Advanced Robotics

ISSN: 0169-1864 (Print) 1568-5535 (Online) Journal homepage: www.tandfonline.com/journals/tadr20

It takes two, not one: context-aware nonverbal behaviour generation in dyadic interactions

Nguyen Tan Viet Tuyen & Oya Celiktutan

To cite this article: Nguyen Tan Viet Tuyen & Oya Celiktutan (2023) It takes two, not one: context-aware nonverbal behaviour generation in dyadic interactions, Advanced Robotics, 37:24, 1552-1565, DOI: 10.1080/01691864.2023.2279595

To link to this article: https://doi.org/10.1080/01691864.2023.2279595

Published online: 02 Dec 2023.

ADVANCED ROBOTICS
2023, VOL. 37, NO. 24, 1552–1565
https://doi.org/10.1080/01691864.2023.2279595

FULL PAPER

OPEN ACCESS Check for updates

It takes two, not one: context-aware nonverbal behaviour generation in dyadic interactions

Nguyen Tan Viet Tuyen and Oya Celiktutan

Centre for Robotics Research, Department of Engineering, King's College London, London, UK

ABSTRACT

Nonverbal behaviours are integral parts of human social interaction. Equipping social robots with human 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 the current works only consider social signals of a single character to model co-speech gestures in non-interactive settings. To address this shortcoming, this paper introduces a context-aware Generative Adversarial Network, intending to produce social cues for robots. The approach captures both intra- and interpersonal social signals of two interlocutors to model body gestures in dyadic interaction. We conducted a series of experiments to validate the proposed solution under different interaction settings. First, the experimental results conducted in the JESTKOD dataset demonstrate the contribution of encoding context, namely the behaviours of the interaction partner, in the prediction of target person's gestures in agreement situations. Second, the experiments conducted in the new LISI-HHI dataset show that combining Discriminator and Context Encoder results in a gesture generation framework that is effective across various social communication contexts.

ARTICLE HISTORY
Received 15 January 2023
Revised 7 August 2023
Accepted 6 October 2023

KEYWORDS
Nonverbal behaviours; co-speech gesture; dyadic interaction; human-human interaction; human-robot interaction

1. Introduction

Social robots are envisioned to have a profound impact on many sectors, 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 tend to use a wide range of nonverbal cues to signal their emotions, intentions, or verbal contents of their speech to their interaction partners. Motivated by this, imitating nonverbal communication has been an active area of research to enhance the clarity of the human-robot interaction (HRI) interfaces and the sense of rapport, hence maximize the user trust and acceptance of them.

A considerable effort has gone into designing nonverbal interaction skills for social robots. For humanoid robot platforms, nonverbal cues are commonly inspired by human behaviours. One of the main reasons is to ensure communicative messages, encoded in robots' body movements, are interpretable by humans [1]. Previous work on nonverbal generation can be briefly categorized into two groups: (1) the rule-based approach and (2) the data driven-approach. Early methods have focussed on rule-based approaches [2, 3], requiring the design of interaction logic manually, which is limited, not transferable to unforeseen 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 [4–6] have been a promising solution to address these shortcomings. However, so far, only a handful of works [7–12] aim to model 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 formalized 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 consider the interaction partner's way of speaking and acting to be able to create socially suitable behaviours for robots.

CONTACT Nguyen Tan Viet Tuyen tan_viet_tuyen.nguyen@kcl.ac.uk; Oya Celiktutan oya.celiktutan@kcl.ac.uk

This paper introduces a context-aware Generative Adversarial Network (GAN) towards modelling robots' nonverbal behaviours in dyadic interactions. The approach takes speech features of a target person together with nonverbal signals of their interaction partner, modelled by a novel Context Encoder, to produce appropriate body gestures supporting for social interaction. We comprehensively validated the proposed framework on two datasets, namely JESTKOD and LISI-HHI, covering human dyadic interactions in affective contexts and social communication contexts, respectively. The main contributions of this paper are: (1) a novel co-speech gesture generation framework that captures both intra- and interpersonal social signals to model body gestures of the target person in dyadic interactions; (2) a series of experiments carried out in different scenarios to examine the impact of interaction context on generated cues; and (3) a newly created LISI-HHI dataset which aims to serve as a high-accuracy multimodal database for HRI community and related research domains. The experimental results conducted in the LISI-HHI dataset aims to serve as a benchmark of the context-aware nonverbal behaviour synthesis task.

The rest of this paper is organized as follows. In Section 2, we review previous studies on nonverbal behaviours generation inspired by data-driven approach. Section 3 describes the proposed end-to-end learning framework in detail. It is followed by a series of experiments conducted to verify the proposed network. We validate the model performance on an affective interaction dataset in Section 5 and a social communication database in Section 6. As a proof concept, we demonstrate the proposed framework on the Pepper humanoid robot in Section 7. Finally, the experimental results and future work are summarized in Section 8.

2. Data-driven nonverbal behaviour synthesis

2.1. Nonverbal behaviour synthesis from intrapersonal social signals

The data-driven approach provides a solution to transfer human nonverbal communication skills to robots in an end-to-end manner using large-scale datasets of human behaviours [13, 14]. The approach receives social signals (e.g, speech audio, speech text) of a target person to model their co-speech gestures conveying their emotions or intentions. Different learning frameworks have been introduced to capture the relationship between human audio [6, 15], speech text [4, 5, 16] and human co-speech gestures. The network architecture could be constructed in various ways, ranging from auto-regressive [17], encoder-decoder [6], Long Short Term Memory (LSTM) [15] to generative adversarial network (GAN) [16]. Although these approaches are promising solutions to address the shortcomings of the rule-base approach, they only consider social signals of a single character to model co-speech gestures in non-interactive settings.

2.2. Nonverbal behaviour synthesis from intra- and interpersonal social signals

In small-group social interaction, an essential aspect of communication is the dynamic exchange of nonverbal signals among interlocutors, with the aim of adapting to interacting social norms [18], for building or breaking a common ground [19–21]. This factor suggests that when modelling human or social robots' nonverbal behaviours in small-group interaction settings, particularly dyadic interaction, both intra- and interpersonal nonverbal signals should be taken into consideration. However, only a few studies [7–12] aim to generate behaviours by taking into consideration the interaction context, namely, the nonverbal signals of the interaction partner.

The problem of modelling human facial expressions in an interaction between an interviewee and an interviewer could be addressed by a conditional GAN framework [7] or a variational autoencoder (VAE) [8]. On the other hand, the idea of forecasting nonverbal cues was demonstrated by a residual attention network [10] to forecast human upper body motions or a GAN network [11] to predict interlocutors' upper body gestures and their facial landmarks. In the scenario of triadic interaction, the authors [9] introduced a generative framework that observes nonverbal signals of all interlocutors to forecasts nonverbal signals of a target person. However, none of these approaches has investigated the problem of co-speech gesture synthesis in dyadic interaction and, importantly, the effect of interaction contexts on generated actions. Motivating from that, our early work [12] introduced a context-aware co-speech gesture generation framework and verified the impact of affective context on synthesized gestures. In this paper, we further extended the work [12] by incorporating the early approach with a new loss function, a modified network architecture, and an updated audio feature extraction towards enhancing the model performance. In addition to the experiment conducted in affective interaction contexts, we further demonstrated the approach in social communication contexts using our newly created LISI-HHI dataset [22]. By demonstrating the idea on two different databases representing for two different settings, this paper aims to understand the impact of interaction contexts on the context-aware GAN approach comprehensively.

