2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 23-27, 2022, Kyoto, Japan

Controlling the Impression of Robots via GAN-based Gesture Generation

Bowen Wu[1,2], Jiaqi Shi[1,2], Chaoran Liu[1,3], Carlos T. Ishi[1,3], and Hiroshi Ishiguro[2,3]

***Abstract*— As a type of body language, gestures can largely affect the impressions of human-like robots perceived by users. Recent data-driven approaches to the generation of co-speech gestures have successfully promoted the naturalness of produced gestures. These approaches also possess greater generalizability to work under various contexts than rule-based methods. However, most have no direct control over the human impressions of robots. The main obstacle is that creating a dataset that covers various impression labels is not trivial. In this study, based on previous findings in cognitive science on robot impressions, we present a heuristic method to control them without manual labeling, and demonstrate its effectiveness on a virtual agent and partially on a humanoid robot through subjective experiments with 50 participants.**

2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-7927-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IROS47612.2022.9981535

I. INTRODUCTION

Humanoid robots are useful to human society, such as television announcers, salesclerk, or senior companion. These situations require multimodal interactions between robots and humans. During such interactions, users form impressions of the robots. Such impressions may be strongly connected with the acceptance of robots by an individual or society [1]. To improve the acceptance of robots by users, it is imperative to develop methods that can effectively control the impressions of them.

According to studies in cognitive science, various aspects of robots can affect their impressions, such as appearance, verbal and non-verbal behavior [2]–[7]. Among non-verbal behaviors, gestures are a crucial factor that affects the impressions of robots. Various studies have verified that changing certain gestures results in different impressions [8], [9]. Although the foundation for controlling the impressions of robots can be established, these studies cannot be directly embedded into a robot's control system, since the goal is not the automatic generation of gestures.

Using the open-source tools for recording human motions, co-speech gesture datasets were created [10]–[14], which are used to train data-driven gesture generation models. These models have been evaluated mainly on virtual agents through subjective experiments where evaluators assign scores or rank gestures for different systems [14]–[21]. Compared to rule-based methods where gestures are manually designed for specific scenarios, data-driven models can be generalized to speech inputs that are not in the training data, to provide a powerful component for achieving a multimodal scheme for social robots. Nonetheless, these studies did not consider the control of the impressions of robots. While these models are data-driven, there is no available dataset that includes impressions as labels for training them; the amount of necessary data is too considerable to create such a dataset. Taking the big-five personality as an example, if we set three discrete categories for each personality, then to cover every combination of personalities, at least 3^5 individual's data are necessary. Moreover, data collection of more than one person for each combination is needed to discover commonality. Therefore, modeling the impressions purely by collecting more data is a daunting task.

In order to solve the above stated issues, i.e., the manually-based approaches and the infeasible collection of labeled data for controlling the impression of robots, we propose a pseudo labeling method based on the above existing cognitive science approaches, which automatically divides gesture data into several categories. Gestures belonging to different categories can evoke different impressions in users. The pseudo labeled data were used to train a Generative Adversarial Networks (GANs) -based model to provide a data-driven model with the ability to generate gestures for different impressions. We considered extroversion and excitement the objective impressions since they have been elaborately studied in cognitive science. We summarized gesture traits that affect extroversion and excitement in related studies, e.g., the speed or the amplitude of motion. Based on these traits, the gestures in the dataset are categorized into several classes as pseudo labels, which are then used as additional input in our proposed model. To verify our model's effectiveness, we conducted subjective experiments and showed that it generated gestures that affect perceived extroversion and excitement of a virtual agent and a humanoid robot. Since the social robot is multimodal, i.e., speech and gestures, we also observed that gestures changed the perceived impression of speech and vice-versa.

The rest of this paper is organized as follows. In Section II, we present studies related to the present study. Section III provides an explanation of our proposed method. In Section IV, we describe our experiment that evaluated our proposed model and discuss the results in Section V. We made our implementation public available at `https://github.com/wubowen416/Controlling-the-Impression-of-Robots-via-GAN-based-Gesture-Generation`.

*This work was supported in part by JST, Moonshot RD under Grant JPMJMS2011 (methodology conceptualization), and Grant-in-Aid for Scientific Research on Innovative Areas JP20H05576 (evaluation experiments).

[1]Guardian Robot Project, RIKEN, Japan, `carlos.ishi@riken.jp`

[2]Department of Engineering Science, Osaka University, Japan, {`wu.bowen, shi.jiaqi, ishiguro`}`@irl.sys.es.osaka-u.ac.jp`

[3]Hiroshi Ishiguro Labs, ATR, Japan, `chaoran.liu@atr.jp`

II. RELATED WORK

A. Deep-learning-based co-speech gesture generation

Deep-learning-based models have achieved superior results on gesture generation tasks. Hasegawa et al. [22] found that Mel-Frequency Cepstral Coefficient (MFCC) features of the audio are effective to model the relationship between speech and gestures when using long short term memory (LSTM). Kucherenko et al. [20], [23] argues that MFCC is superior among other type of audio features in terms of the precision of the generated gestures. Yoon et al. [13], [17] proposed models to exploit speech text. Other than deterministic generation, probabilistic generation has been realized by generative models. A combination of Generative Flow (GF) and Variational Auto-Encoder (VAE) model was proposed by Taylor et al. [18] for modeling conditional gesture distribution. GANs are also used to model the conditional gestures distribution when conditioned on speech signals [21], [24]. Although Alexanderson et al. [16] proposed to explicitly control the style of generated gestures using GF-based model, i.e., the speed, the radius and the height of hand positions, they did not consider the possible impact on the impression of robots elicited by this control. While GANs and GF are both probabilistic models, GF usually have a larger number of parameters compared with the generator of GANs due to its weak non-linearity in each layer, which reduces the inference speed, making it impractical to be used in real-life applications. In this study, we trained a conditional GAN-based model to directly control the speed and amplitude of generated gestures and analyzed its impact in terms of the impressions of robots.

