217, 2012.

[31] K. Amaya, A. Bruderlin, and T. Calvert. Emotion from motion. In *Graphics interface*, volume 96, pages 222–229, 1996.

[32] A. Yamaguchi, Y. Yano, S. Doki, and S. Okuma. A study of emotional motion description by motion modification and adjectival expressions. In *Cybernetics and Intelligent Systems, 2006 IEEE Conference on*, pages 1–6. IEEE, 2006.

[33] C. Pelachaud. Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7):630–639, 2009.

[34] Junchao Xu, Joost Broekens, Koen Hindriks, and Mark A. Neerincx. Bodily mood expression: Recognize moods from functional behaviors of humanoid robots. In *Proceedings of the 5th International Conference on Social Robotics - Volume 8239*, ICSR 2013, pages 511–520, Berlin, Heidelberg, 2013. Springer-Verlag.

[35] J. Xu, J. Broekens, K. Hindriks, and M.A. Neerincx. The relative importance and interrelations between behavior parameters for robots' mood expression. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 558–563. IEEE, 2013.

[36] M. Gielniak and A. Thomaz. Enhancing interaction through exaggerated motion synthesis. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 375–382. ACM, 2012.

[37] Greet Van de Perre, Hoang-Long Cao, Albert De Beir, Pablo Gómez Esteban, Dirk Lefeber, and Bram Vanderborght. Generic method for generating blended gestures and affective functional behaviors for social robots. *Autonomous Robots*, 42(3):569–580, Mar 2018.

[38] G. van de Perre, M. van Damme, D. Lefeber, and B. Vanderborght. Development of a generic method to generate upper-body emotional expressions for different social robots. *Advanced Robotics*, 29(9):597–609, 2015.

[39] G. Van de Perre, A. De Beir, H.-L. Cao, P.G. Esteban, D. Lefeber, and B. Vanderborght. Reaching and pointing gestures calculated by a generic gesture system for social robots. *Robot. Auton. Syst.*, 83(C):32–43, September 2016.

[40] Y.-H. Lin, C.-Y. Liu, H.-W. Lee, S.-L. Huang, and T.-Y. Li. Evaluating emotive character animations created with procedural animation. In *Intelligent Virtual Agents*, pages 308–315. Springer, 2009.

[41] C.L. Nehaniv, K. Dautenhahn, J. Kubacki, M. Haegele, C. Parlitz, and R. Alami. A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction. In *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, pages 371–377. IEEE, 2005.

[42] B. Hartmann, M. Mancini, and C. Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In *Proceedings of International Gesture Workshop*, pages 188–199, 2005.

[43] M.E. Foster, R. Alami, O. Gestranius, O. Lemon, M. Niemelä, J.-M. Odobez, and A.K. Pandey. The MuMMER project: Engaging human-robot interaction in real-world public spaces. In *Proceedings of the Eighth International Conference on Social Robotics (ICSR 2016)*, November 2016.

[44] D. Howell. *Statistical methods for psychology.* Cengage Learning, 2012.

[45] 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*, pages 289–300, 1995.

[46] L.J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 13(3):297–334, 1951.

[47] J. Stanley. Reliability. In R. L. Thorndike, editor, *Educational Measurement. Second edition.* American Council on Education, 1971.

[48] A. Kendon. *Conducting Interaction.* Cambridge University Press, 1990.

[49] E. Hall. *The silent language.* Doubleday, 1959.

[50] V. Gazzola, G. Rizzolatti, B. Wicker, and C. Keysers. The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage*, 35(4):1674–1684, 2007.

[51] 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*, 5(3):313–323, 2013.

[52] C. Nass and S. Brave. *Wired for speech: How voice activates and advances the human-computer relationship.* MIT Press, 2005.