3. Context-aware generative adversarial network

3.1. 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 $P_{fo}^{0:T}$.

3.2. 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 Figure 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 combined into a contextual vector, namely c_{ob}^t. s_{fo}^t, c_{ob}^t, and together with the previously generated pose $\hat{P}_{fo}^{t-1}$ is injected into $G_{Encoder}$. 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 into D for identifying samples to be either fake or real. In the sequel, the proposed network architecture is described in detail.

3.3. Context encoder E

Context Encoder is designed to encode social signals simultaneously collected from the interaction 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.

3.3.1. Motion encoder E_M

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.

3.3.2. Speech encoder E_S

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 first extract the MFCCs and related low-level speech features. MFCCs are well known to encode signal frequencies according to how humans perceive sounds, and such low-level features are widely utilized in speech recognition or identification tasks [23]. In addition to MFCCs, the prosodic features representing the energy of speech are utilized as it encompasses intonation, rhythm, and other information about the speech outside of the specific words spoken (e.g. semantics and syntax). Speech prosody is a common candidate for modelling human beat gestures [24]. Similar to the *Motion Encoder*, *Speech Encoder* processes input speech features frame-by-frame. *Speech Encoder* is constructed with 4 Convolutional (CONV) layers, 1 LTSM layers, and 1 FC layer.

3.4. Generator G

Generator G consists of *Speech Encoder*, $G_{Encoder}$, and $G_{Decoder}$. *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 $G_{Encoder}$. At the initial time stamp ($t = 0$), a seed pose P_{fo}^{init} is injected into $G_{Encoder}$ instead of the previously generated pose $\hat{P}_{fo}^{t-1}$. $G_{Encoder}$ 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. As illustrated in Figure 1, for better modelling the velocity of generated motion, a residual connection is added between the previously generated pose and the new output pose produced by $G_{Decoder}$. 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.

Note that *Generator* can also be used independently without the need of integrating with *ContextEncoder* and *Discriminator*. In this case, G receives $A_{fo}^{0:T}$ to predict the co-speech gesture $\hat{P}_{fo}^{0:T}$. Further details are presented in Sections 5 and 6.

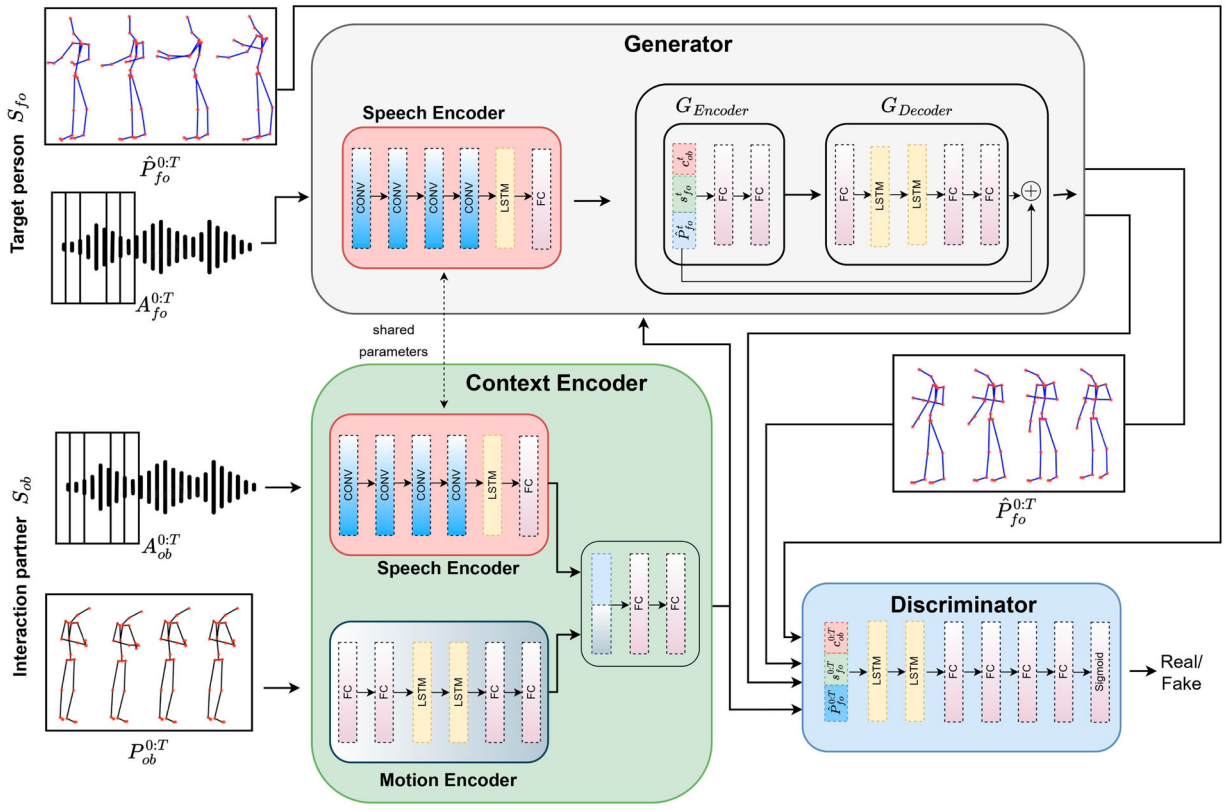

Figure 1. The proposed framework based on conditional GAN to generate body gestures for a target person taking into consideration the target person's speech (or audio) and their interaction partner's nonverbal signals encoded by the *Context Encoder*.

3.5. Discriminator D

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 Figure 1 is trained with the loss functions L_G and L_D defined in Equations (1) and (2), respectively. The training procedure is summarized in Algorithm 1. $\Delta P_{fo}^{1:T}$ and $\Delta \hat{P}_{fo}^{1:T}$ represents the velocity of ground truth motion $P_{fo}^{1:T}$ and the generated one $\hat{P}_{fo}^{1:T}$, respectively. α, β, and γ are weight parameters to manipulate the corresponding loss terms. Note that the newly implemented velocity loss can be considered as an improvement of the loss function L_G introduced in [5]. By incorporating velocity loss into the total loss L_G, along with adversarial loss and position loss, the new approach enhances the smoothness of generated motions.

$$\mathcal{L}_G = \alpha * ||P_{fo}^{0:T} - \hat{P}_{fo}^{0:T}||_2^2 + \beta * ||\Delta P_{fo}^{1:T} - \Delta \hat{P}_{fo}^{1:T}|| + \gamma * log\left(1 - D\left(c_{ob}^{0:T}, s_{fo}^{0:T}, \hat{P}_{fo}^{0:T}\right)\right) \quad (1)$$

$$\mathcal{L}_D = -log\left(D\left(c_{ob}^{0:T}, s_{fo}^{0:T}, P_{fo}^{0:T}\right)\right) - log\left(1 - D\left(c_{ob}^{0:T}, s_{fo}^{0:T}, \hat{P}_{fo}^{0:T}\right)\right) \quad (2)$$

4. Evaluation metrics

The following metrics are used to validate the accuracy and the quality of generation actions based on the related literature [6, 15, 25]. 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 assessing 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 Equation (3), where T

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**
    $c_A^t \leftarrow E_S(A_{ob}^t)$;
    $c_P^t \leftarrow E_M(A_{ob}^t)$;
    $c_{ob}^t \leftarrow \text{concat}(c_A^t, c_P^t)$
    $s_{fo}^t \leftarrow E_S(A_{fo}^t)$;
    $\hat{P}^t \leftarrow G(c_{ob}^t, s_{fo}^t, \hat{P}^{t-1})$
  end for
  $y_r \leftarrow D(c_{ob}^{0:T}, s_{fo}^{0:T}, P_{fo}^{0:T})$
  $y_f \leftarrow D(c_{ob}^{0:T}, s_{fo}^{0:T}, \hat{P}_{fo}^{0:T})$
  Update D with $\mathcal{L}_D$
  Update G, E_S, and E_M with $\mathcal{L}_G$
end for

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.

