2D or not 2D: How Does the Dimensionality of Gesture Representation Affect 3D Co-Speech Gesture Generation?

Téo Guichoux, Laure Soulier, Catherine Pelachaud, Nicolas Obin

2D or not 2D: How Does the Dimensionality of Gesture Representation Affect 3D Co-Speech Gesture Generation?

Téo Guichoux[1,2], Laure Soulier[1], Nicolas Obin[2], Catherine Pelachaud[1,3]
[1] Sorbonne Université, ISIR, F-75005 Paris, France ; [2] Sorbonne Université, IRCAM, Stms Lab, F-75003, Paris France ; [3] CNRS
{teo.guichoux, laure.soulier, catherine.pelachaud}@isir.upmc.fr

ABSTRACT

Co-speech gestures are fundamental for communication. The advent of recent deep learning techniques has facilitated the creation of lifelike, synchronous co-speech gestures for Embodied Conversational Agents. "In-the-wild" datasets, aggregating video content from platforms like YouTube via human pose detection technologies, provide a feasible solution by offering 2D skeletal sequences aligned with speech. Concurrent developments in lifting models enable the conversion of these 2D sequences into 3D gesture databases. However, it is important to note that the 3D poses estimated from the 2D extracted poses are, in essence, approximations of the ground-truth, which remains in the 2D domain. This distinction raises questions about the impact of gesture representation dimensionality on the quality of generated motions. Our study examines the effect of using either 2D or 3D joint coordinates as training data on the performance of speech-to-gesture deep generative models.

CCS CONCEPTS

• **Computing methodologies** → **Shape representations**; **Neural networks**;

KEYWORDS

Co-speech gesture generation, Pose Representation, Diffusion Models, Sequence modeling

ACM Reference Format:
Téo Guichoux[1,2], Laure Soulier[1], Nicolas Obin[2], Catherine Pelachaud[1,3]. 2024. 2D or not 2D: How Does the Dimensionality of Gesture Representation Affect 3D Co-Speech Gesture Generation?. In *ACM International Conference on Intelligent Virtual Agents (IVA '24), September 16–19, 2024, GLASGOW, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3652988.3673934

1 INTRODUCTION

In human communication, gestures play an integral role by conveying intentions and emphasizing points [13]. Recent studies [1–3, 5, 7, 14, 16, 18–22] aim to create similar gestures for Embodied Conversational Agents (ECA), using learning algorithms and extensive human motion datasets. Data on 2D motion is easily accessible from "in-the-wild" monocular videos, which are videos freely accessible online. To gather 3D gestures from such videos, one can use a pre-trained model to convert these 2D poses to 3D. However, it still faces the bottleneck of the 2D body pose representation. In "in-the-wild" datasets lifted in 3D, the source is still 2D monocular videos. Deciding whether to convert 2D poses to 3D beforehand or to train a model to generate 2D gestures and then lift them to 3D in post-processing, needs to be addressed. It remains unclear what the impact of the dimensionality of the skeleton representation is on the training of the generative model and the quality of the generated gestures. In this work, we compare two training settings to evaluate the influence of data dimensionality on the performance of two speech-to-gesture generative models by considering either 2D or 3D joint coordinates, as illustrated in Figure 1. More particularly, our contributions are the following:

• We propose an evaluation pipeline to investigate the impact of the dimensionality of the pose representation on the performance of two co-speech gesture generative models [20, 22]. We train both models to generate sequences of body poses represented in 2D coordinates which are then lifted to 3D using VideoPose3D [15].

• We empirically compare the quality of the gestures generated in 2D lifted to 3D to the gestures directly generated in 3D using evaluation metrics commonly used in co-speech gesture generation tasks [12, 20, 22].