B. Human impressions of robots

Personality is a critical facilitator of interpersonal relationships, and human-robot interactions [25]–[27]. The traits of extroversion and introversion are impressive aspect and one of the most popular measures of human personality [28]. Kim et al. [8] designed different personality types for a robot by changing the size, speed, and frequency of gestures and found that a robot with an extroverted personality made a more enjoyable and capable impression on users than introverted robots. Neff et al. [29] described a method of controlling the perceived extroversion of a virtual agent by varying language generation, gesture rate, and movement performance parameters. McRorie et al. [30] created four different characters with various facial and gesture parameters, and evaluated the perception of the agents' personalities, including extroversion-introversion, neuroticism-emotional stability, and psychoticism. Mileounis et al. [31] used a humanoid robot called NAO that is teleoperated through a computer to investigate the effect of extroversion-introversion on social intelligence and revealed that extroversion creates the impression of greater socially intelligence. Deshmukh et al. explored the relationship between speed and amplitude of a gesture and Godspeed scores [32] and demonstrated that higher amplitude and speed are related to higher extroversion and neuroticism [33]. Dou et al. [34] conducted experiments under shopping scenarios to study the effects of robotic voice and gestures on the perceived personalities. Their results showed that specific informative gestures resulted in extroverted impressions. Emotion is another crucial factor that impacts the impressions of robots in human-robot interaction, and some studies have investigated the emotions that can be conveyed using the body movements of robots. In [35], gestures were displayed to participants to convey Ekman's six basic emotions, and the results suggest that social robots can achieve emotion communication merely by simple head and arm movements. Experiments [36] show that participants can successfully recognize emotions after watching videos of a robot's facial expressions; gestures were also effective for emotion recognition in addition to facial expressions. These studies investigated the traits that impact the human impressions of robots. Based on these previous results, we developed an automatic co-speech gesture generation system that can control these traits to control the impressions of robots.

III. METHODOLOGY

In our proposed model, the gesture generation is to model the conditional gesture distribution when given the speech and a trait label. Specifically, using the extracted features from a speech segment and a trait label, we sample gestures from the estimated gesture distribution learned by a deep neural network.

A. Data pre-processing

We used prosodic features as the speech features, including voice pitch and power. The feature estimation interval was 10ms. For the voice pitch, the conventional autocorrelation-based method was used to estimate the value of fundamental frequency (F0), which were logarithmic in semitone intervals [37]. The power were in dB units. Therefore, the speech features are 2-dimensional. Although the dataset we chose contains gesture data for the whole body, we only used the upper body joints to train our model because the movements in its gestures are correlated more with the impressions of robots under co-speech gesture generation settings and for comparison with the baseline model. Therefore, there are 12 target joints, each of which has 3 rotation values: roll, pitch, and yaw. The dimension of gestures is 36 in total.

B. Pseudo labels

The speed and the amplitude of gestures are the traits that we controlled. For the speed trait, the joint rotation values were converted to coordinate values in 3D space using forward kinematics. The bone lengths and initial coordinate values necessary for the calculation are provided by the dataset. After obtaining the x, y, and z coordinate values of all the joints, the velocity is calculated:

$$vel = \frac{1}{T}\sum_{t=1}^{T}(\sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2 + (z_t - z_{t-1})^2}) \tag{1}$$

where t is a specific frame of a gesture sequence, T is the total length of a gesture segment.

We define the amplitude trait as the average maximum distance between any two frames of the two hand positions, based on the calculated coordinate values:

$$amp = \frac{1}{2}(\max_{i,j \in T}(dist(lh_i, lh_j)) + \max_{i,j \in T}(dist(rh_i, rh_j))) \quad (2)$$

where lh and rh stand for the left and right hands. $dist$ is defined as

$$dist(a, b) = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2 + (a_z - b_z)^2} \quad (3)$$

where x, y, and z represent coordinates along different axes in 3D.

After the speed and the amplitude are calculated for all the gesture segments in the dataset, they are normalized by Eq. (4) across all the samples respectively, since the speed and amplitude have different scales.

$$x_{normalized} = (x - \mu)/\sigma \quad (4)$$

where μ is the mean, σ is the standard deviation.

We defined three classes for the gesture traits: low, $neutral$, and $high$. low means that the speed and the amplitude of a specific gesture sample are below the average of the dataset, $neutral$ means that they are at the dataset's average, and $high$ means that they are above the average.

to divide the gesture samples in the dataset to different classes, we applied K-means [38] to all of the gestures by setting the number of categories to three and with the distance metric:

$$l(x, y) = \frac{1}{2}\sqrt{(vel_x - vel_y)^2 + (amp_x - amp_y)^2} \quad (5)$$

The divided results are encoded as one-hot labels and assigned to the corresponding gesture samples for subsequent processing. Note that we use K-means only for dividing, rather than clustering. Dividing can be performed by other algorithms or manually.

C. Generator and discriminator

Fig. 1 shows an overview of our proposed model. While our proposed model resembles a previously proposed model [21], the main difference is that our proposed model has additional conditional input for gesture traits both for the generator and the discriminator.