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

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 [26].

5. Experimental results in affective contexts

5.1. JESTKOD – a dataset of dyadic interactions in affective contexts

The proposed approach was validated on the JESTKOD dataset [27], a time-synchronized 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 a target person from speech while taking into consideration the contextual information simultaneously acquired from an interaction partner. The JESTKOD dataset covers a wide range of agreement and disagreement discussions on different topics (e.g. movies, sport, music, etc.) with 10 participants (4 females, 6 males). The dataset was collected in such a way that the participants' profiles were considered to put them into proper conversational topics to create agreement and 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.

Table 1. Low-level features extracted from audio input.

Feature	Dimension
Mel Frequency Cepstral Coefficients (MFCCs)	13
delta-MFCC (1st)	13
delta-MFCC (2nd)	13
Spectral (centroid, bandwidth, rolloff, poly features)	8
Energy	1
Total	48

5.2. Dataset preprocessing

We divided the dataset into training and testing sets. To better understand the contribution of affective contexts to generated motions, we trained and evaluated the approach on two separate interaction tasks, namely, agreement and disagreement scenarios. 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). On each interaction session, from the audio recordings, we extracted low-level features as illustrated in Table 1 with a total dimension of 48. In terms of motion data, on each motion frame, 63 features representing 21 joints of human body motion in *Roll*, *Pitch*, and *Yaw* were selected ($P^{0:T} \in \mathbb{R}^{63\times T}$). Speech features and motion features were normalized by taking into consideration their corresponding min-max values over the whole time sequence. Finally, data was split into a set of training instances using a time window $T = 6$ (secs) and a sliding window $\Delta T = 2$ (secs). On each motion instance, we stored as an initial pose P^{init} of the motion sequence $P^{0:T}$ and used it as a seed pose as discussed in Section 3.4.

5.3. Ablation studies

The network was firstly trained on the training set of agreement scenarios as mentioned in Section 5.2. The training data was fed to the network with a batch size of 1024. We use the Adam optimizer with a learning rate $\alpha = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate was decayed after completing the first 700 training epochs, it was then reduced with a decay factor 0.9 for every next 20 epochs. In the loss function L_G, we set $\alpha = 5$, $\beta = 5$,

Table 2. Key components of ablation models.

No	Model	Components			
			Context Encoder *E*		
		Generator *G*	E_M	E_S	Discriminator *D*
1	full model	✓	✓	✓	✓
2	Tuyen el al. [12]	✓	✓	✓	✓
3	without D	✓	✓	✓	none
4	without *E* and *D*	✓	none	none	none
5	Speech to Gesture [15]	✓	none	none	none

and $\gamma = 1$. All values of these parameters 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. Table 2 summarizes the key components of 5 implemented models: (1) the *full model* is composed of G, E_M, E_S, and D as introduced in Section 3. Compared to [12], the network architectures of E, G, and D of *full model* were improved by updating several hidden layers to better present output features. Audio inputs were described by a higher number of relevant low-level features as shown in Table 1. Indeed, L_G was incorporated with the velocity loss to better encourage the smoothness of generated motions. (2) the model is comprised of G, E_M, E_S, and D as introduced in [12]. (3) the approach *without D* was implemented by removing D out of the proposed framework. In other words, the adversarial loss was not contributed to the loss function L_G. (4) the model *without E and D* was designed by removing both D and E. (5) the *Speech to Gesture* network is introduced in [15], Similar to the *without E and D* framework, *Speech to Gesture* receives $A_{ob}^{0:T}$ as an input for modelling speech gestures $P_{fo}^{0:T}$. 5 models were trained on the JESTKOD training set of agreement and disagreement scenarios using the same training pipeline mentioned above.

5.4. The impact of affective context on body gestures in dyadic interaction

The results shown in Table 3 indicated that the *full model* and the network [5] demonstrate a similar performance in terms of *APE* scores in *Agreement* and *Disagreement* scenarios. However, motions produced by *full model* have lower *Acceleration* and *Jerk* values. The result can be interpreted taking into consideration the improved loss function L_G of *full model*, which aims to enhance the smoothness of generated actions.

A closer look at the *APE* scores reported in Table 3(a ,b), except for the *full model* and the approach [12] in which the difference in terms of *APE* values is negligible, other models implemented in the scenarios of *Agreement* always showed better performance with respect to all metrics defined in Section 4 as compared to the same network architecture employed in disagreement scenarios. In other words, 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 smaller *Acceleration* and *Jerk* values obtained. The differences of APE values were even more obvious in the case of *Speech to Gesture* and *without E and D* 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 modelling 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 affective interactions [20, 21, 28]. For instance, the non-conscious behavioural mimicry can be detected when interlocutors have affiliative motivations during interaction [28], or the synchrony of movements in dyadic interactions is established between people who has pre-existing friendship [21]. Vice versa, the synchrony of behaviours has been observed to decrease in situations in which the relationship between interlocutors is not well established [20]. The aforementioned studies provide empirical evidence that the impact of moderating variables on the interlocutors' nonverbal behaviours is unavoidable in affective dyadic interactions. Specifically, considering 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

Table 3. Performances of 5 implemented models in terms of *APE*, *Acceleration*, *Jerk*, and *FGD*.

(a) Agreement Scenarios				
No	Model	APE (degree)	Acceleration (degree/s^2)	Jerk (degree/s^3)
1	full model	3.852 ± 1.988	4.675 ± 0.716	121.367 ± 19.885
2	Tuyen el al. [12]	3.966 ± 1.961	5.064 ± 0.870	134.418 ± 26.040
3	without *D*	4.518 ± 1.878	40.829 ± 7.546	909.959 ± 158.248
4	without *E* and *D*	4.949 ± 1.818	147.026 ± 37.627	4037.629 ± 976.420
5	Speech to Gesture [15]	6.470 ± 1.789	201.008 ± 41.942	6769.455 ± 1447.205
(b) Disagreement Scenarios				
No	Model	APE (degree)	Acceleration (degree/s^2)	Jerk (degree/s^3)
1	full model	3.888 ± 2.220	3.304 ± 0.325	87.501 ± 10.198
2	Tuyen el al. [12]	3.891 ± 2.207	6.270 ± 1.448	170.298 ± 41.463
3	without *D*	5.363 ± 1.885	50.407 ± 7.298	1150.785 ± 169.108
4	without *E* and *D*	5.603 ± 2.285	163.651 ± 49.227	4445.693 ± 1281.226
5	Speech to Gesture [15]	7.400 ± 2.008	240.050 ± 42.073	8061.153 ± 1434.751

Note: The results are reported on the test set of: (a) agreement scenarios and (b) disagreement scenarios of the JESTKOD dataset.