• We conducted a user study where participants were asked to choose between gestures generated in 2D then lifted to 3D and gestures directly generated in 3D, providing a direct comparison of perceived quality. The code is provided at the following repository: `https://github.com/TGuichoux/2D-or-not-2D`

2 METHODOLOGY

2.1 Pipeline

To evaluate the inductive bias caused by the dimensionality of the gesture representation (2D or 3D) and the 2D-to-3D conversion, we trained co-speech gesture generators in both 2D and 3D settings. We employed a 3D lifter for post-processing the 2D generated sequences to be able to compare them to the 3D generated sequences. The two gesture generators selected for this study are DiffGesture [22] and Trimodal [20].

1) DiffGesture is defined as a DDPM that generates sequences of poses out of noise, conditioned on raw speech audio.

2) Trimodal is an encoder-decoder model trained in an adversarial scheme, that translates speech audio and text into 3D gestures, conditioned on speaker identity.

DiffGesture and Trimodal are designed to generate 3D gestures. We adapted these architectures to account for 2D body pose sequences by changing the input and output dimensions of the denoising network and recurrent decoder network respectively. Specifically, we removed the depth axis of the body pose coordinates. We define *DiffGesture 2D + VP3D* and *Trimodal 2D + VP3D* as DiffGesture and Trimodal trained on 2D motion data whose outputs

IVA '24, September 16–19, 2024, GLASGOW, United Kingdom

ACM ISBN 979-8-4007-0625-7/24/09
https://doi.org/10.1145/3652988.3673934

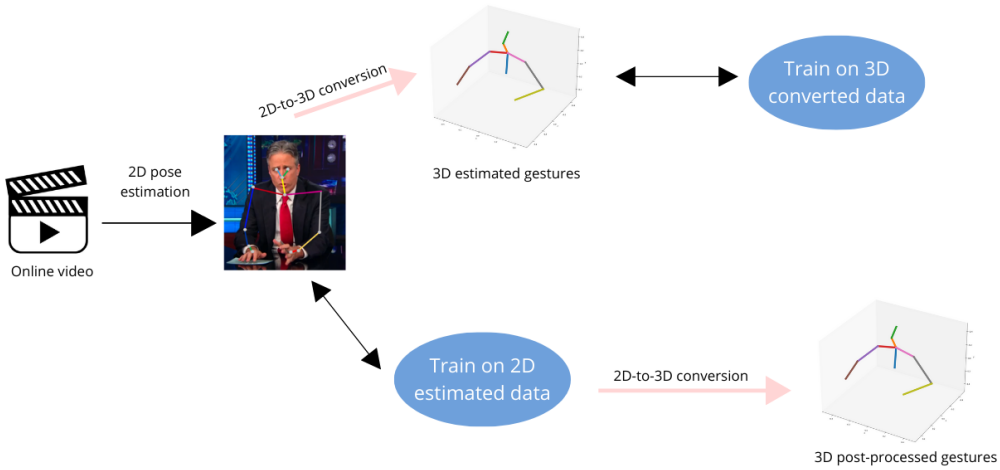

Figure 1: The two considered training settings for co-speech gesture generation from online videos.

are then lifted to 3D using VideoPose3D and we compare them to the original models, *DiffGesture 3D* and *Trimodal 3D* [20, 22]. Both models are trained on the TED Gesture-3D dataset [20], which is a dataset compiling 97h of paired speech-3D motion data. For 2D-to-3D conversion, we employed VideoPose3D [15], a Temporal Convolutional Network that maps a sequence of 2D poses to its 3D counterparts. We retrained VideoPose3D [15] on the TED Gesture-3D dataset to be able to input body poses when only the upper part of the body is considered.

2.2 Objective Evaluation metrics

We numerically evaluate our models with three commonly used metrics in the co-speech gesture generation field.

The Fréchet Gesture Distance (FGD): The FGD defined by Yoon et al. [20] computes the 2-Wasserstein distance between two distributions leveraging latent features extracted with a pose encoder. Similar distributions result in a high FGD value.

The Beat Consistency Score: The Beat Consistency Score (BC) [10, 11] measures the temporal consistency between kinematic and audio beats of a paired audio-motion sequence.

