2022 31st IEEE International Conference on Robot and
Human Interactive Communication (RO-MAN)
August 29 - September 2, 2022. Naples, Italy.

2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) | 978-1-7281-8859-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/RO-MAN53752.2022.9900760

# Agree or Disagree? Generating Body Gestures from Affective Contextual Cues during Dyadic Interactions

Nguyen Tan Viet Tuyen and Oya Celiktutan

***Abstract*— Humans naturally produce nonverbal signals such as facial expressions, body movements, hand gestures, and tone of voice, along with words, to communicate their messages, opinions, and feelings. Considering robots are progressively moving out from research laboratories into human environments, it is increasingly desirable that they develop a similar social intelligence. Therefore, equipping social robots with nonverbal communication skills has been an active research area for decades, where data-driven, end-to-end learning approaches have become predominant in recent years, offering scalability and generalisability. However, most of these approaches consider a single character, modelling intrapersonal dynamics only. In this paper, we propose a method based on conditional Generative Adversarial Networks, intending to generate behaviours for a robot in affective dyadic interactions. Our method takes as an input the audio of a target person together with the nonverbal signals of their interacting partner, modelled by a novel *Context Encoder*, to generate appropriate body gestures. We evaluate our method on the multimodal JESTKOD dataset that comprises dyadic interactions under agreement and disagreement scenarios. The experimental results show that *Context Encoder* can better contribute to the prediction of co-speech gestures in agreement situations.**

## I. INTRODUCTION

Social robots will progressively become widespread in many aspects of our daily lives, including education, healthcare, workplace, and home. All of such practical applications require that humans and robots interact and collaborate with each other seamlessly. Along with verbal communication, successful social interaction is closely coupled with the exchange of nonverbal cues, such as gaze, facial expressions, body movements, and hand gestures. Humans perform social interaction in an instinctive and adaptive manner, with no effort. For robots to be successful in our social landscape, they should therefore engage in social interactions in a human-like manner, with increasing levels of autonomy. Motivated by this, imitating nonverbal communication has been an active area of research to enhance the clarity of the human-robot interaction interfaces and the sense of rapport, hence maximise the user trust and acceptance of them.

Early methods have focused on rule-based approaches [1], requiring the design of interaction logic manually, which is notoriously difficult, taking into account the complexity of social interactions. Once fixed, it will be limited, not transferrable to unseen interaction contexts, and not robust to unpredicted inputs from the robot's environment (e.g., sensor noise). Therefore, data-driven, end-to-end learning approaches [2], [3], [4] has been a promising solution to address these shortcomings. However, so far only a handful of works [5], [6], [7], [8] aim to generate behaviours by taking into account the interaction context, namely, the nonverbal signals of the interaction partner. Although social interaction is an open-ended concept, it can be formalised through two main processes: (i) Perception – perception process involves receiving visual stimuli about the behaviours of others, or the state of the interaction; and (ii) Action – action process is the generation of a behaviour by taking into account all aspects of interaction including current perceived states and history. Therefore, it is necessary to integrate what the interaction partner says and how they say it to be able to create socially suitable behaviours for robots.

In this paper, we propose a method based on conditional Generative Adversarial Networks, with an ultimate goal of generating behaviours for a robot in affective dyadic interactions. Our method takes as input the audio of a target person together with the nonverbal signals of their interaction partner, modelled by a novel Context Encoder, to generate appropriate body gestures. Differently from existing works, we particularly focus on agreement and disagreement scenarios, due to their prevalence in a wide range of daily interaction situations. Previous research has shown that nonverbal signals play a crucial role in the communication and interpretation of agreement and disagreement [9], [10], [11]. The main goal of this paper is to examine the impact of affective interactional context on the generation of body gestures, which has implications for social robots to adapt based on the changing context of the interaction.

## II. RELATED WORK

Communicative gestures are naturally performed when speaking and they are applied to convey the communicator's emotion, intention, or verbal contents of their speech. The majority of the works has used audio or text or both of them to develop body gestures. Generative Adversarial Network (GAN) has been widely used to address this problem. Among these approaches, Ahn *et al*. [2] proposed a Sequence to Sequence (Seq2Seq) model based on a GAN. The other works [3] developed methods based on conditional GANs: Tuyen *et al*. [3] designed a conditional GAN with Convolution Neural Network (CNN) operations. Bhattacharya *et al*. [10] used GANs to synthesise co-speech gestures with affective expressions.