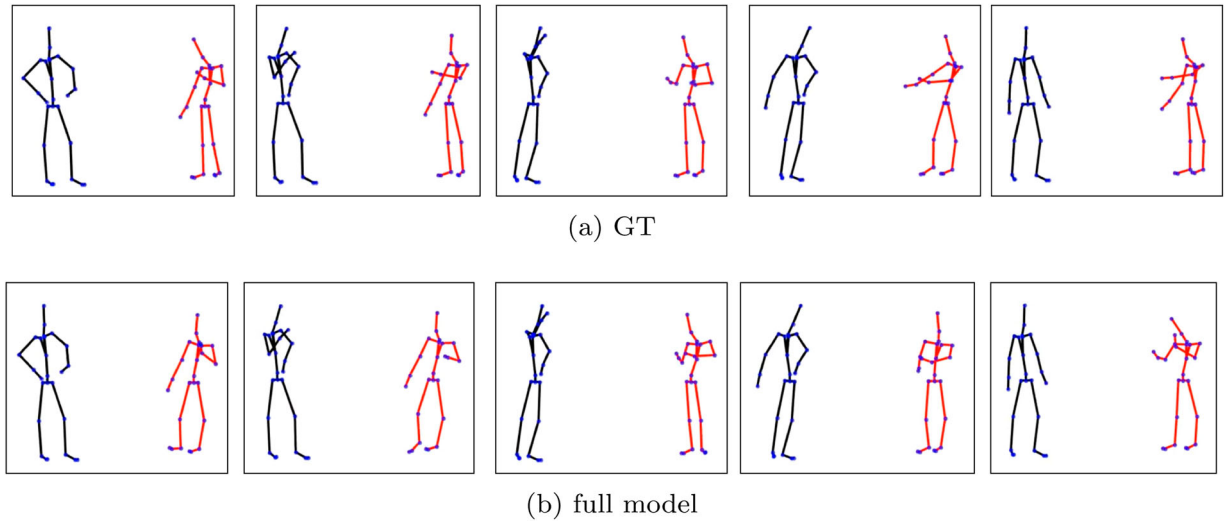

Figure 2. Sample generated body gestures (colored in red – right side) from the *agreement* scenario: (a) ground truth; (b) full model. The human skeleton coloured in black represents the body motion of the interaction partner P_{ob}.

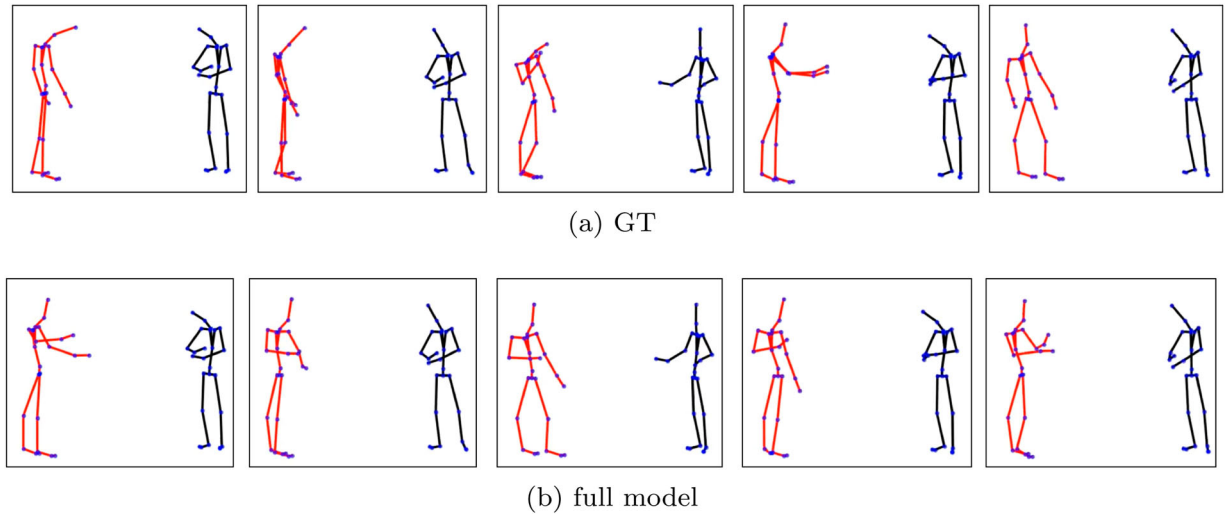

Figure 3. Sample generated body gestures (colored in red – left side) from the *disagreement* scenario: (a) ground truth; (b) full model. The human skeleton coloured in black represents the body motion of the interaction partner P_{ob}.

encourages the dynamic exchange of nonverbal signals during interactions. As a result, information encoded from our proposed *Context Encoder* can better contribute to the prediction of co-speech gestures.

Figures 2 and 3 present examples of generated co-speech body gestures derived from the test set of agreement and disagreement scenarios, respectively. Here, human motions represented by 3*D* angle rotations were converted into joint coordinates and presented in 3*D* coordinate 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'* motion related to the disagreement scenario as illustrated in Figure 3. As also highlighted in [29], this is a common behaviour used to communicate a disagreement or confusion to the interaction partner in controversial conversations.

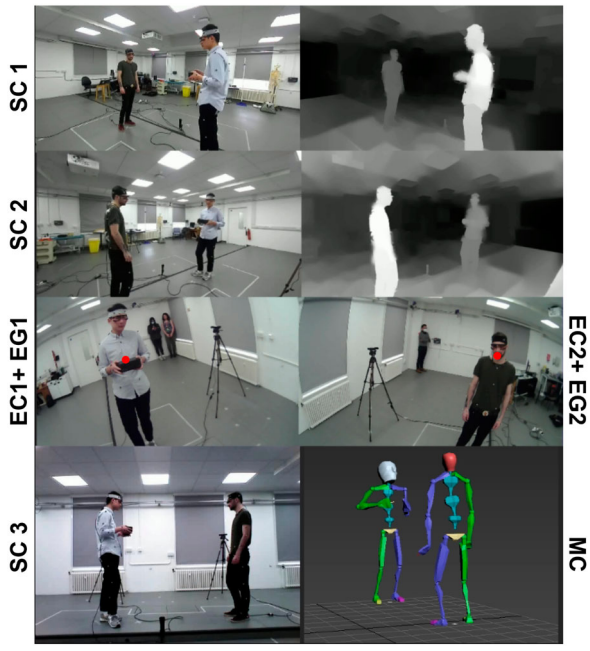

Figure 4. Multimodal data featuring in the LISI-HHI dataset. (**SC1:** Statistic RGBD 1, **SC2:** Statistic RGBD 2, **EC1:** Egocentric RGB 1, **EG1:** Eye-gaze of person P1, **EC2:** Egocentric RGB 2,**EG2:** Eye-gaze of person P2, **SC3:** Statistic RGBD 3, **MC:** motion data).