Diversity: The Diversity measure also leverages the latent features extracted with a pose encoder [9]. Diversity is computed by randomly selecting two sets of features from a distribution and calculating the distance between the mean of both sets.

3 USER STUDY

3.1 Protocol

We created videos [1] showing the co-speech gestures animation of the upper body of an articulated humanoid. Each video features two stimuli for pairwise comparison as it has been shown to reduce the cognitive load of the users [4, 17, 20]. The first animation is displayed on the left side of the screen with the second animation masked. The second animation is shown while the first one is masked. In each video, both animations used the same model (either DiffGesture or Trimodal), one with direct 3D gestures and the other with 2D gestures converted to 3D. We qualitatively evaluated the impact of the 2D-to-3D lifting, by including baseline videos (referred to as *Human GT*). These videos paired 3D pseudo-ground-truth gestures to those created by converting the 2D versions of these gestures to 3D using the retrained version of VideoPose3D. The pseudo-ground-truth gestures originate from the test set of the TED Gesture-3D dataset.

After viewing each video, participants were asked to select the animation they preferred in terms of human-likeness, aliveness, and speech synchrony. For each question, there were four possible answers: "*Clearly left*", "*Fairly left*", "*Fairly right*", and "*Clearly right*". Each response is assigned an integer value: +2 and +1 for a clear or slight preference for gestures directly generated in 3D, -2 and -1 for a clear or slight preference for lifted 3D gestures.

We recruited 67 participants on the Prolific platform [6, 8]. Among the participants who took the test, 7 failed the attention checks. The users were 36.7+/-11.8 years old and there were 37 females and 30 males and the median completion time was 17 minutes. We obtained 30 valid responses for each stimulus. The participants were paid 3£ if they passed the attention checks.

4 EXPERIMENTAL RESULTS

4.1 Objective evaluation

The results of our objective experiments are reported in Table 1. When comparing the results of *DiffGesture 2D + VP3D* and *Trimodal 2D + VP3D* to those of *DiffGesture 3D* and *Trimodal 3D*, we can notice that the models trained on 2D gestures perform worse than the original 3D models in terms of FGD, and BC. There is also a slight drop in diversity for Trimodal. We assume that the one-to-many relationship between 2D and 3D keypoints is mostly responsible for the performance drop of *DiffGesture 2D + VP3D* and *Trimodal 2D + VP3D* for the FGD. As VideoPose3D is deterministic, to one 2D pose it will systematically predict the same 3D pose although there exists multiple possibilities. Hence, the distribution resulting from lifting 2D sequences is tighter than the distribution directly

[1] Supplementary videos used are provided at the following website: https://sites.google.com/view/iva-2d-or-not-2d.

Table 1: Objective results of the experiments on the TED Gesture-3D dataset [20]. Up arrows indicate that a higher result is better whereas down arrows indicate that a lower result is better.

Methods	TED Gesture		
	FGD ↓	BC ↑	Diversity ↑
Evaluation on the 3D gesture space			
Ground Truth 3D	0	0.702	102.339
DiffGesture 3D [22]	1.947	0.678	101.436
DiffGesture 2D + VP3D	3.121	0.551	100.822
Trimodal 3D [20]	3.964	0.733	95.253
Trimodal 2D + VP3D	6.374	0.610	93.017

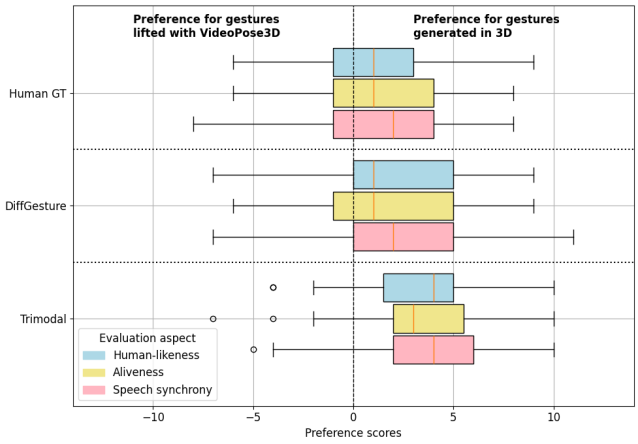

Figure 2: Evaluation study results. A positive score means that gestures generated directly in 3D are preferred over 2D gestures lifted to 3D. Reciprocally, a negative score means that 2D gestures lifted to 3D are preferred over direct 3D gestures. A score close to 0 means that the preference is unclear.

