Context-Aware Body Gesture Generation for Social Robots

Nguyen Tan Viet Tuyen and Oya Celiktutan

Abstract— **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. However, the majority of these approaches considers 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 preliminary results show that taking into account the context contributes significantly to the accuracy in disagreement scenarios.**

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 [1]. 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 [2], 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 [3], [4], [5], [6] has been a promising solution to address these shortcomings. However, so far only a handful of works [7], [8], [9], [10], [11] 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 [12].

II. RELATED WORK

In recent years, there has been a growing interest in gesture generation using data-driven, end-to-end approaches to determine the relationships between non-verbal signals of a communicator (or a robot) and their speech [13]. In particular, co-speech 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*. [3] proposed a Sequence to Sequence (Seq2Seq) model based on a GAN. The other works [4], [6] developed methods based on conditional GANs: Tuyen *et al*. [4] designed a conditional GAN with Convolution Neural Network (CNN) operations. Wu *et al*. [6] proposed to use conditional GAN and unrolled GAN jointly. Bhattacharya *et al*. [14] used GANs to synthesise co-speech gestures with affective expressions.

The aforementioned approaches have focused on the communicator in isolation, without considering an interaction context. Huang and Khan [7] 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*. [8] suggested a VAE network to handle the generation of facial cues between a user and an embodied agent. There is only a handful of

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).

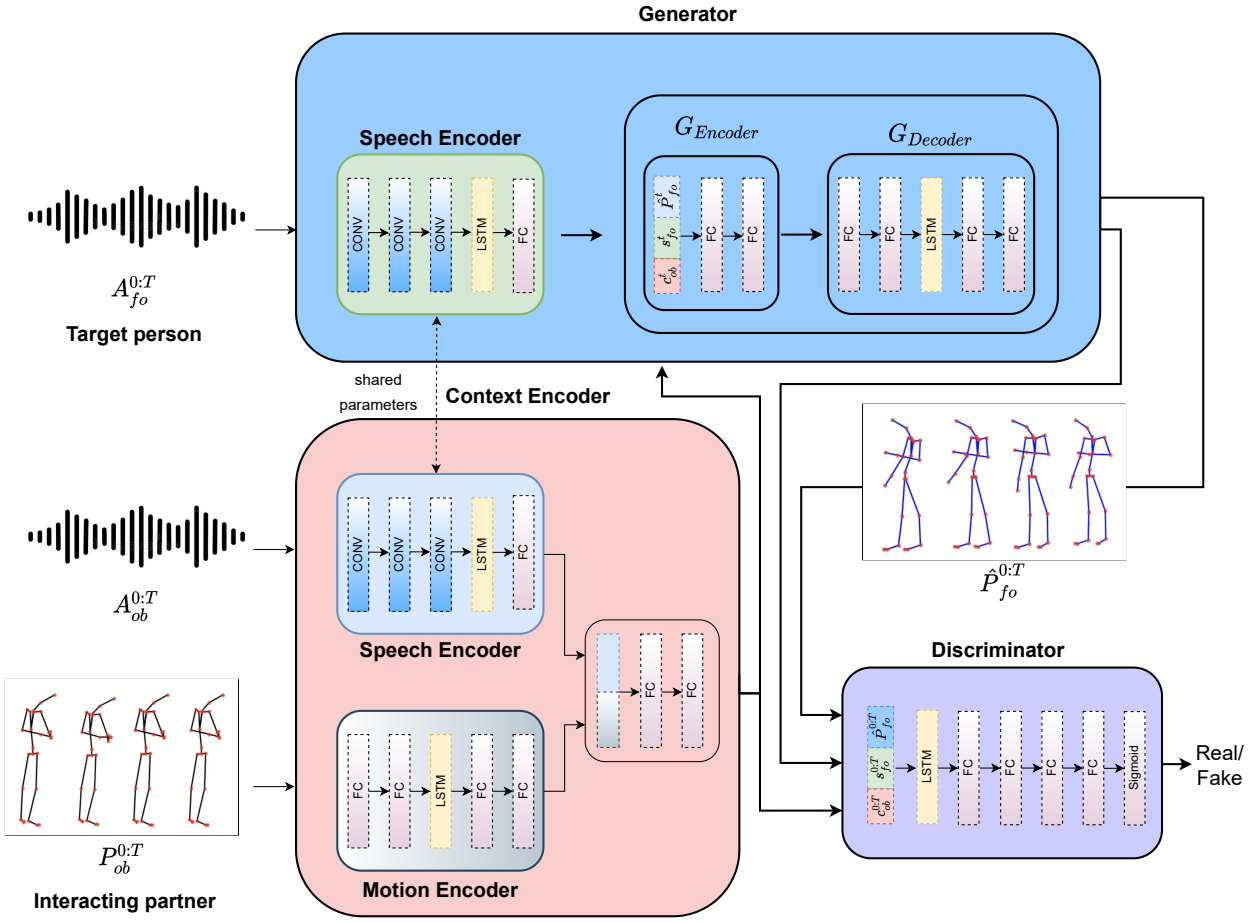


Fig. 1: The proposed framework based on conditional GAN to generate body gestures for a target person from their speech (or audio) and affective contextual cues, namely, their interaction partner's nonverbal signals encoded by the *Context Encoder*.

approaches aiming to generate body gestures during dyadic interactions [9]. In terms of triadic human communication, Joo *et al*. [10] presented a generative approach that acquires non-verbal signals from interacting partners and encodes them into latent vectors. Ahuja *et al*. [11] 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 (e.g., agreement versus disagreement) on the generated cues.