6. Experimental results in social communication contexts

6.1. LISI-HHI – a multimodal dataset of dyadic interactions in social communication contexts

The LISI-HHI (Learning to Imitate Social Interaction – Human-Human Interaction) dataset [22] consists of multimodal signals, including multiple RGBD views, eye gaze, audio, and motion data. Figure 4 illustrates an example of multimodal data collected from the dataset. The experiment was conducted in a motion capture room, where all sensors are synchronized together in the time domain. The LISI-HHI dataset complements the previous databases by incorporating a multi-sensory setup with a novel design of social communication scenarios. Without creating a dataset limited to a specific context, for instance, agree and disagree discussions [27], theatrical narratives [30], LISI-HHI covers a wider range of human daily communication scenarios, which are practical to transfer into social HRI. To the best of our knowledge, LISI-HHI is among the few available databases that cover a high number of modalities, camera views, participants, and social interaction sessions. Putting all together, LISI-HHI aims to serve as a high-accuracy and multimodal dataset for many different research domains, especially HRI.

With the aim of collecting a diverse set of verbal and nonverbal behaviours in different communication contexts, participants are not given any narrations, and no constraints are put into them regarding their way of speaking and acting. LISI-HHI comprises 5 designed scenarios, including: (1) small talk; (2) meal planning; (3) tangram game; (4) role playing; (5) way finding. The LISI-HHI dataset covers a total of 160 interaction sessions performed by 64 participants (38 females, 26 males). Each pair of participants were instructed to conduct 5 interaction sessions under 5 different scenarios mentioned above. The dataset is composed of a total of 8.3 hours.

In addition to the experiment conducted in affective contexts discussed in Section 5, the LISI-HHI dataset was utilized to verify the model performance in social communication contexts. Regarding motion data, body gestures are presented by a sequence of joint angles in the JESKOD, while they are defined as joint coordinates in LISI-HHI. Putting all together, conducting experiments with two datasets, recorded in two different dyadic contexts and defined by two different motion types, allowed us to validate the proposed approach comprehensively.

6.2. Dataset preprocessing

From the audio recording, we first extracted low-level audio features resulting in a total dimension of 48 ($A^{0:T} \in \mathbb{R}^{48\times T}$) as explained in Table 1. The audio vectors were then normalized based on their min-max values over the whole time sequence, similar to the preprocessing of the JESTKOD dataset. Concerning motion data, in the LISI-HHI dataset, human gestures are defined by 39 joints in 3*D* Cartesian space. To eliminate the differences in body size, we reconstructed human

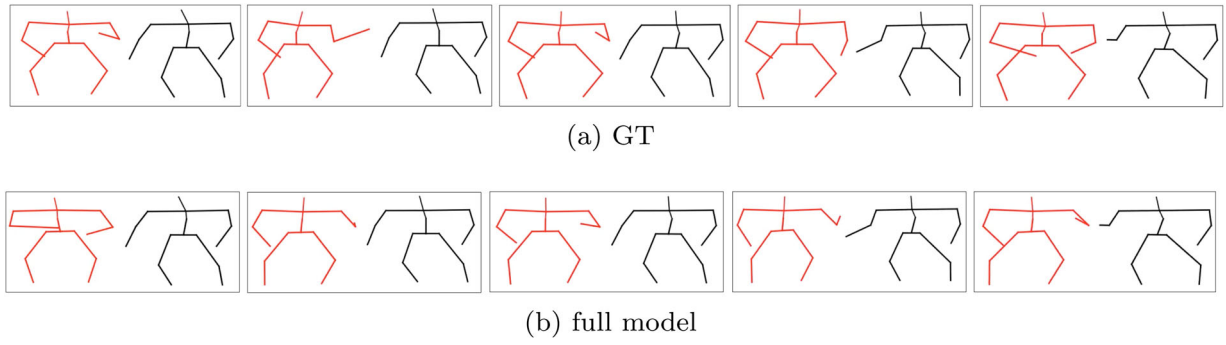

Figure 5. Sample generated body gestures (coloured in red – left side) from the scenario *Small Talk*.

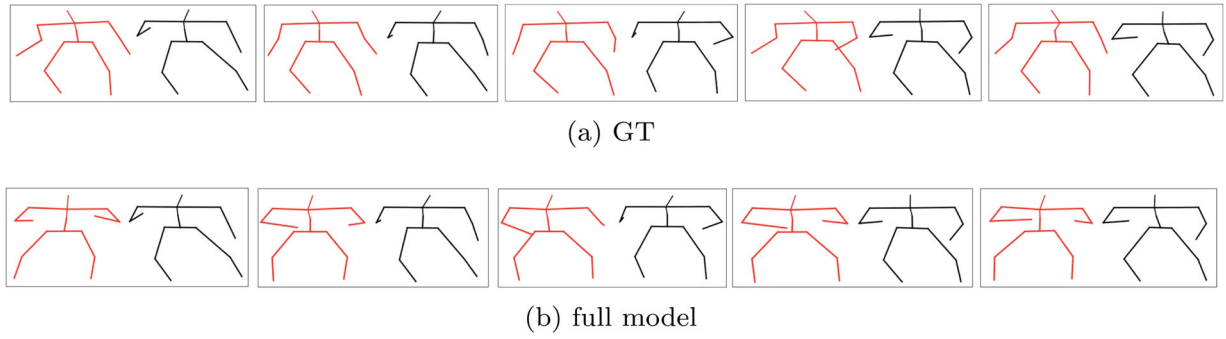

Figure 6. Sample generated body gestures (coloured in red – left side) from the scenario *Meal Planning*.

joints with respect to their top-chest joint coordinates. Motion values were then normalized, taking into consideration their min-max values over the whole time sequence. From 39 joints of the dataset, 30 main joints were selected out to represent a whole human body pose, which results in a 90-dimensional motion vector ($P^{0:T} \in \mathbb{R}^{90\times T}$).

The LISI-HHI dataset consists of 5 social communication scenarios, each of them is composed of 32 sessions. To better examine the effects of interaction context on generated motions, the network was trained and evaluated on individual scenarios. For each scenario, 25 sessions were used for training, and 7 sessions were utilized for testing.

6.3. The effects of social communication context on motion synthesis

An ablation study was conducted with 5 implemented frameworks. Their model components are illustrated in Table 2. We sequentially trained them on each of 5 scenarios. Due to a higher joint dimension $P^{0:T} \in \mathbb{R}^{90\times T}$, the training process was conducted with a batch size of 512. Apart from the batch size of 512, we used the same training parameters and strategies as introduced in Section 5.3. Table 4 summarizes the performance of 5 implemented frameworks across 5 different social communication scenarios. Figures 5–9 present examples of generated co-speech body gestures derived from the testing set of the LISI-HHI dataset.

The results depicted in Table 4 indicates that differences between *full model* and the approach introduced in [12] were not noticeable in term of *APE*. However, *Acceleration* and *Jerk* of generated motions produced by *full model* tend to be lower than the ones generated by the approach [12]. The differences in *Acceleration* and *Jerk* values could be explained by the differences in their loss functions L_G. In *full model*, the addition of velocity loss to L_G enhances the smoothness of generated motion, resulting lower values of *Acceleration* and *Jerk*.

Compared to *full model* and the network in [12], the performances of *without D*, *without E and D*, and *Speech to Gestures* were significantly reduced, even though *without D* demonstrated a comparable performance with *full model* and the network in [12] in some scenarios. It is interesting to observe that *APE* values significantly increased by removing *D* from the framework. Taking into consideration the setting of LISI-HHI dataset in which a wider range of body and hand gestures were exhibited to support communicators' messages during dyadic conversations, the adversarial loss from *D* could provide *G* informative feedback to better imitate the distribution of human communicative gestures, with a final goal of producing fake gestures $\hat{P}_{fo}^{0:T}$ as much similar as the real ones $P_{fo}^{0:T}$. In other words, despite the fact that

Table 4. Performances of 5 implemented models in terms of *APE*, *Acceleration*, *Jerk*, and *FGD*.