Denoting the extracted speech features as $s \in \mathcal{R}^{T*2}$, the trait one-hot label as $c \in [0, 1]^3$ with $\sum_{i=1}^{3} c_i = 1$, the previous poses as $p_{prev} \in \mathcal{R}^{T_{prev}*36}$, the noise vector as $z \in \mathcal{R}^{dz}$ where dz is the dimension of the noise vector, and the generator is f_g, the output of f_g is defined as output gestures $g \in \mathcal{R}^{T*36}$, where T is the length of one segment:

$$g = f_g(s, c, p_{prev}, z) \quad (6)$$

Since z and c do not possess the time dimension, they are replicated T times and concatenated with the speech features. In the experiment, we set $T = 34$ and $T_{prev} = 4$, both of which are consistent with [21]. Therefore, the input speech feature segment has 34 frames (1.7 seconds) with four frames (0.2 seconds) that overlap between consecutive segments. The gestures in each segment also have 34 frames.

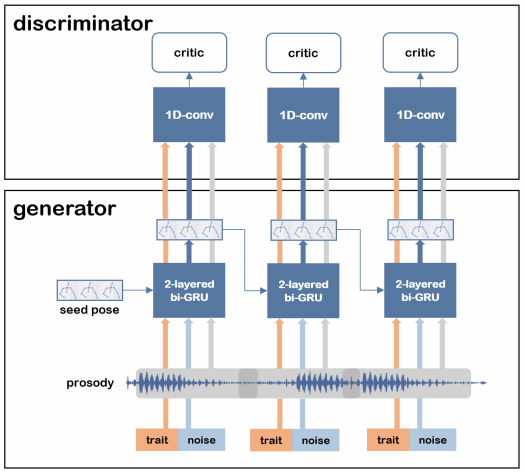

Fig. 1. Overview of proposed model.

The overlap period between consecutive gesture segments is averaged over the last four of the previous gesture segment and the first four frames of the next gesture segment as in [13], [17], [21]. Finally, a low-pass filter is used to post-process the generated rotation values as in [21], [24] to avoid jerky motions.

The discriminator consists of 1-dimensional convolution layers, whose input is the concatenation of the gesture segment, the speech feature segment, and the trait label. The output is a scalar that indicates the distance between the real data distribution and the generated distribution. The discriminator's output is defined as:

$$critic = f_d(s, c, g) \quad (7)$$

where f_d denotes the discriminator as a function, and $critic \in \mathcal{R}$.

D. Training

We used a combination of different losses as proposed in [21], including adversarial loss, continuity loss, and gradient penalty (GP)[39]:

$$\mathcal{L} = \mathcal{L}_{critic} + \lambda_{gp} * \mathcal{L}_{gp} + \lambda_c * \mathcal{L}_{continuity} \quad (8)$$

$$\mathcal{L}_{critic} = \max_{f_d} \min_{f_g} \frac{1}{M} \sum_{i=1}^{M} f_d(s^{(i)}, c^{(i)}, y^{(i)}) - f_d(s^{(i)}, c^{(i)}, f_g(s^{(i)}, c^{(i)}, p_{prev}^{(i)}, z)) \quad (9)$$

$$\mathcal{L}_{gp} = \frac{1}{M} \sum_{i=1}^{M} (\|\nabla_{f_d}^{(i)}\|_{L2} - 1)^2 \quad (10)$$

$$\mathcal{L}_{continuity} = \frac{1}{M} \sum_{i=1}^{M} \text{Huber}(g_{:k}^{(i)}, y_{-k:}^{(i-1)}) \quad (11)$$

where y is the ground truth motion, M is the number of training samples, k is a hyper-parameter to define how many frames are considered in the continuity loss. Huber loss is defined in [40].

We trained our model by setting the learning rate to 10^{-4} for both the generator and discriminator. The batch size was 128. The lambda for continuity loss was set to 1 and for the GP is set to 10. The distribution from which we sample noise vector is a Gaussian with zero mean and unit variance.

IV. Experiment

We trained the proposed model on a multimodal Japanese dataset [12]. Gestures were recorded using Motion Capture Software. The dataset contains 1094 motion and audio pairs. The total length of the data was around five hours.

We compared our proposed method with different control groups through video clips. The details are shown below:

- **Ground truth (GT)**: GT are the data recorded from human using Motion Capture Software, i.e., the test set. A low-pass filter was applied to avoid jerky motions.
- **Baseline**: A probabilistic gesture generation model [21] was used as our baseline, where only prosodic feature were used as input. This model was chosen because our model resembles it, and it has no explicit control of the impressions, matching our purpose. This model is the state-of-the-art for the dataset used in this study.
- **Proposed model**. Our proposed model generates gestures for different label input, i.e., low, $neutral$, and $high$, conditioned on speech input. Therefore, gestures are generated for all the labels for the evaluation. For convenience, in the rest of this paper, we refer to the proposed model with low, $neutral$ and $high$ trait labels as m_{low}, $m_{neutral}$ and m_{high}, respectively.

To subjectively evaluate the joint rotation values, we used the protocol provided by a previous study [21] to visualize the rotation values on an avatar using Unity[1] (Fig. 2 left), and created videos using the self-contained recording tool. We also visualized the result on a desktop humanoid robot named CommU (Fig. 2 right). CommU has only four degree of freedoms (DOFs) for hand gestures (2 DOFs for each hand), which are the pitch and yaw of the shoulder joints, for both sides. To implemented gestures on CommU, we first calculated the coordinates of each joint using forward kinematics. Then we calculated the pitch and yaw for shoulder joints using the coordinates of the shoulders and hands for both sides following inverse kinematics. The pitch and yaw are constrained to avoid collision with the body and sent to CommU via TCP/IP protocol as formatted commands. The videos were taken by a camera.

Totally, five groups were evaluated: GT, baseline, m_{low}, $m_{neutral}$, and m_{high}. We first randomly chose three utterance samples from the test set, then used the speech signals to generate gestures using the baseline and the proposed model with different trait labels, resulting in 12 samples. Added to the ground truth (GT), 15 videos were recorded both for avatar and CommU, called the "with-audio" condition. Additionally, to investigate possible mismatches of the impressions between gestures and speech, we split the video into muted and audio-only for evaluation. The muted video is called the "no-audio" condition, and the audio-only sample is called the "audio-only" condition. If the evaluation result for the audio-only agrees with that of the muted video, the audio and the gestures are considered as matched, otherwise they are considered as mismatched.