generated in 3D, explaining the high FGD of the gestures generated in 2D lifted to 3D. There is a drop in BC between the 3D models and their 2D counterparts. It can be that post-processing 2D gestures using VideoPose3D tends to over-smooth the resulting 3D gestures therefore reducing the number of kinematic beats.

4.2 Subjective evaluation

The results of our user study are presented in Figure 2, which highlights pairwise preferences between gestures generated directly in 3D or gestures generated in 2D lifted to 3D using VideoPose3D. A score significantly greater than zero indicates that gestures created directly in 3D are better perceived than those initially generated in 2D and subsequently converted to 3D using VideoPose3D. On the other hand, a score near zero means that the preference is unclear. We conducted a statistical analysis to determine the significance of our user study results. We used Student t-tests to check if the average scores for human likeness, aliveness, and speech synchrony were significantly different from zero. We also conducted Welch t-tests to determine the significance level of the model-wise mean

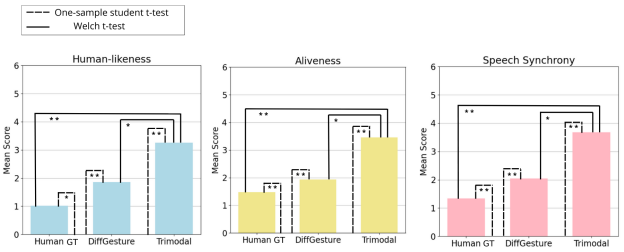

Figure 3: Statistical comparison of mean scores for each model and each aspect. The lines represent a significant superiority between the two values. Dotted lines correspond to Student t-tests and plain lines to Welch t-tests. * means p-value < 0.05 while ** means p-value < 0.01

score comparisons. The results of the Student t-tests and Welch tests are depicted in Figure 3.

First, for the DiffGesture and Trimodal models, all scores are significantly higher than zero, with a p-value less than 0.01. This suggests that gestures created directly in 3D by these methods are more effective than those initially created in 2D and then converted to 3D for all three aspects.

For the Human baseline (Human GT), converting 2D gestures to 3D demonstrates a minimal yet statistically significant impact on the perception of their human-likeness animation, with a score above zero (p-value of 0.020) and a confidence interval of the mean score close to zero. This suggests that while VideoPose3D influences the perceived human likeness, the effect is subtle. In contrast, the conversion process significantly affects gesture quality in terms of aliveness and speech synchrony, as evidenced by scores significantly higher than zero (p-values of 0.001 and 0.004, respectively). This indicates a notable degradation in these aspects due to the use of VideoPose3D for lifting 2D gestures to 3D.

We can conclude that training a model to generate 2D gestures and then converting these gestures to 3D deteriorates the overall animation quality in terms of human likeness, aliveness, and speech synchrony. The 2D-to-3D conversion of gestures has a small yet significant impact on the perception of human-likeness. Hence, the drop in human-likeness quality for gestures generated in 2D and then lifted to 3D may come from the training of the generative model itself since the 2D gesture representation may not allow the generation of highly human-like gestures once converted to 3D.

5 CONCLUSION

We conducted a study to measure the impact of the gesture representation dimensionality on the quality of 3D co-speech gesture generation. We selected two baseline models trained in two different settings: generation of 3D gestures and generation of 2D gestures with subsequent 2D-to-3D conversion. To evaluate the different approaches, we performed an objective evaluation and a large-scale user study involving 67 participants. From the results of our objective and subjective studies, we can conclude that the 3D representation of gestures during training is better for generating high-quality 3D motion synchronized with speech.