(a) Scenario 1 – Small talk

No	Model	APE (cm)	Acceleration (cm/s^2)	Jerk (cm/s^3)
1	full model	3.245 ± 1.514	95.374 ± 36.929	1073.883 ± 355.641
2	Tuyen el al. [12]	3.143 ± 1.543	97.767 ± 35.905	1108.112 ± 348.140
3	without D	4.723 ± 1.724	109.198 ± 29.484	1295.615 ± 330.364
4	without E and D	7.798 ± 3.570	138.821 ± 19.595	1768.489 ± 259.950
5	Speech to Gesture [15]	12.791 ± 8.518	146.659 ± 52.284	1884.253 ± 514.419

(b) Scenario 2 – Meal planning

No	Model	APE (cm)	Acceleration (cm/s^2)	Jerk (cm/s^3)
1	full model	3.215 ± 1.184	96.776 ± 29.192	1098.390 ± 285.767
2	Tuyen el al. [12]	3.292 ± 1.198	96.775 ± 28.778	1096.135 ± 283.586
3	without D	10.380 ± 7.269	168.508 ± 52.462	2459.331 ± 839.200
4	without E and D	12.043 ± 7.818	166.812 ± 16.783	2367.214 ± 231.778
5	Speech to Gesture [15]	13.935 ± 9.317	154.368 ± 35.170	1937.501 ± 625.290

(c) Scenario 3 – Tangram game

No	Model	APE (cm)	Acceleration (cm/s^2)	Jerk (cm/s^3)
1	full model	5.696 ± 3.134	102.334 ± 28.500	1166.565 ± 312.031
2	Tuyen el al. [12]	6.178 ± 2.338	155.633 ± 54.812	1805.334 ± 593.727
3	without D	10.366 ± 7.581	162.776 ± 68.921	2334.365 ± 923.377
4	without E and D	19.898 ± 10.007	153.555 ± 28.499	2015.589 ± 390.419
5	Speech to Gesture [15]	23.453 ± 12.215	220.623 ± 71.180	2805.664 ± 605.881

(d) Scenario 4 – Role playing

No	Model	APE (cm)	Acceleration (cm/s^2)	Jerk (cm/s^3)
1	full model	3.784 ± 3.227	91.299 ± 34.296	1015.391 ± 350.710
2	Tuyen el al. [12]	3.867 ± 3.190	92.718 ± 32.963	1036.304 ± 337.131
3	without D	7.052 ± 3.632	138.532 ± 36.820	1721.840 ± 403.241
4	without E and D	9.043 ± 3.408	139.858 ± 27.865	1848.502 ± 396.816
5	Speech to Gesture [15]	13.753 ± 2.612	142.529 ± 45.177	1754.926 ± 602.244

(f) Scenario 5 – Way finding

No	Model	APE (cm)	Acceleration (cm/s^2)	Jerk (cm/s^3)
1	full model	4.221 ± 2.336	146.281 ± 83.594	1629.757 ± 923.318
2	Tuyen el al. [12]	4.429 ± 2.280	146.348 ± 76.281	1645.545 ± 833.795
3	without D	10.869 ± 6.991	160.828 ± 37.050	2280.928 ± 550.136
4	without E and D	11.712 ± 6.131	222.550 ± 28.340	3080.263 ± 431.174
5	Speech to Gesture [15]	30.997 ± 11.224	200.174 ± 71.431	2487.976 ± 602.155

Note: The results are reported on the test set of: (a) Scenario 1, (b) Scenario 2, (c) Scenario 3, (d) Scenario 4, and (f) Scenario 5 of the LISI-HHI dataset.

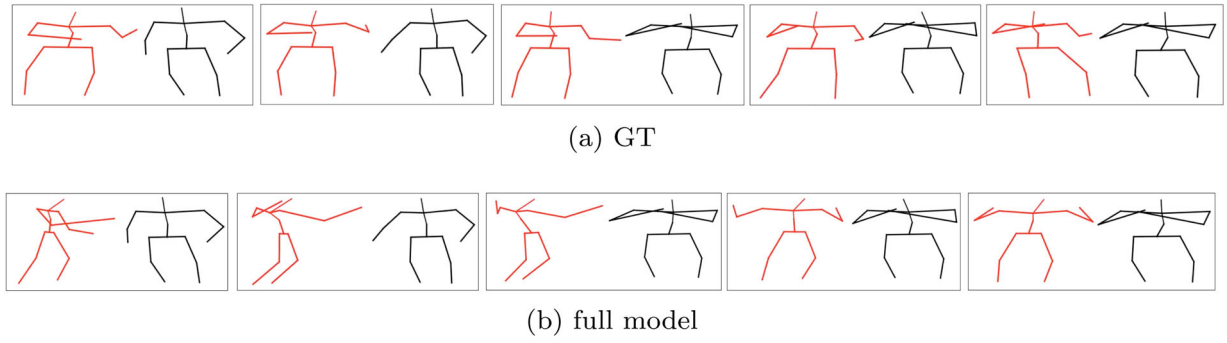

Figure 7. Sample generated body gestures (coloured in red – left side) from the scenario *Tangram game*.

Context Encoder E contributes to the prediction of generated action, *E* should be incorporated with *D* to form an efficient context-aware generative framework operating in such social communication contexts.

The performance of all implemented models exhibited variations across different scenarios. In particular, *APE*, *Acceleration*, and *Jerk* were higher in the scenarios of *Tangram game* and *Way finding* for all networks. A closer look at Figure 7 illustrating an example of *Tangram game*, a higher range of body movements can be seen as compared to the other scenarios. In *Tangram game*, iconic and metaphoric gestures are commonly utilized

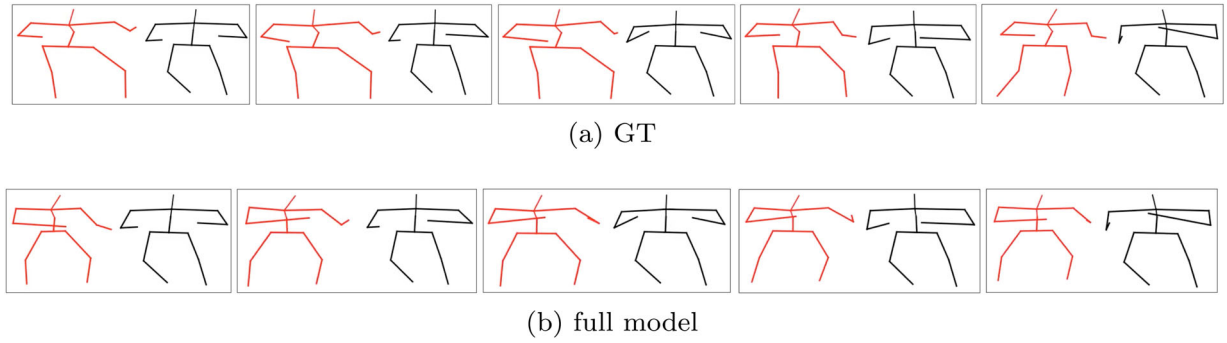

Figure 8. Sample generated body gestures (coloured in red – left side) from the scenario *Role playing*.