III. PROPOSED APPROACH

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

To solve this research problem, we develop a co-speech gesture generative framework with context awareness based on the conditional GAN [4]. Fig. 1 illustrates the proposed framework which consists of *Context Encoder E*, *Generator G*, and *Discriminator D*. *Context Encoder* is designed to encode social signals simultaneously collected from the interacting partner in dyadic interaction into a contextual vector and comprises of *Motion Encoder* and *Speech Encoder*.

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

TABLE I: Accuracy in terms of *APE*, *Acceleration*, and *Jerk* with respect to the two interaction scenarios, namely, agreement and disagreement.

Scenario	Model	APE ($degree$)	Acceleration ($degree/s^2$)	Jerk ($degree/s^3$)
Agreement	Full Model	3.966 ± 1.961	5.064 ± 0.870	134.418 ± 26.040
Agreement	w/o Context Encoder + Discriminator	4.917 ± 1.810	145.680 ± 38.366	3999.423 ± 995.027
Disagreement	Full Model	3.891 ± 2.207	6.270 ± 1.448	170.298 ± 41.463
Disagreement	w/o Context Encoder + Discriminator	5.752 ± 2.253	166.135 ± 45.301	4518.250 ± 1197.977

IV. EXPERIMENTAL RESULTS

A. Dataset

The proposed approach was validated on the JESTKOD dataset [15], 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 agreement/disagreement situations. It consists of 56 dyadic interactions in agreement and 42 sessions in disagreement with a total duration of 154 and 105 minutes, respectively.

B. Evaluation Metrics

We implement the commonly used metrics in the literature [16], [5], [17] to validate the accuracy and the quality of generation actions. *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. 1, 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.

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

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 [18].

C. Ablation Study

To analyse the performance with respect to the interaction context, we trained separate networks for two scenarios, namely, agreement and disagreement using the training set from the JESTKOD dataset. The models were then evaluated on the testing set using the aforementioned evaluation metrics. Table I presents the results of implemented models. Looking at the results, we observe two trends. Our preliminary results show that *Context Encoder* improves the performance across both scenarios in terms of all evaluation metrics, validating the importance of taking into account the nonverbal signals of the interaction partner. In both implementations, either with *Context Encoder* or without, models perform slightly better under the disagreement condition as compared to the agreement condition. However, the contribution of the *Context Encoder* is more evident in the case of disagreement scenario. Fig. 2b and Fig. 3b show examples of human motions produced by the fully implemented model compared to the ground truth.

V. CONCLUSION

In this paper, we proposed an approach to generate body gestures in dyadic interactions. Our method is based on conditional Generative Adversarial Networks, which takes the 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. Our preliminary results show that *Context Encoder* can better contribute to the prediction of co-speech gestures in disagreement situations, implying the importance of interaction context. As a future work, we will demonstrate the idea of modeling body gestures with context awereness on a humanoid robot. Our current method relies on motion capture data, which is not feasible for real-world applications. Our ultimate goal is to extend this research idea to work with nonverbal cues estimated by robots' off-the-shelf modules to enable the real-time exchange of social signals in human-robot interaction.

REFERENCES

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

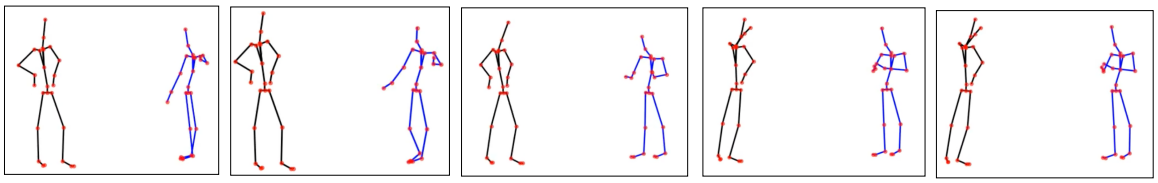
[2] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: the behavior expression animation toolkit," in *Life-Like Characters*. Springer, 2004, pp. 163–185.

[3] H. Ahn, T. Ha, Y. Choi, H. Yoo, and S. Oh, "Text2action: Generative adversarial synthesis from language to action," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 5915–5920.