REFERENCES

[1] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Computer Graphics Forum* 39, 2 (2020), 487–496. https://doi.org/10.1111/cgf.13946 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13946

[2] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. *ACM Transactions on Graphics* 42, 4 (2023). https://doi.org/10.1145/3592458

[3] Tenglong Ao, Zeyi Zhang, and Libin Liu. [n. d.]. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. *ACM Trans. Graph.* ([n. d.]), 18 pages. https://doi.org/10.1145/3592097

[4] Andrew P. Clark, Kate L. Howard, Andy T. Woods, Ian S. Penton-Voak, and Christof Neumann. 2018. Why rate when you could compare? Using the "EloChoice" package to assess pairwise comparisons of perceived physical strength. *PLOS ONE* 13, 1 (01 2018), 1–16. https://doi.org/10.1371/journal.pone.0190393

[5] Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. 2023. Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 755–762. https://doi.org/10.1145/3577190.3616117

[6] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS One* 18, 3 (March 2023), e0279720.

[7] Mireille Fares, Catherine Pelachaud, and Nicolas Obin. 2023. Zero-Shot Style Transfer for Multimodal Data-Driven Gesture Synthesis. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. 1–4. https://doi.org/10.1109/FG57933.2023.10042658

[8] Patrik Jonell, Taras Kucherenko, Ilaria Torre, and Jonas Beskow. 2020. Can we trust online crowdworkers?: Comparing online and offline participants in a preference test of virtual agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20)*. ACM. https://doi.org/10.1145/3383652.3423860

[9] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. 2019. Dancing to Music. *CoRR* abs/1911.02001 (2019). arXiv:1911.02001 http://arxiv.org/abs/1911.02001

[10] Buyu Li, Yongchi Zhao, and Lu Sheng. 2021. DanceNet3D: Music Based Dance Generation with Parametric Motion Transformer. *CoRR* abs/2103.10206 (2021). arXiv:2103.10206 https://arxiv.org/abs/2103.10206

[11] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. 2021. Learn to Dance with AIST++: Music Conditioned 3D Dance Generation. *CoRR* abs/2101.08779 (2021). arXiv:2101.08779 https://arxiv.org/abs/2101.08779

[12] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022. Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10462–10472.

[13] David Mcneill. 1992. Hand and Mind: What Gestures Reveal About Thought. *University of Chicago Press* 27 (1992). https://doi.org/10.2307/1576015

[14] Shivam Mehta, Siyang Wang, Simon Alexanderson, Jonas Beskow, Eva Szekely, and Gustav Eje Henter. 2023. Diff-TTSG: Denoising probabilistic integrated speech and gesture synthesis. In *12th ISCA Speech Synthesis Workshop (SSW2023)*. ISCA. https://doi.org/10.21437/ssw.2023-24

[15] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Hendric Voß and Stefan Kopp. 2023. AQ-GT: a Temporally Aligned and Quantized GRU-Transformer for Co-Speech Gesture Synthesis. *arXiv preprint arXiv:2305.01241* (2023).

[17] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. 2022. A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents. *IEEE Transactions on Human-Machine Systems* 52, 3 (June 2022), 379–389. https://doi.org/10.1109/thms.2022.3149173

[18] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023. DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models. arXiv:2305.04919 [cs.HC]

[19] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. 2023. QPGesture: Quantization-Based and Phase-Guided Motion Matching for Natural Speech-Driven Gesture Generation. arXiv:2305.11094 [cs.HC]

[20] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics* 39, 6 (2020).

[21] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2018. Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots. *CoRR* abs/1810.12541 (2018). arXiv:1810.12541 http://arxiv.org/abs/1810.12541

[22] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. 2023. Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10544–10553.

Received 12th April, 2024