[1] https://unity.com/

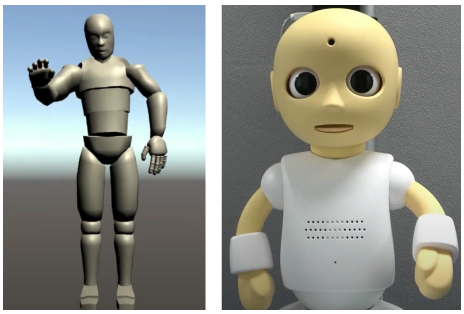

Fig. 2. Snapshots of avatar (left) and CommU (right).

A. Materials and procedure

We made a questionnaire with six sections for the subjective evaluation. In the first section, after a short description of the experiment, participants provided their gender, age, and the time they started the questionnaire for data screening. In the second section, muted videos of the avatar are presented in a randomized order. After watching each video, participants assigned a score to each of the following questions: (1) What kind of personality do you think the avatar/robot has? (2) How excited do you think the avatar/robot is? The score ranged from 1 to 7 for both questions. 1 stands for extremely introverted or not excited and 7 stands for extremely extroverted or excited. In the third section, the audio clips are presented in a randomized order. After listening to each clip of audio, participants assigned a score to each of the same questions in the second section. In the fourth section, the with-audio videos of the avatar are presented in a randomized order. After watching each one, participants assigned a score to each of the same questions in the second section. In the fifth section, the with-audio videos of the humanoid CommU are presented in a randomized order. After watching each video, participants assigned a score to each of the same questions in the second section.

We recruited through a cloud sourcing service 50 participants (24 males, 26 females, all native Japanese speakers, average age = 36.02, standard deviation = 8.38 years old). All took a reasonable duration to complete the questionnaire (average = 28 minutes, standard deviation = 7 minutes, maximum = 52 minutes, minimum = 18 minutes). The total length of the videos in the experiment was 13 minutes.

B. Result

We averaged the scores across all three samples for the two questions in the questionnaire and used them as the final score for each model, both for the avatar and the humanoid CommU. Since two objective impressions are involved in our experiment, i.e., extroversion and excitement, we presented them separately. We statistically tested the results by

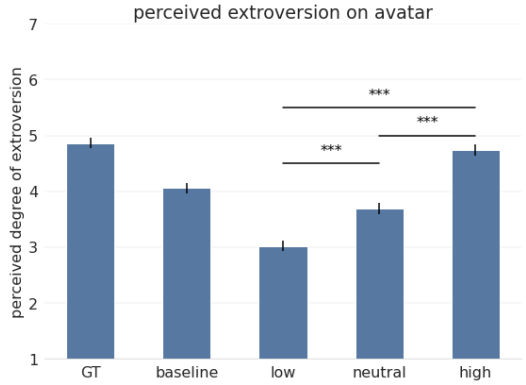

Fig. 3. Subjective scores for extroversion under with-audio condition for avatar. Only p-values within different groups of proposed model are annotated. Error bar represents standard error. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. (GT: $M = 4.86$, $SE = 0.09$. baseline: $M = 4.05$, $SE = 0.1$. m_{low}: $M = 3.01$, $SE = 0.09$. $m_{neutral}$: $M = 3.69$, $SE = 0.09$. m_{high}: $M = 4.74$, $SE = 0.1$.)

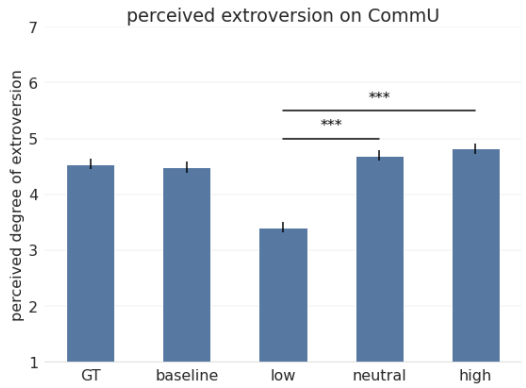

Fig. 4. Subjective scores for extroversion under with-audio condition for CommU. Only p-values within different groups of proposed model are annotated. Error bar represents standard error. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. (GT: $M = 4.54$, $SE = 0.09$. baseline: $M = 4.49$, $SE = 0.1$. m_{low}: $M = 3.41$, $SE = 0.09$. $m_{neutral}$: $M = 4.69$, $SE = 0.09$. m_{high}: $M = 4.82$, $SE = 0.09$.)

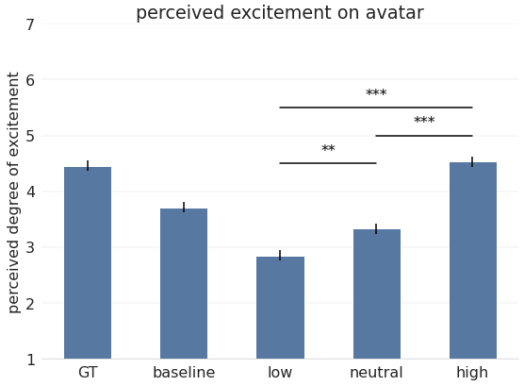

Fig. 5. Subjective scores for excitement under with-audio condition for avatar. Only p-values within different groups of proposed model are annotated. Error bar represents standard error. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. (GT: $M = 4.45$, $SE = 0.09$. baseline: $M = 3.71$, $SE = 0.09$. m_{low}: $M = 2.85$, $SE = 0.09$. $m_{neutral}$: $M = 3.33$, $SE = 0.09$. m_{high}: $M = 4.53$, $SE = 0.1$.)