The authors are with the Centre for Robotics Research, Department of Engineering, King's College London, London WC2R 2LS, United Kingdom {tan_viet_tuyen.nguyen; oya.celiktutan}@kcl.ac.uk

This work has been supported by the "LISI - Learning to Imitate Nonverbal Communication Dynamics for Human-Robot Social Interaction" project, funded by the Engineering and Physical Sciences Research Council (Grant Ref.: EP/V010875/1).

Apart from GAN, other approaches used include Long Short-Term Memory Networks (LSTM) [12], variational auto-encoders (VAE) [13], autoregressive models [14], encoder-decoder networks [4] based on Recurrent Neural Networks [15], and so on. While this long list indicates the need for a comparative study to investigate advantages and disadvantages of these gesture generators, Zabala *et al*. [16] showed that GAN works better as compared to VAE in terms of both quantitative and qualitative metrics, which is also the adopted approach in this paper.

The aforementioned approaches have focused on the communicator in isolation, without considering an interaction context. Huang and Khan [5] focused on the problem of facial expressions produced during interactions between an interviewee and an interviewer, and they introduced a framework based on conditional GAN. The method generated the interviewer's facial gestures that were appropriately contextualized and responsive to the interviewee's facial expressions. Similarly, Feng *et al*. [6] suggested a VAE network to handle the generation of facial cues between a user and an embodied agent. There is only a handful of approaches aiming to generate body gestures during dyadic interactions. In terms of triadic human communication, Joo *et al*. [7] presented a generative approach that acquires nonverbal signals from interacting partners and encodes them into latent vectors. The encoded features were utilized to estimate the target person's body gestures. Ahuja *et al*. [8] proposed the Dyadic Residual Attention Model to forecast the pose of an agent. Particularly, they designed an adaptive attention mechanism to select intrapersonal and interpersonal dynamics for generating the agent's future poses. However, none of these approaches has investigated the impact of interaction context on the generated cues.

## III. METHODOLOGY

### A. Problem Statement

We define the problem of speech-driven gesture generation with context awareness as follows: in a dyadic interaction between a target person $S_{fo}$ and an interaction partner $S_{ob}$, $A_{fo}^{0:T}$ denotes the speech audio of $S_{fo}$ in a temporal time window, namely $t \in [0, T]$. $P_{ob}^{0:T}$ and $A_{ob}^{0:T}$ are the co-speech gesture and the speech audio simultaneously observed from $S_{ob}$ within the same spatial and temporal window. This research aims to find a mapping function $F$ that receives $A_{fo}^{0:T}$, $P_{ob}^{0:T}$, and $A_{ob}^{0:T}$ as inputs, and predict an output co-speech gesture of $S_{fo}$, namely $P_{fo}^{0:T}$.

### B. Model Overview

To address the research question in the aforementioned section, this paper introduces a co-speech gesture generative framework with context awareness as shown in Fig. 1. The framework consists of *Context Encoder E*, *Generator G*, and *Discriminator D*. At the timestamp $t$ ($t \in [0, T]$), the training pipeline is started by encoding $P_{ob}^{t}$ into $c_P^t$, $A_{ob}^{t}$ into $c_A^t$ and $A_{fo}^{t}$ into $s_{fo}^t$. Then, $c_P^t$ and $c_A^t$ are then combined into a contextual vector, namely $c_{ob}^t$. $s_{fo}^t$, $c_{ob}^t$, and the previously generated pose $\hat{P}_{fo}^{t-1}$ are injected to $G_{Encoder}$. The internal representation encoded by $G_{Encoder}$ is then fed to $G_{Decoder}$ for producing the next motion frame $\hat{P}_{fo}^{t}$. This process is repeated until $t = T$. Finally, the generated co-speech gesture $\hat{P}_{fo}^{0:T}$ and their corresponding speech feature vector $s_{fo}^{0:T}$, contextual vector $c_{ob}^{0:T}$ are injected to $D$ for identifying samples to be either fake or real. In the sequel, the proposed network architecture is described in detail.

### C. Context Encoder

*Context Encoder* is designed to encode social signals simultaneously collected from the interacting partner in dyadic interaction into a contextual vector. *Context Encoder* consists of *Motion Encoder* and *Speech Encoder*. Here, $c_P^t$ encoded by *Motion Encoder* and $c_A^t$ encoded by *Speech Encoder* are combined into $c_{ob}^t$. $c_{ob}^t$ represents the contextual information extracted from the interaction partner $P_{ob}^t$ at the current timestamp $t$.

#### 1) Motion Encoder

The network receives the motion sequence $P_{ob}^{0:T}$ of the interaction partner $P_{ob}$ as input and delivers the output feature vector $c_P^{0:T}$. *Motion Encoder* is constructed with a sequence of fully connected (FC) layers and Long-Short Term Memory (LSTM) layers. *Motion Encoder* iteratively encodes $P_{ob}^{0:T}$ into $c_P^{0:T}$ frame-by-frame.