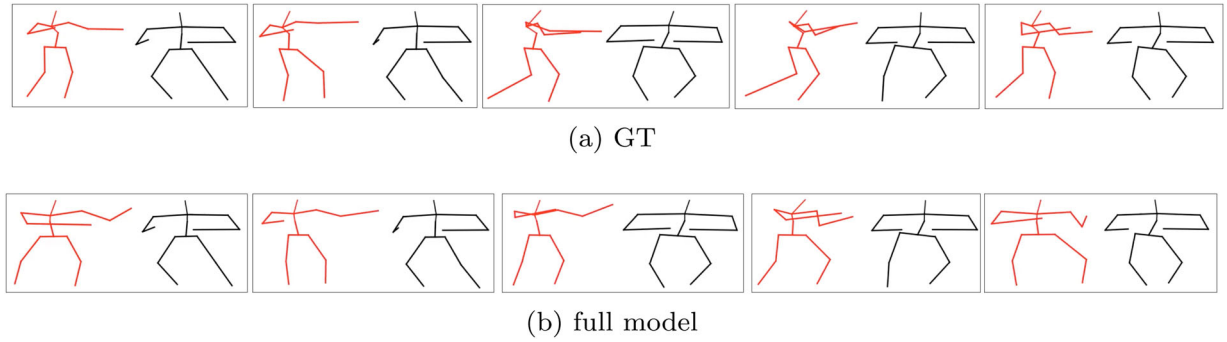

Figure 9. Sample generated body gestures (coloured in red – left side) from the scenario *Way finding*.

by participants to better explain semantic contents of their speech, for instance, shapes of tangram cards. Such nonverbal behaviours tend to produce a higher range of movements compared to beat gestures [31] as displayed in Figures 5 and 8. Consequently, a higher position error is predictable in that scenario. Similarly, in the example of *Way finding* shown in Figure 9, iconic and pointing gestures were commonly performed by participants for navigating and localizing purposes. This context encourages participants to perform more energetic hand movements resulting higher *Acceleration* and *Jerk* values. In general, the variations of motion accuracy across 5 scenarios highlighted the influence of communication contexts on the performance of co-speech generative networks. To some extent, the experimental results in Table 4 demonstrated that the combination of E and D as in the *full model* can mitigate the effect of interaction contexts on the accuracy of actions produced by G. It also suggests

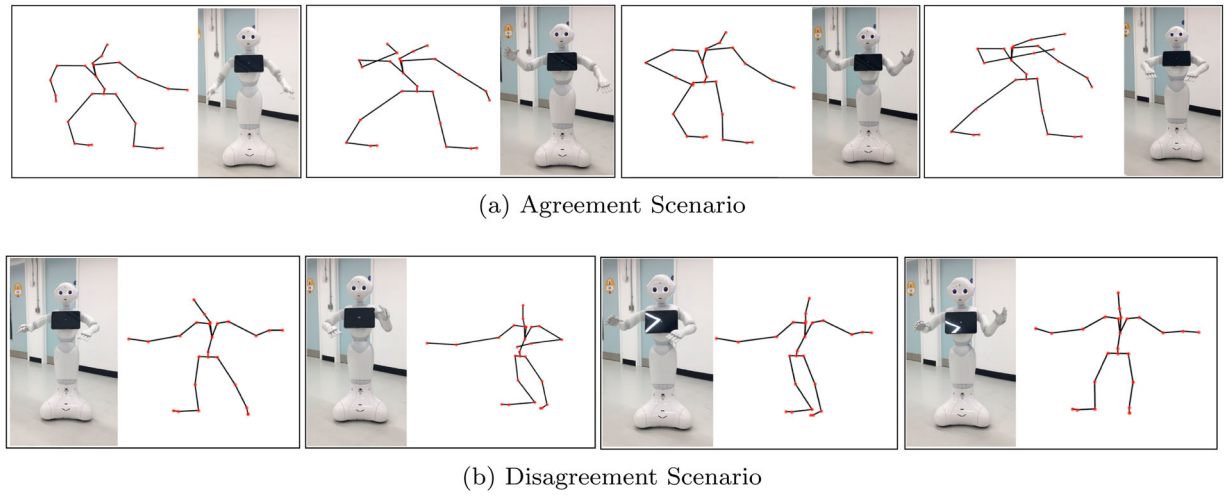

Figure 10. Transferring the generated motion of the target person P_{fo} into the Pepper social robot. The human skeleton coloured in black represents for the body motion of the interaction partner P_{ob}.

that discussions about the performance of co-speech gesture framework should consider the nature of interaction context settings.

7. Transferring human gestures to social robots

The generated motions in dyadic human-human interaction can be transferred into social robots, being robots' nonverbal gestures supporting 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} in affective contexts 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 was then transferred into the Pepper robot's motion space using the transformation model introduced in [32]. Consequently, the robot's motion is presented by a list of the robot's joint angles over the time sequence. Figure 10 presents generated actions collected in the test set of agreement and disagreement interactions.

8. Conclusion

This paper introduces a context-aware GAN approach towards modelling robots' nonverbal behaviours in dyadic interactions. The framework consists of *Context Encoder*, *Generator*, and *Discriminator*. The approach receives speech features of a target person together with nonverbal signals of their interaction partner, modelled by a novel Context Encoder, to generate appropriate co-speech gestures supporting for dyadic interaction. A series of experiments were conducted to validate the proposed framework comprehensively. We first evaluated our method against agreement and disagreement situations using the JESTKOD dataset. The experimental results show that *Context Encoder* can better contribute to the prediction of co-speech gestures in agreement situations, implying the importance of interaction context. To verify model performance in social communication settings, we conducted an experiment using our new LISI-HHI dataset. The experimental results confirm the contribution of *Context Encoder* to the accuracy of generated gestures. The results also highlight the combination of *Discriminator* and *Context Encoder* to form an efficient co-speech generation network that can work across different social communication settings. As a proof concept, we demonstrated the idea of modelling body gestures with context awareness on the Pepper robot.

In small group interaction, especially dyadic interaction, an essential aspect of communication is the dynamic exchange of nonverbal signals among interlocutors. The interaction context could influence interlocutors' way of speaking and acting, either for adapting to interaction social norms, building or breaking a common ground. Consequently, this social factor should be considered when modelling nonverbal signals in dyadic interactions, particularly for generating appropriate robots' gestures in social HRI settings.

From human behaviour studies, interpersonal coordination in dyadic interactions can be observed either in the same time window or with several seconds lag [33–35]. In other words, the contribution of interaction context to generated gestures should be investigated not only in the nonverbal behaviour synthesis task, as discussed in this paper, but also in the forecasting task. Hence, potential avenues for future research include demonstrating gesture synthesis and gesture forecasting within HRI scenarios, as well as investigating their effectiveness using both objective and subjective evaluation techniques.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was 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 (EPSRC) [grant ref.: EP/V010875/1].

Notes on contributors

Nguyen Tan Viet Tuyen is a Research Associate in Centre for Robotics Research, Department of Engineering, King's College London, United Kingdom. He received the M.Sc degree and Ph.D. degree in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in 2018 and 2021, respectively. His research interests include social and cognitive robotics, human-robot interaction, machine learning, and mechatronics.