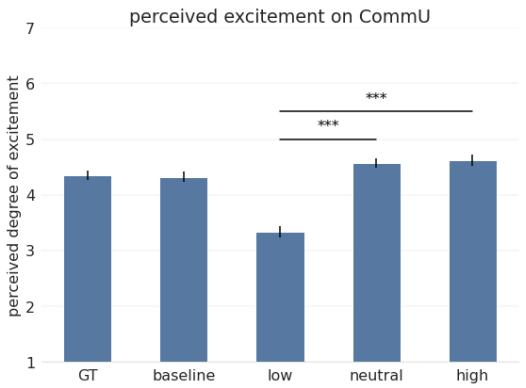

Fig. 6. Subjective scores for excitement under with-audio condition for CommU. Only p-values within different groups of proposed model are annotated. Error bar represents standard error. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. (GT: $M = 4.35$, $SE = 0.09$. baseline: $M = 4.32$, $SE = 0.09$. m_{low}: $M = 3.33$, $SE = 0.1$. $m_{neutral}$: $M = 4.57$, $SE = 0.09$. m_{high}: $M = 4.62$, $SE = 0.1$.)

first performing an analysis of variance (ANOVA). Then a Tukey's honestly significant difference (Tukey HSD) was used to investigate the significant differences between the groups pair-wisely. The alpha was set to 0.05. Additionally, we performed a power analysis to calculate the statistical power (SP) for pairs whose difference is significant according to their p-value to ensure that the results are sufficiently predictive for the single measurement on impressions. The lower limit of acceptance was set to 0.8. An SP smaller than the threshold means that the sample size is insufficient to conclude whether there is a significant difference by the specific p-value. We calculated Cohen's d as the effect size for each pair, which indicates how large the difference is between groups. The larger it is, the stronger is the correlation between variables, i.e., the treatment and the effect.

1) Extroversion: For the avatar, Fig. 3 shows the perceived extroversion on avatar for all the previously defined systems. As expected, a gradation of the perceived extroversion for the proposed model was obtained that matches the purpose of the proposed pseudo-labeling method. The degree of the perceived extroversion of m_{low} is lower than that of $m_{neutral}$ ($p < .001$, $SP = 1.0$, $d = 1.501$). The degree of the perceived extroversion of $m_{neutral}$ is lower than that of m_{high} ($p < .001$, $SP = 1.0$, $d = 0.9$). Compared with GT, both the degrees of the perceived extroversion of m_{low} and $m_{neutral}$ are lower than GT ($p < .001$, $SP = 1.0$, $d = 1.654$, and $p < .001$, $SP = 1.0$, $d = 1.031$, respectively). Additionally, the difference between GT and m_{high} is not statistically significant ($p = .89$, $d = 0.105$). This suggests that the proposed model can only lower the perceived degree of extroversion with respect to GT. Compared to the baseline, the proposed model lowered the perceived degree of extroversion by choosing m_{low} ($p < .001$, $SP = 1.0$, $d = 0.903$) or raised it by choosing m_{high} ($p < .001$, $SP = 0.955$, $d = 0.581$). There is no significant difference between baseline and $m_{neutral}$ ($p < 0.05$ and $d = 0.313$; however, since $SP = 0.771$ is smaller than 0.8, the null

hypothesis was not rejected). Surprisingly, the degree of the perceived extroversion of the baseline is lower than that of GT ($p < .001$, $SP = 0.997$ $d = 0.702$), which contradicts our expectation that the baseline would be the same as GT.

For CommU, Fig. 4 shows the perceived extroversion on CommU for all the previously defined systems. The degree of the perceived extroversion of m_{low} is lower than that of $m_{neutral}$ ($p < .001$, $SP = 1.0$, $d = 1.158$) and m_{high} ($p < .001$, $SP = 1.0$, $d = 1.265$). Compared with GT, the degree of the perceived extroversion of m_{low} is lower than GT ($p < .001$, $SP = 1.0$, $d = 1.005$). Additionally, the difference between GT and m_{high} is not statistically significant ($p = .21$ and $d = 0.249$), as well as the difference between GT and $m_{neutral}$ ($p = .74$, $d = 0.137$). This suggests that the proposed model can only lower the perceived degree of extroversion for the CommU compared with GT, which is consistent with the avatar results. Compared with the baseline, the proposed model lowered the perceived degree of extroversion by choosing m_{low} ($p < .001$, $SP = 1.0$, $d = 0.938$).

2) Excitement: For the avatar, Fig. 5 shows the perceived degree of excitement on avatar for all the previously defined systems. As expected, a gradation of the perceived degree of excitement for the proposed model was obtained that matches the purpose of the proposed pseudo-labeling method. The degree of the perceived degree of excitement of m_{low} is lower than that of $m_{neutral}$ ($p < .005$, $SP = 0.813$, $d = 0.43$). The degree of the perceived degree of excitement of $m_{neutral}$ is lower than that of m_{high} ($p < .001$, $SP = 1.0$, $d = 1.044$). Compared with GT, both the perceived degrees of excitement of m_{low} and $m_{neutral}$ are lower than GT ($p < .001$, $SP = 1.0$, $d = 1.423$, and $p < .001$, $SP = 1.0$, $d = 0.993$, respectively). Additionally, the difference between GT and m_{high} is not statistically significant ($p = .9$, $d = 0.063$). This suggests that the proposed model can only lower the perceived degree of excitement with respect to GT. Compared with the baseline, the proposed model lowered the perceived degree of excitement by choosing m_{low} ($p < .005$, $SP = 1.0$, $d = 0.784$) and $m_{neutral}$ ($p < .05$, $SP = 0.852$, $d = 0.348$) or raised it by choosing m_{high} ($p < .001$, $SP = 0.998$, $d = 0.715$). Surprisingly, the degree of the perceived extroversion of the baseline is lower than that of GT ($p < .001$, $SP = 0.991$, $d = 0.659$), which contradicts our expectation that the baseline would be the same as GT.