#### 2) Speech Encoder

The network handles the speech audio $A_{ob}^t$ as input and produces the audio feature vector $c_A^{0:T}$. From the raw audio speech, we firstly extract the MFCCs and speech prosodic features. MFCCs are well known to encode signal frequencies according to how humans perceive sounds [17] while prosodic features encompass intonation, rhythm, and other information about the speech outside of the specific words spoken (e.g. semantics and syntax) [18]. Similar to the *Motion Encoder*, *Speech Encoder* processes input speech features frame-by-frame. *Speech Encoder* is constructed with 3 Convolutional (CONV) layers, 2 LTSM layers, and 1 FC layer.

### D. Generator

*Generator G* consists of *Speech Encoder*, $G_{Encoder}$, and $G_{Decoder}$. It should be noticed that *Speech Encoder* implemented in $G$ inherits the same network architecture as the one implemented in $E$, and they share the same weight parameters. Here, at a time stamp $t$, *Speech Encoder* receives the audio speech $A_{fo}^t$ as an input and encodes it into $s_{fo}^t$. It is followed by feeding $s_{fo}$, $c_{ob}^t$, and the previously generated pose $\hat{P}_{fo}^{t-1}$ into $G_{Encoded}$. $G_{Encoder}$ is designed with a sequence of FC layers to encode the input vector into an internal representation $h_e^t$. Finally, $h_e^t$ is fed to $G_{Decoder}$ for generating the next motion frame $\hat{P}_{fo}^t$. We designed $G_{Decoder}$ with a sequence of FC layers and LSTM layers. For better modeling the velocity of generated motion, a residual connection is added between the input and the output of each LSTM cell of $G_{Decoder}$. This approach allows $G_{Decoder}$ to model the differences between $\hat{P}_{fo}^{t-1}$ and $\hat{P}_{fo}^t$ that encourages the continuity of generated motions. *Generator* can also be used independently without the need of integrating with $ContextEncoder$ and $Discriminator$.

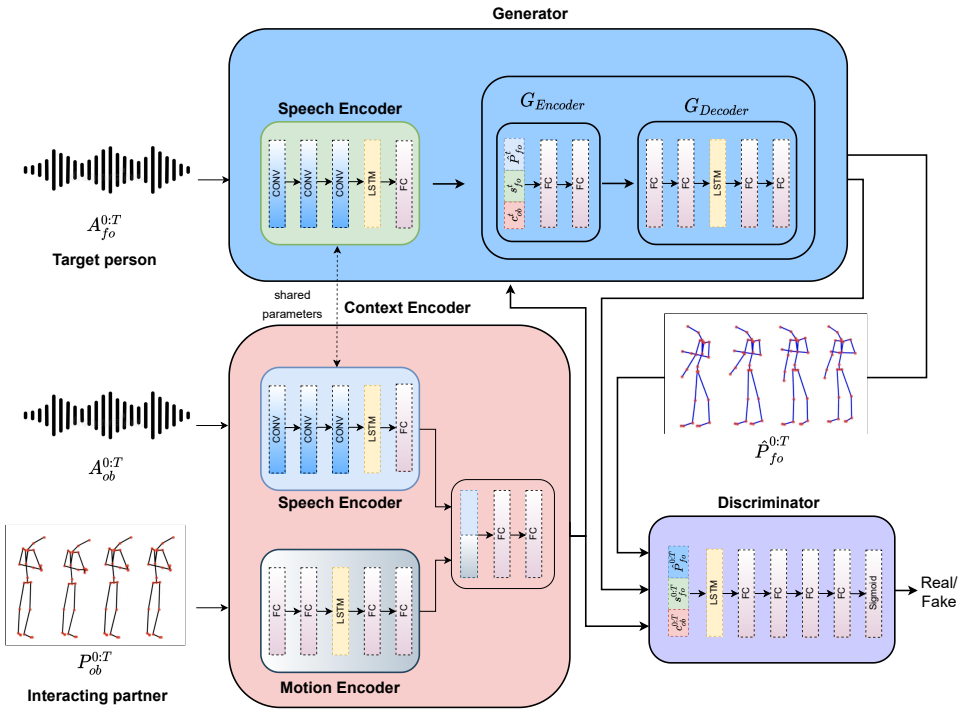



Fig. 1: The proposed framework based on conditional GAN to generate body gestures for a target person from their speech (or audio) and affective contextual cues, namely, their interaction partner's nonverbal signals encoded by the *Context Encoder*.