Oya Celiktutan received her Ph.D. degree in electrical and electronic engineering from Bogazici University, Turkey, in 2013. In 2018, she joined King's College London, United Kingdom, where she is an Associate Professor in the Centre for Robotics Research, Department of Engineering and is the Head of the Social AI & Robotics Laboratory. Her primary research interest is machine learning applied to computer vision, human behaviour understanding and generation, and human-robot interaction.

ORCID

Nguyen Tan Viet Tuyen http://orcid.org/0000-0001-8000-6485
Oya Celiktutan http://orcid.org/0000-0002-7213-6359

References

[1] Saunderson S, Nejat G. How robots influence humans: a survey of nonverbal communication in social human–robot interaction. Int J Soc Robot. 2019;11(4):575–608. doi: 10.1007/s12369-019-00523-0

[2] Cassell J, Vilhjálmsson HH, Bickmore T. Beat: the behavior expression animation toolkit. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques; 2001. p. 477–486.

[3] Marsella S, Xu Y, Lhommet M, et al. Virtual character performance from speech. In: Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation; 2013. p. 25–35.

[4] Ahn H, Ha T, Choi Y, et al. Text2action: generative adversarial synthesis from language to action. In: ICRA; IEEE; 2018. p. 5915–5920.

[5] Tuyen NTV, Elibol A, Chong NY. Conditional generative adversarial network for generating communicative robot gestures. In: RO-MAN; IEEE; 2020. p. 201–207.

[6] Kucherenko T, Hasegawa D, Henter GE, et al. Analyzing input and output representations for speech-driven gesture generation. In: IVA; 2019. p. 97–104.

[7] Huang Y, Khan SM. Dyadgan: generating facial expressions in dyadic interactions. In: CVPR Workshops; 2017. p. 11–18.

[8] Feng W, Kannan A, Gkioxari G, et al. Learn2smile: learning non-verbal interaction through observation. In: IROS; IEEE; 2017. p. 4131–4138.

[9] Joo H, Simon T, Cikara M, et al. Towards social artificial intelligence: nonverbal social signal prediction in a triadic interaction. In: CVPR; 2019. p. 10873–10883.

[10] Ahuja C, Ma S, Morency L, et al. To react or not to react: end-to-end visual pose forecasting for personalized avatar during dyadic conversations. In: ICMI; ACM; 2019. p. 74–84.

[11] Tuyen NTV, Celiktutan O. Context-aware human behaviour forecasting in dyadic interactions. In: Understanding social behavior in dyadic and small group interactions; PMLR; 2022. p. 88–106.

[12] Tuyen NTV, Celiktutan O. Agree or disagreef generating body gestures from affective contextual cues during dyadic interactions. In: 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN); IEEE; 2022. p. 1542–1547.

[13] Xu J, Mei T, Yao T, et al. Msr-vtt: a large video description dataset for bridging video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 5288–5296.

[14] Yoon Y, Ko WR, Jang M, et al. Robots learn social skills: end-to-end learning of co-speech gesture generation for humanoid robots. In: 2019 International Conference on Robotics and Automation (ICRA); IEEE; 2019. p. 4303–4309.

[15] Hasegawa D, Kaneko N, Shirakawa S, et al. Evaluation of speech-to-gesture generation using bi-directional lstm network. In: IVA; 2018. p. 79–86.

[16] Tuyen NTV, Elibol A, Chong NY. A gan-based approach to communicative gesture generation for social robots. In: 2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO); IEEE; 2021. p. 58–64.

[17] Kucherenko T, Jonell P, van Waveren S, et al. Gesticulator: a framework for semantically-aware speech-driven gesture generation. In: Proceedings of the 2020 International Conference on Multimodal Interaction; 2020. p. 242–250.

[18] Lakin JL, Jefferis VE, Cheng CM, et al. The chameleon effect as social glue: evidence for the evolutionary significance of nonconscious mimicry. J Nonverbal Behav. 2003;27(3):145–162. doi: 10.1023/A:1025389814290

[19] Noy L, Dekel E, Alon U. The mirror game as a paradigm for studying the dynamics of two people improvising motion together. Proc Nati Acad Sci. 2011;108(52):20947–20952. doi: 10.1073/pnas.1108155108

[20] Miles LK, Griffiths JL, Richardson MJ, et al. Too late to coordinate: contextual influences on behavioral synchrony. Eur J Soc Psychol. 2010;40(7):1200–1211. doi: 10.1002/ejsp.v40:7

[21] Fujiwara K, Kimura M, Daibo I. Rhythmic features of movement synchrony for bonding individuals in dyadic interaction. J Nonverbal Behav. 2020;44(1):173–193. doi: 10.1007/s10919-019-00315-0

[22] Tuyen NTV, Georgescu AL, Di Giulio I, et al. A multimodal dataset for robot learning to imitate social human-human interaction. In: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction; 2023. p. 238–242.

[23] Murty KSR, Yegnanarayana B. Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Process Lett. 2005;13(1):52–55. doi: 10.1109/LSP.2005.860538

[24] Studdert-Kennedy M. Hand and mind: what gestures reveal about thought. Lang Speech. 1994;37(2):203–209. doi: 10.1177/002383099403700208

[25] Ahuja C, Morency LP. Language2pose: natural language grounded pose forecasting. In: 3DV; IEEE; 2019. p. 719–728.

[26] Uno Y, Kawato M, Suzuki R. Formation and control of optimal trajectory in human multijoint arm movement. Biol Cybern. 1989;61(2):89–101. doi: 10.1007/BF00204593

[27] Bozkurt E, Khaki H, Keçeci S, et al. The jestkod database: an affective multimodal database of dyadic interactions. Lang Resour Eval. 2017;51(3):857–872. doi: 10.1007/s10579-016-9377-0

[28] Lakin JL, Chartrand TL. Using nonconscious behavioral mimicry to create affiliation and rapport. Psychol Sci. 2003;14(4):334–339. doi: 10.1111/1467-9280.14481

[29] Bousmalis K, Mehu M, Pantic M. Spotting agreement and disagreement: a survey of nonverbal audiovisual cues and tools. In: ACII workshops; IEEE; 2009. p. 1–9.

[30] Metallinou A, Yang Z, Lee C, et al. The usc creativeit database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations. Lang Resour Eval. 2016;50(3):497–521. doi: 10.1007/s10579-015-9300-0

[31] McNeill D. Hand and mind: what gestures reveal about thought. University of Chicago Press; 1992.

[32] Tuyen NTV, Jeong S, Chong NY. Emotional bodily expressions for culturally competent robots through long term human-robot interaction. In: 2018 IEEE/RSJ

International Conference on Intelligent Robots and Systems (IROS); IEEE; 2018. p. 2008–2013.
[33] Heyes C. Automatic imitation. Psychol Bull. 2011;137(3): 463–483. doi: 10.1037/a0022288
[34] Leander NP, Chartrand TL, Bargh JA. You give me the chills: embodied reactions to inappropriate amounts of behavioral mimicry. Psychol Sci. 2012;23(7):772–779. doi: 10.1177/0956797611434535
[35] Hale J, Ward JA, Buccheri F, et al. Are you on my wavelength? Interpersonal coordination in dyadic conversations. J Nonverbal Behav. 2020;44:63–83. doi: 10.1007/s10919-019-00320-3