For CommU, Fig. 6 shows the perceived excitement on CommU for all the previously defined systems. The degree of the perceived excitement of m_{low} is lower than that of $m_{neutral}$ ($p < .001$, $SP = 1.0$, $d = 1.101$) and m_{high} ($p < .001$, $SP = 1.0$, $d = 1.097$). Compared with GT, the degree of the perceived excitement of m_{low} is lower than GT ($p < .001$, $SP = 1.0$, $d = 0.902$). Additionally, the difference between GT and m_{high} is not statistically significant ($p = .25$, $d = 0.238$), as well as the difference between GT and $m_{neutral}$ ($p = .48$, $d = 0.2$). This suggests that the proposed model can only lower the perceived degree of excitement for CommU compared with GT, which is consistent with the results of avatar. Compared with baseline, the proposed model lowered the perceived degree of excitement by choosing m_{low} ($p < .001$, $SP = 1.0$, $d = 0.902$).

V. Discussion

A. Impact of gestures on impressions

1) Avatar: As shown in Fig. 3, the proposed model can control extroversion and excitement by managing the gesture's speed and amplitude. Additionally, while the baseline obtained a score of 4, which means neither introverted nor extroverted, a gradation was obtained from introvert to extrovert . The evaluation of excitement shows similar results (Fig. 5). Thus, we conclude that our model can cover a range of extroversion and excitement for the avatar. However, none of the scores of the three groups of the proposed model exceeds GT, indicating that the proposed model can only lower the perception of extroversion or excitement from GT. This suggests that extrapolation methods should be considered to cover a wider range in future work. Additionally, the difference in the impressions between GT and baseline suggests that the latter failed to capture the former's impression, which indicates that the baseline's learning algorithm is not optimal.

2) CommU: As shown in Fig. 4, the model can only lower the extroversion or excitement scores, which is consistent with the avatar results. However, no significant difference was observed between $m_{neutral}$ and m_{high}, which contradicts the obtained avatar results. This may be due to fewer DOFs and CommU's actuator constraints, which makes CommU less expressive in terms of gestures. However, more experiments are necessary to clarify the reason.

B. Mismatch between speech and gesture

To investigate the effects of different modalities on the impressions of extroversion and excitement, we analyzed the assigned scores of the audio-only condition (i.e., without motion) and the no-audio (muted) and with-audio conditions accompanied with motions generated by the proposed model with the avatar. Note that since the no-audio condition is unavailable for CommU, the results in this subsection are based on the impressions of the avatar.

1) Extroversion: Fig. 7 shows the perceived degree of extroversion for different conditions. Between audio and m_{low}, since the perceived degree of extroversion of the audio-only is higher than no-audio ($p < .001$, $SP = 1.0$, $d = 1.53$), they are mismatched. The perceived degree of extroversion becomes lower when combined with gestures, i.e., with-audio condition ($p < .001$, $SP = 1.0$, $d = 0.951$). Similarly, between audio and m_{high}, since the perceived degree of extroversion of the audio-only is higher than no-audio ($p < .001$, $SP = 1.0$, $d = 1.014$), they are mismatched. The perceived degree of extroversion becomes higher when combined with gestures, i.e., the with-audio condition ($p < .001$, $SP = 0.893$, $d = 0.528$). This effect also exists for the neutral condition. The perceived degree of extroversion of audio-only is higher than the no-audio ($p < .05$, $SP = 0.807$, $d = 0.327$), and thus they are mismatched. The perceived degree of extroversion becomes

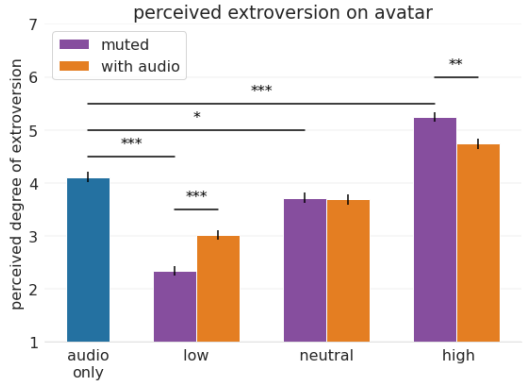

Fig. 7. Subjective scores for extroversion under audio-only, no-audio, and with-audio conditions for avatar. All groups with significant differences are annotated with their p-values. Error bar represents standard error. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. (audio-only: $M = 4.11$, $SE = 0.1$. no-audio m_{low}: $M = 2.34$, $SE = 0.09$. no-audio $m_{neutral}$: $M = 3.72$, $SE = 0.1$. no-audio m_{high}: $M = 5.25$, $SE = 0.08$.)

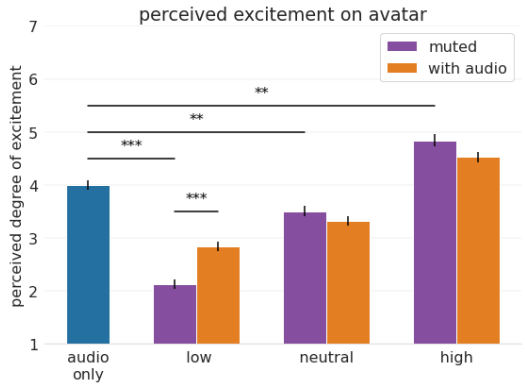

Fig. 8. Subjective scores for excitement under audio-only, no-audio, and with-audio conditions for the avatar. All groups with significant differences are annotated with their p-values. Error bar represents standard error. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. (audio-only: $M = 4.0$, $SE = 0.09$. no-audio m_{low}: $M = 2.13$, $SE = 0.09$. no-audio $m_{neutral}$: $M = 3.51$, $SE = 0.1$. no-audio m_{high}: $M = 4.84$, $SE = 0.11$.)

higher when combined with gestures, i.e., the with-audio condition ($p < .05$, $SP = 0.879$, $d = 0.363$). These changes suggest that gestures whose perceived degree of extroversion are different from their corresponding audio alter the avatar's overall extroversion.