In this circumstance, $G$ receives $A_{fo}^{0:T}$ to predict the co-speech gesture $\hat{P}_{fo}^{0:T}$. This approach is further verified in IV-C.

### E. Discriminator

During the training phase, both real $P_{fo}^{0:T}$ and fake co-speech gestures $\hat{P}_{fo}^{0:T}$ are injected into the *Discriminator* $D$. Additionally, $D$ also takes both speech feature $s_{fo}^{0:T}$ of the target user $P_{fo}$ and the contextual vector $c_{ob}^{0:T}$ of the interaction partner $P_{ob}$ into consideration for producing the adversarial loss $y$. Here, $D$ is able to work as a smart adaptive loss function where $s_{fo}^{0:T}$ delivers information allowing $D$ to validate the speech synthesis while $c_{ob}^{0:T}$ contains information for verifying the context synchrony. $D$ is designed with 2 LSTM layers and followed by a sequence of FC layers. Output values from the last FC layer are passed through a sigmoid function to produce a probability indicating whether the input motion is real or fake.

Overall, the framework demonstrated in Fig. 1 is trained with the loss functions $L_G$ and $L_D$ defined in Eq. 1 and Eq. 2, respectively. $\alpha$ and $\beta$ are parameters to control the weights of the loss terms. The training procedure is summarized in Algorithm 1.

$$\begin{aligned}\mathcal{L}_G =& \alpha * \frac{1}{T}\sum_{t=0}^{T} ||P_{fo}^t - \hat{P}_{fo}^t||_2^2 \\ &+ \beta * log(1 - D(c_{ob}^{0:T}, s_{fo}^{0:T}, \hat{P}_{fo}^{0:T}))\end{aligned} \tag{1}$$

$$\begin{aligned}\mathcal{L}_D =& - log(D(c_{ob}^{0:T}, s_{fo}^{0:T}, P_{fo}^{0:T})) \\ &- log(1 - D(c_{ob}^{0:T}, s_{fo}^{0:T}, \hat{P}_{fo}^{0:T}))\end{aligned} \tag{2}$$

**Algorithm 1** The proposed algorithm for the training phase

**Input:** $P_{ob}^{0:T}$, $A_{ob}^{0:T}$, $P_{fo}^{0:T}$, $A_{fo}^{0:T}$

**for** s=0 to training step S **do**
  **for** t=0 to T **do**
    $c_A^t \leftarrow$ *SpeechEncoder*$(A_{ob}^t)$;
    $c_P^t \leftarrow$ *MotionEncoder*$(A_{ob}^t)$;
    $c_{ob}^t \leftarrow$ concat$(c_A^t, c_P^t)$
    $s_{fo}^t \leftarrow$ *SpeechEncoder*$(A_{fo}^t)$;
    $\hat{P}^t \leftarrow$ *Generator*$(c_{ob}^t, s_{fo}^t, \hat{P}^t)$
  **end for**
  $y_r \leftarrow$ *Discriminator*$(c_{ob}^{0:T}, s_{fo}^{0:T}, P_{fo}^{0:T})$
  $y_f \leftarrow$ *Discriminator*$(c_{ob}^{0:T}, s_{fo}^{0:T}, \hat{P}_{fo}^{0:T})$
  Update *Discriminator* with $\mathcal{L}_D$
  Update *Generator*, *Context Encoder* with $\mathcal{L}_G$
**end for**

## IV. EXPERIMENTAL RESULTS AND DISCUSSION

### A. Dataset and Pre-processing

The proposed approach was validated on the JESTKOD dataset [20], a time-synchronised speech and gesture dataset in affective dyadic interactions. The body data was collected by a motion capture system and was defined by Euler angles. This dataset allows us to model the full body gesture of an target person from co-speech (i.e., audio), while taking into consideration the contextual information simultaneously acquired from an interaction partner. The JESTKOD dataset covers a wide range conversational scenarios in different topics (e.g., movies, sport, music, etc.) carried out with 10 participants (4 females, 6 males). The dataset was collected in a such way that the participants' profiles were considered to put them into proper conversational topics to create

TABLE I: Key components of ablation models.

| No | Model | Components | | | |
|---|---|---|---|---|---|
| | | Generator | Context Encoder | | Discriminator |
| | | | Motion Encoder | Speech Encoder | |
| 1 | full model | ✓ | ✓ | ✓ | ✓ |
| 2 | w/o Discriminator | ✓ | ✓ | ✓ | none |
| 3 | w/o Context Encoder + Discriminator | ✓ | none | none | none |
| 4 | Speech to Gesture [19] | ✓ | none | none | none |

agreement/disagreement situations. For instance, soccer can initiate controversial discussions between two participants supporting different teams. The dataset consists of 56 dyadic interactions in agreement and 42 sessions in disagreement with a total duration of 154 and 105 minutes, respectively.