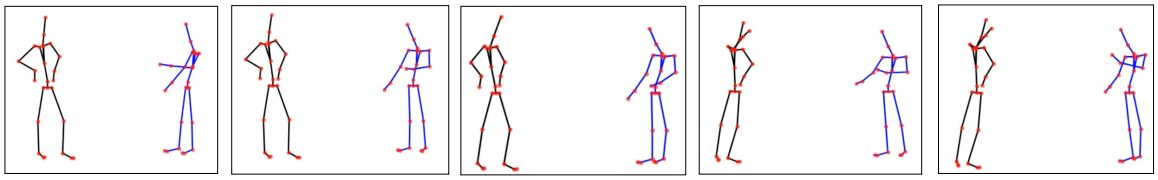
[4] N. T. V. Tuyen, A. Elibol, and N. Y. Chong, "Conditional generative adversarial network for generating communicative robot gestures," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 201–207.

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

[6] 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.

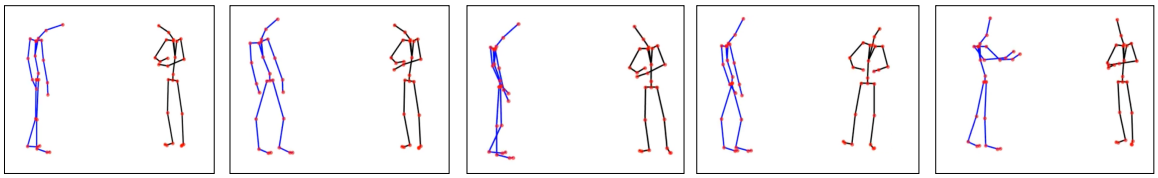


(a) GT

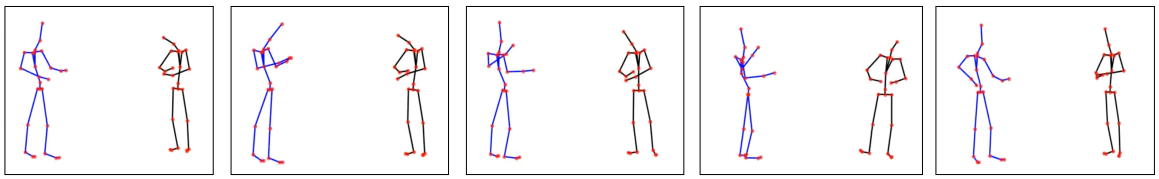


(b) Full model

Fig. 2: Sample generated body gestures (colored in blue - right side) from the *agreement* scenario: (a) ground truth; (b) full model. The human skeleton colored in black represents for the body motion of the interacting partner P_{ob}.



(a) GT



(b) Full model

Fig. 3: Sample generated body gestures (colored in blue - left side) from the *disagreement* scenario: (a) ground truth; (b) full model. The human skeleton colored in black represents for the body motion of the interacting partner P_{ob}.

[7] Y. Huang and S. M. Khan, "Dyadgan: Generating facial expressions in dyadic interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 11–18.

[8] W. Feng, A. Kannan, G. Gkioxari, and C. L. Zitnick, "Learn2smile: Learning non-verbal interaction through observation," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 4131–4138.

[9] N. T. V. Tuyen and O. Celiktutan, "Context-aware human behaviour forecasting in dyadic interactions," in *Understanding Social Behavior in Dyadic and Small Group Interactions*. PMLR, 2022, pp. 88–106.

[10] H. Joo, T. Simon, M. Cikara, and Y. Sheikh, "Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 873–10 883.

[11] 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 *International Conference on Multimodal Interaction, ICMI 2019, Suzhou, China, October 14-18, 2019*, W. Gao, H. M. Meng, M. A. Turk, S. R. Fussell, B. W. Schuller, Y. Song, and K. Yu, Eds. ACM, 2019, pp. 74–84. [Online]. Available: https://doi.org/10.1145/3340555.3353725

[12] K. Bousmalis, M. Mehu, and M. Pantic, "Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools," in *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, 2009, pp. 1–9.

[13] Y. Liu, G. Mohammadi, Y. Song, and W. Johal, "Speech-based gesture generation for robots and embodied agents: A scoping review," in *Proceedings of the 9th International Conference on Human-Agent Interaction*, ser. HAI '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 31–38. [Online]. Available: https://doi.org/10.1145/3472307.3484167

[14] U. Bhattacharya, E. Childs, N. Rewkowski, and D. Manocha, *Speech2AffectiveGestures: Synthesizing Co-Speech Gestures with Generative Adversarial Affective Expression Learning*. New York, NY, USA: Association for Computing Machinery, 2021, p. 2027–2036. [Online]. Available: https://doi.org/10.1145/3474085.3475223

[15] 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.

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

[17] C. Ahuja and L.-P. Morency, "Language2pose: Natural language grounded pose forecasting," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 719–728.

[18] 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.