Interestingly, the impact on the perceived extroversion by the gesture can also be interpreted as their impact on the gesture by the audio. The decrease from audio-only to with-audio m_{low} and the increase from audio-only to with-audio m_{high} can also be explained as the increase from no-audio m_{low} to m_{low} ($p < .001$, $SP = 0.969$, $d = 0.6$), and the decrease from no-audio m_{high} to m_{high} ($p < .005$, $SP = 0.867$, $d = 0.456$). However, this is not the case for $m_{neutral}$ ($p = .9$, $d = 0.028$). It can be explained by observing that although there is a difference between audio-only and no-audio $m_{neutral}$, it is much smaller than that between no-audio m_{low} and audio-only, and between no-audio m_{high} and audio-only, which weakens the effect.

2) Excitement: Fig. 8 shows the perceived degree of excitement for different conditions. Between audio and m_{low}, since the perceived degree of excitement of the audio-only is higher than the no-audio ($p < .001$, $SP = 1.0$, $d = 1.65$), they are mismatched. The perceived degree of excitement becomes lower when combined with gestures, i.e., the with-audio condition ($p < .001$, $SP = 1.0$, $d = 1.015$). Similarly, between audio and m_{high}, since the perceived degree of excitement of audio-only is higher than no-audio ($p < .01$, $SP = 1.0$, $d = 1.65$), they are mismatched. The perceived degree of excitement becomes higher when combined with gestures, i.e., the with-audio condition ($p < .01$, $SP = 1.0$, $d = 1.015$). This effect also exists for the neutral condition. The perceived degree of excitement of the audio-only is higher than the no-audio ($p < .01$, $SP = 0.833$, $d = 0.411$), and thus they are mismatched. The perceived degree of excitement becomes higher when combined with gestures, i.e., the with-audio condition ($p < .01$, $SP = 0.994$, $d = 0.59$). These changes suggests that gestures whose perceived degree of excitement are different from their corresponding audio can alter the overall excitement of the avatar.

Similarly to extroversion, the impact on the perceived excitement by the gesture can also be interpreted as the impact on the gesture by the audio. However, there is neither a significant difference between no-audio $m_{neutral}$ and with-audio $m_{neutral}$ ($p = .85$, $d = 0.152$), nor between no-audio m_{high} to with-audio m_{high} ($p = .25$, $d = 0.246$), indicating that the audio has no impact on the gestures generated by $m_{neutral}$ and m_{high}. This may be explained by that the difference of the perceived excitement between them is relatively small. The differences between the audio-only and the no-audio $m_{neutral}$ and between audio-only and no-audio m_{high} are smaller than the difference between audio-only and no-audio m_{low}, weakening the effect.

VI. CONCLUSION

We proposed a conditional GAN-based co-speech gesture generation model that exploits cognitive heuristics while maintaining the flexibility of data-driven methods. The experimental results on VR avatars/humanoid robots showed that our model controlled the perceived extroversion and excitement, and such findings are consonant with those from human cognitive behavior. Furthermore, our proposed model highlights the potential of fine control with a "black box" data-driven method. The burden of collecting various big data can be moderated by adopting the findings from cognitive science. Future works will further investigate the combination of learned methods and the revelations of cognitive science.

REFERENCES

[1] M. Destephe, M. Brandao, T. Kishi, M. Zecca, K. Hashimoto, and A. Takanishi, "Walking in the uncanny valley: Importance of the attractiveness on the acceptance of a robot as a working partner," *Frontiers in Psychology*, vol. 6, p. 204, 2015.

[2] Y. Yamashita, H. Ishihara, T. Ikeda, and M. Asada, "Appearance of a robot influences causal relationship between touch sensation and the personality impression," in *Proceedings of the International Conference on Human Agent Interaction*, 2017, pp. 457–461.

[3] R. Tamagawa, C. I. Watson, I. H. Kuo, B. A. MacDonald, and E. Broadbent, "The effects of synthesized voice accents on user perceptions of robots," *International Journal of Social Robotics*, vol. 3, no. 3, pp. 253–262, 2011.

[4] I. Torre, J. Goslin, L. White, and D. Zanatto, "Trust in artificial voices: A" congruency effect" of first impressions and behavioural experience," in *Proceedings of the Technology, Mind, and Society*, 2018, pp. 1–6.

[5] S. Ryoko, F. Chie, K. Takatsugu, S. Kaori, H. Yuki, O. Motoyuki, and O. Natsuki, "Does talking to a robot in a high-pitched voice create a good impression of the robot?" in *ACIS*. IEEE, 2012, pp. 19–24.

[6] C. Thepsoonthorn, K.-i. Ogawa, and Y. Miyake, "The relationship between robot's nonverbal behaviour and human's likability based on human's personality," *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.

[7] G. Hoffman, G. E. Birnbaum, K. Vanunu, O. Sass, and H. T. Reis, "Robot responsiveness to human disclosure affects social impression and appeal," in *International conference on Human-robot interaction*, 2014, pp. 1–8.