We divided the dataset into training and testing set. Specifically, for agreement scenarios, 41 sessions were used for training and 15 sessions were utilized for testing. For disagreement scenarios, the training set includes 30 sessions while the testing set consists of 12 sessions. The recordings of motion and speech were down-sampled into a common frame rate of 20 frames per second (fps). From the audio recordings, we extracted 26 Mel-frequency cepstral coefficients (MFCCs) and 4 speech prosodic features resulting in a total dimension of 30 ($A^{0:T} \in R^{30\times T}$). In terms of motion data, on each motion frame, 63 features representing a whole human body pose were selected ($P^{0:T} \in R^{63\times T}$). Finally, extracted speech and motion features were normalized by taking into consideration of their corresponding min-max values over the whole time sequence.

### B. Evaluation Metrics

The following metrics are established for validating the accuracy and the quality of generation actions based on the related literature [19], [4], [21]. In short, *Average Position Error* is used to to measure the differences between ground truth and the predicted motions while *Acceleration* and *Jerk* are implemented for validating the smoothness of the actions.

**Average Position Error (APE)** : *APE* measures the average distance between the predicted joint angles and the ground truth ones as given in Eq. 3, where $T$ denotes the time sequence of motion, $D$ is the total number of joints. The closer *APE* scores to 0, the more similar to the ground truth motions.

$$APE(P_{fo}^{0:T}, \hat{P}_{fo}^{0:T}) = \frac{1}{TD}\sum_{T}^{t=0}\sum_{D}^{d=0} ||P_{fo}^{t} - \hat{P}_{fo}^{t}||_2 \qquad (3)$$

**Acceleration and Jerk:** *Acceleration* is calculated based on the rate of change of joint velocity while *Jerk* is defined as the rate of change of *Acceleration*. The two metrics are commonly used for verifying the smoothness of motion; the lower values, the smoother motions are [22].

### C. Ablation Studies

The network was firstly trained on the training set of agreement scenarios as mentioned in IV-A. Specifically, the training data was fed to the proposed framework with a batch size of 1024. We implemented the Adam optimizer with a learning rate of 0.0001 and $\beta_1, \beta_2$ were set to 0.9 and 0.999, respectively. After 700 training epochs, we decayed the learning rate with a decay factor of 0.9 for every next 20 epochs. The weighting losses ($\alpha = 5, \gamma = 1$) of $L_G$ and $L_D$ were chosen empirically. The network was trained for 1000 epochs. In the first 50 warm-up epochs, the adversarial loss was not included in $L_G$. This training pipeline was repeated for the JESTKOD training set of disagreement scenarios.

In addition to the full model consisting of *Generator*, *Context Encoder*, and *Discriminator*, ablation experiments were conducted to verify the impact of individual model components on the generated actions. Table I summarizes the key components of 4 implemented models. We also implemented the *Speech to Gesture* network [19] that receives $A_{ob}^{0:T}$ as an input for modelling speech synthesis gestures $P_{fo}^{0:T}$. All models were trained on the JESTKOD training set of agreement and disagreement scenarios using the same training pipeline as mentioned above.

At the testing phase, the generated motions were validated using evaluation metrics established in IV-B. Table IIa and Table IIb summarize the results of implemented models. Although both *Speech to Gesture* and *w/o Context Encoder+Discriminator* are only constructed by generative sequence models that consists of speech encoder and motion decoder, *w/o Context Encoder+Discriminator* shows better performance than *Speech to Gesture* in terms of *Acceleration* and *Jerk*, in particular. A closer look at the *Speech to Gesture* approach [19], the model heavily relies on post-processing for temporally smoothing the generated motion. On the other hand, our motion decoder is designed with residual connections to foster the continuity of generated motions. As a result, *w/o Context Encoder+Discriminator* is able to produce co-speech motions with lower *Acceleration* and *Jerk* scores.

For modeling co-speech gestures of a communicator in dyadic interactions, in addition to their audio, social signals encoded from their interaction partner are essential modalities that should not be neglected. The results depicted in Table IIa and Table IIb indicate that contextual vectors provided by *Context Encoder* allow *w/o Discriminator* to generate speech-driven gestures closer to the ground truth motion. Apart from the fact that our designed *Discriminator* can verify the synthesis between $s_{fo}^{0:T}$ and $\hat{P}_{fo}^{0:T}$, the information encoded in $c_{ob}^{0:T}$ may allow $D$ further verify the synthesis between $c_{ob}^{0:T}$ and $\hat{P}_{fo}^{0:T}$. Those settings suggest that the adversarial loss provided by *Discriminator* can serve as a smart adaptive loss function for enhancing the generated motion $\hat{P}_{fo}^{0:T}$. The results presented in Table IIa and Table IIb demonstrate that *APE* score can be further reduced by employing *Discriminator* in the context-aware generative framework.

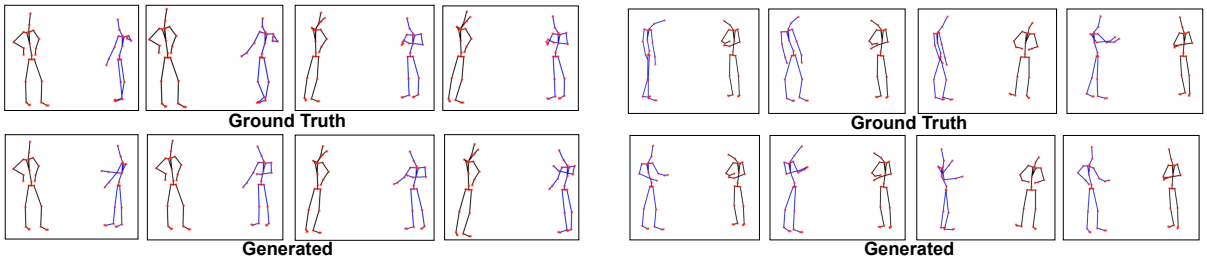
TABLE II: Performances of four implemented models in terms of *APE*, *Acceleration*, and *Jerk*. The results are reported on the testing set of: (a) agreement scenarios and (b) disagreement scenarios.

| Agreement Scenarios | | | | |
|---|---|---|---|---|
| No | Model | APE ($degree$) | Acceleration ($degree/s^2$) | Jerk ($degree/s^3$) |
| 1 | full model | $3.966 \pm 1.961$ | $5.064 \pm 0.870$ | $134.418 \pm 26.040$ |
| 2 | w/o Discriminator | $4.637 \pm 1.889$ | $45.573 \pm 9.692$ | $1014.541 \pm 198.861$ |
| 3 | w/o Context Encoder + Discriminator | $4.917 \pm 1.810$ | $145.680 \pm 38.366$ | $3999.423 \pm 995.027$ |
| 4 | Speech to Gesture [19] | $6.470 \pm 1.789$ | $201.008 \pm 41.942$ | $6769.455 \pm 1447.205$ |

(a) *agreement* scenarios

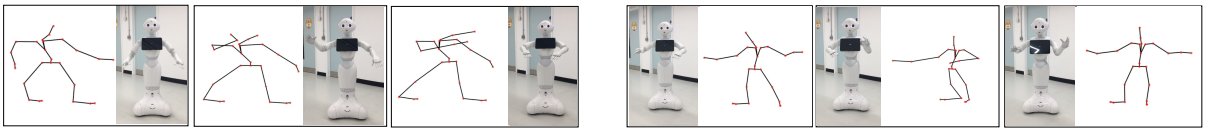
| Disagreement Scenarios | | | | |
|---|---|---|---|---|
| No | Model | APE ($degree$) | Acceleration ($degree/s^2$) | Jerk ($degree/s^3$) |
| 1 | full model | $3.891 \pm 2.207$ | $6.270 \pm 1.448$ | $170.298 \pm 41.463$ |
| 2 | w/o Discriminator | $5.280 \pm 1.850$ | $55.596 \pm 8.639$ | $1253.642 \pm 194.994$ |
| 3 | w/o Context Encoder + Discriminator | $5.752 \pm 2.253$ | $166.135 \pm 45.301$ | $4518.250 \pm 1197.977$ |
| 4 | Speech to Gesture [19] | $7.400 \pm 2.008$ | $240.050 \pm 42.073$ | $8061.153 \pm 1434.751$ |

(b) *disagreement* scenarios




(a) *agreement* scenario (b) *disagreement* scenario

Fig. 2: Sample generated body gestures (colored in blue) from: (a) *agreement* scenario and (b) *disagreement* scenario. The human skeleton colored in black represents for the body motion of the interacting partner $P_{ob}$.



(a) *agreement* scenario. (b) *disagreement* scenario.

Fig. 3: Transferring the generated motion of the target person $P_{fo}$ into the Pepper social robot. The human skeleton colored in black represents for the body motion of the interacting partner $P_{ob}$.