[8] H. Kim, S. S. Kwak, and M. Kim, "Personality design of sociable robots by control of gesture design factors," in *International Symposium on Robot and Human Interactive Communication*. IEEE, 2008, pp. 494–499.

[9] K. Bergmann, F. Eyssel, and S. Kopp, "A second chance to make a first impression? how appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time," in *International conference on intelligent virtual agents*. Springer, 2012, pp. 126–138.

[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[11] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *IEEE conference on computer vision and pattern recognition*, 2018, pp. 7297–7306.

[12] K. Takeuchi, S. Kubota, K. Suzuki, D. Hasegawa, and H. Sakuta, "Creating a gesture-speech dataset for speech-based automatic gesture generation," in *International Conference on Human-Computer Interaction*. Springer, 2017, pp. 198–202.

[13] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots," in *International Conference on Robotics and Automation*. IEEE, 2019, pp. 4303–4309.

[14] Y. Ferstl, M. Neff, and R. McDonnell, "Multi-objective adversarial gesture generation," in *Motion, Interaction and Games*, 2019, pp. 1–10.

[15] C. T. Ishi, D. Machiyashiki, R. Mikata, and H. Ishiguro, "A speech-driven hand gesture generation method and evaluation in android robots," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3757–3764, 2018.

[16] S. Alexanderson, G. E. Henter, T. Kucherenko, and J. Beskow, "Style-controllable speech-driven gesture synthesis using normalising flows," in *Computer Graphics Forum*, vol. 39, no. 2. Wiley Online Library, 2020, pp. 487–496.

[17] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Transactions on Graphics*, vol. 39, no. 6, pp. 1–16, 2020.

[18] S. Taylor, J. Windle, D. Greenwood, and I. Matthews, "Speech-driven conversational agents using conditional flow-vaes," in *European Conference on Visual Media Production*, 2021, pp. 1–9.

[19] T. Kucherenko, R. Nagy, P. Jonell, M. Neff, H. Kjellström, and G. E. Henter, "Speech2properties2gestures: Gesture-property prediction as a tool for generating representational gestures from speech," in *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 2021, pp. 145–147.

[20] T. Kucherenko, D. Hasegawa, N. Kaneko, G. E. Henter, and H. Kjellström, "Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation," *International Journal of Human–Computer Interaction*, pp. 1–17, 2021.

[21] B. Wu, C. Liu, C. T. Ishi, and H. Ishiguro, "Probabilistic human-like gesture synthesis from speech using gru-based wgan," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 194–201.

[22] D. Hasegawa, N. Kaneko, S. Shirakawa, H. Sakuta, and K. Sumi, "Evaluation of speech-to-gesture generation using bi-directional lstm network," in *18th International Conference on Intelligent Virtual Agents*, 2018, pp. 79–86.

[23] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 97–104.

[24] B. Wu, C. Liu, C. T. Ishi, and H. Ishiguro, "Modeling the conditional distribution of co-speech upper body gesture jointly using conditional-gan and unrolled-gan," *Electronics*, vol. 10, no. 3, p. 228, 2021.

[25] L. Robert, "Personality in the human robot interaction literature: A review and brief critique," in *Proceedings of the 24th Americas Conference on Information Systems*, 2018, pp. 16–18.

[26] J. Hwang, T. Park, and W. Hwang, "The effects of overall robot shape on the emotions invoked in users and the perceived personalities of robot," *Applied ergonomics*, vol. 44, no. 3, pp. 459–471, 2013.

[27] B. Tay, Y. Jung, and T. Park, "When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction," *Computers in Human Behavior*, vol. 38, pp. 75–84, 2014.

[28] L. Robert, R. Alahmad, C. Esterwood, S. Kim, S. You, and Q. Zhang, "A review of personality in human–robot interactions," *SSRN 3528496*, 2020.

[29] M. Neff, Y. Wang, R. Abbott, and M. Walker, "Evaluating the effect of gesture and language on personality perception in conversational agents," in *International Conference on Intelligent Virtual Agents*. Springer, 2010, pp. 222–235.

[30] M. McRorie, I. Sneddon, G. McKeown, E. Bevacqua, E. De Sevin, and C. Pelachaud, "Evaluation of four designed virtual agent personalities," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 311–322, 2011.

[31] A. Mileounis, R. H. Cuijpers, and E. I. Barakova, "Creating robots with personality: The effect of personality on social intelligence," in *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 2015, pp. 119–132.

[32] A. Deshmukh, B. Craenen, A. Vinciarelli, and M. E. Foster, "Shaping robot gestures to shape users' perception: The effect of amplitude and speed on godspeed ratings," in *Proceedings of the 6th International Conference on Human-Agent Interaction*, 2018, pp. 293–300.

[33] B. 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 *27th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2018, pp. 699–704.

[34] X. Dou, C.-F. Wu, K.-C. Lin, and T.-M. Tseng, "The effects of robot voice and gesture types on the perceived robot personalities," in *International Conference on Human-Computer Interaction*. Springer, 2019, pp. 299–309.

[35] J. Li and M. Chignell, "Communication of emotion in social robots through simple head and arm movements," *International Journal of Social Robotics*, vol. 3, no. 2, pp. 125–142, 2011.

[36] S. Costa, F. Soares, and C. Santos, "Facial expressions and gestures to convey emotions with a humanoid robot," in *International Conference on Social Robotics*. Springer, 2013, pp. 542–551.

[37] C. T. Ishi, H. Ishiguro, and N. Hagita, "Automatic extraction of paralinguistic information using prosodic features related to f0, duration and voice quality," *Speech communication*, vol. 50, no. 6, pp. 531–543, 2008.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[39] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *34th International Conference on Machine Learning*, 2017, pp. 214–223.

[40] R. Girshick, "Fast r-cnn," in *IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.