### *D. The Impact of Affective Context on Body Gestures in Dyadic Interaction*

Overall, the full model implemented in the agreement scenario (as shown in Table I) always showed better performance with respect to all metrics defined in IV-B as compared to the same network architecture employed in disagreement scenario. A closer look at the $APE$ scores in Table IIa and Table IIb, except for the *full model* where the difference of $APE$ values is smaller, the other three implemented models conducted in agreement scenarios were able to produce co-speech gestures $\hat{P}_{fo}^{0:T}$ more similar to the ground truth motions $P_{fo}^{0:T}$. Indeed, generated motions were smoother with respect to the lower *Acceleration* and *Jerk* obtained. The differences of APE values were even more obvious in the case of *Speech to Gesture* and *w/o Context Encoder + Discriminator* networks in which *Context Encoder* was not implemented. This result suggests that in affective conversations, it is more difficult to model co-speech gestures of the target person $P_{fo}$ since their speech feature $s_{fo}^t$ is not the only factor manipulating their body gesture $\hat{P}_{fo}^{0:T}$. In other words, the impact of interaction context on the prediction of co-speech gestures is unavoidable. Thus, *Context Encoder* should be employed for better modeling the dynamic exchange of social signals in dyadic interaction.

From interpersonal perspectives, there are several moderating variables (e.g., mimicry, synchrony, etc.) that have a high impact on the way human behave, in particular, their body gestures during social interactions [23], [24], [25]. For instance, the non-conscious behavioral mimicry can be detected when interlocutors have affiliative motivations during interaction [23] or the synchrony of movements in dyadic interactions is established between people who has pre-existing friendship [24]. Vice versa, it has been also shown that the synchrony of behaviors has been observed to decrease in situations in which the relationship between interlocutors are not well established [25]. The aforementioned studies provide an empirical evidence that the impact of moderating variables on the interlocutors' nonverbal behaviors is unavoidable in affective dyadic interactions. Specifically, taking into consideration agreement and disagreement scenarios presented in this work, the synchrony and mimicry of nonverbal signals between two interlocutors tend to decrease when they are involved in a controversial communication. Contrarily, when two partners share convergent opinions

for building a common ground, this process encourages the dynamic exchange of nonverbal signals between them during interaction. As a result, information encoded from our proposed *Context Encoder* can better contribute to the prediction of co-speech gestures.

Fig. 2a and Fig. 2b present examples of generated co-speech body gestures derived from the testing set of agreement and disagreement scenarios, respectively. Human motions represented by joint rotations were converted to into joint coordinates and presented in $3D$ space. It can be seen that in this dataset, interlocutors tend to use hand gestures to communicate their messages to their interaction partner, while the lower body remains relatively static. In particular, one of the frequently occurring cues was *"head tilting"* motions related to the disagreement scenario as illustrated in Fig. 2b. As also highlighted in [9], this is a common behaviour used to communicate a disagreement or confusion to the interaction partner in controversial conversations.

### *E. Transferring Human Gestures to Social Robots*

The generated motions in dyadic human-human interaction can be transferred into social robots, being robots' non-verbal gestures supporting for social human-robot interaction. As a proof of concept, we implemented the generated motion $\hat{P}_{fo}^{0:T}$ of the target person $P_{fo}$ on the Pepper robot. The process was started by converting $\hat{P}_{fo}^{0:T}$ into a set of $3D$ human joint coordinates. The motion $\hat{P}_{fo}^{0:T}$ defined in human motion space is then transferred into the Pepper robot's motion space and defined as a list of the robot's joint angles over time sequence. Fig. 3 presents generated actions collected in the testing set of agreement and disagreement interactions.

## V. CONCLUSION AND FUTURE WORK

In this paper, we propose an approach to generate body gestures from affective contextual cues in dyadic interactions. The framework is based on conditional Generative Adversarial Networks, which takes as an audio input of a target person together with the nonverbal signals of their interacting partner, modelled by a novel *Context Encoder*, to generate the body communication gestures of the target person. We evaluate our method against agreement and disagreement situations. The experimental results show that *Context Encoder* can better contribute to the prediction of co-speech gestures in agreement situations, implying the importance of context. As a proof concept, we demonstrated the idea of modeling body gestures with context awareness on the Pepper robot. There are several unexplored points that should be investigated in the future work, including a subjective evaluation of generated motions and transferring this research approach into scenarios of human-robot interaction. Finally, we aim to extend this research idea to work with nonverbal cues estimated by robots' off-the-shelf modules to enable the dynamic exchange of social signals between humans and robots in the real-world interaction applications.

## REFERENCES


[1] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: the behavior expression animation toolkit," in *Life-Like Characters*. Springer, 2004, pp. 163–185.

[2] H. Ahn, T. Ha, Y. Choi, H. Yoo, and S. Oh, "Text2action: Generative adversarial synthesis from language to action," in *ICRA*. IEEE, 2018, pp. 5915–5920.

[3] N. T. V. Tuyen, A. Elibol, and N. Y. Chong, "Conditional generative adversarial network for generating communicative robot gestures," in *RO-MAN*. IEEE, 2020, pp. 201–207.

[4] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *IVA*, 2019, pp. 97–104.

[5] Y. Huang and S. M. Khan, "Dyadgan: Generating facial expressions in dyadic interactions," in *CVPR Workshops*, 2017, pp. 11–18.

[6] W. Feng, A. Kannan, G. Gkioxari, and C. L. Zitnick, "Learn2smile: Learning non-verbal interaction through observation," in *IROS*. IEEE, 2017, pp. 4131–4138.

[7] H. Joo, T. Simon, M. Cikara, and Y. Sheikh, "Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction," in *CVPR*, 2019, pp. 10 873–10 883.

[8] C. Ahuja, S. Ma, L. Morency, and Y. Sheikh, "To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations," in *ICMI*. ACM, 2019, pp. 74–84.

[9] K. Bousmalis, M. Mehu, and M. Pantic, "Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools," in *ACII workshops*. IEEE, 2009, pp. 1–9.

[10] U. Bhattacharya, E. Childs, N. Rewkowski, and D. Manocha, *Speech2AffectiveGestures: Synthesizing Co-Speech Gestures with Generative Adversarial Affective Expression Learning*. New York, NY, USA: ACM, 2021, p. 2027–2036.

[11] K. Bousmalis, M. Mehu, and M. Pantic, "Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools," *Image Vision Comput.*, vol. 31, no. 2, p. 203–221, feb 2013.

[12] J. Ondras, O. Celiktutan, P. Bremner, and H. Gunes, "Audio-driven robot upper-body motion synthesis," *IEEE Transactions on Cybernetics*, vol. 51, no. 11, pp. 5445–5454, 2021.

[13] M. Marmpena, A. Lim, T. S. Dahl, and N. Hemion, "Generating robotic emotional body language with variational autoencoders," in *ACII*, 2019, pp. 545–551.

[14] T. Kucherenko, P. Jonell, S. van Waveren, G. E. Henter, S. Alexandersson, I. Leite, and H. Kjellström, "Gesticulator: A framework for semantically-aware speech-driven gesture generation," in *ICMI*. ACM, oct 2020.

[15] Y. Yoon, W. 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 *ICRA*. IEEE, 2019, pp. 4303–4309.

[16] U. Zabala, I. Rodriguez, J. M. Martínez-Otzeta, I. Irigoien, and E. Lazkano, "Which gesture generator performs better?" in *ICRA*, 2021, pp. 3345–3352.

[17] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE signal processing letters*, vol. 13, no. 1, pp. 52–55, 2005.

[18] M. Studdert-Kennedy, "Hand and mind: What gestures reveal about thought." *Language and Speech*, vol. 37, no. 2, pp. 203–209, 1994.

[19] D. Hasegawa, N. Kaneko, S. Shirakawa, H. Sakuta, and K. Sumi, "Evaluation of speech-to-gesture generation using bi-directional lstm network," in *IVA*, 2018, pp. 79–86.

[20] E. Bozkurt, H. Khaki, S. Keçeci, B. B. Türker, Y. Yemez, and E. Erzin, "The jestkod database: an affective multimodal database of dyadic interactions," *Language Resources and Evaluation*, vol. 51, no. 3, pp. 857–872, 2017.

[21] C. Ahuja and L.-P. Morency, "Language2pose: Natural language grounded pose forecasting," in *3DV*. IEEE, 2019, pp. 719–728.

[22] Y. Uno, M. Kawato, and R. Suzuki, "Formation and control of optimal trajectory in human multijoint arm movement," *Biological cybernetics*, vol. 61, no. 2, pp. 89–101, 1989.

[23] J. L. Lakin and T. L. Chartrand, "Using nonconscious behavioral mimicry to create affiliation and rapport," *Psychological science*, vol. 14, no. 4, pp. 334–339, 2003.

[24] K. Fujiwara, M. Kimura, and I. Daibo, "Rhythmic features of movement synchrony for bonding individuals in dyadic interaction," *Journal of Nonverbal Behavior*, vol. 44, no. 1, pp. 173–193, 2020.

[25] L. K. Miles, J. L. Griffiths, M. J. Richardson, and C. N. Macrae, "Too late to coordinate: Contextual influences on behavioral synchrony," *European Journal of Social Psychology*, vol. 40, no. 1, pp. 52–60, 2010.