Gesture Generation with Diffusion Models Aided by Speech Activity Information

Rodolfo L. Tonoli[*][†]
r105652@dac.unicamp.br
Department of Computer Engineering and Automation, School of Electrical and Computer Engineering,
University of Campinas (UNICAMP)
Campinas, SP, Brazil

Leonardo B. de M. M. Marques[*]
Lucas H. Ueda
lmenezes@cpqd.com.br
lhueda@cpqd.com.br
CQPD
Campinas, SP, Brazil

Paula D. P. Costa[†]
paulad@unicamp.com
Department of Computer Engineering and Automation, School of Electrical and Computer Engineering,
University of Campinas (UNICAMP)
Campinas, SP, Brazil

ABSTRACT

This paper describes a gesture generation model based on state-of-the-art diffusion models. Novel adaptations were introduced to improve motion appropriateness relative to speech and human-likeness. Specifically, the main focus was to enhance gesture responsiveness to speech audio. We explored using a pre-trained Voice Activity Detector (VAD) to obtain more meaningful audio representations. The proposed model was submitted to the GENEA Challenge 2023. Perceptual experiments compared our model, labeled SH, with other submissions to the challenge. The results indicated that our model achieved competitive levels of human-likeness. While appropriateness to the agent's speech score was lower than most entries, there were no statistically significant differences from most models at the confidence level.

CCS CONCEPTS

• **Computing methodologies → Animation**; *Intelligent agents*; Machine learning.

KEYWORDS

Gesture generation, co-speech gestures, diffusion models

ACM Reference Format:
Rodolfo L. Tonoli, Leonardo B. de M. M. Marques, Lucas H. Ueda, and Paula D. P. Costa. 2023. Gesture Generation with Diffusion Models Aided by Speech Activity Information. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23 Companion), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3610661.3616554

1 INTRODUCTION

Human communication is composed of verbal and nonverbal behaviours. Co-speech gestures are one of these behaviours. They are visible actions of any body part produced while speaking and may serve different purposes, such as to provide emphasis or to depict some physical property [30]. Being such a key part of human communication, gestures are employed in embodied agents to simulate real interactions and create believable characters [29]. Otherwise, these agents may be perceived as lifeless or dull.

Recent research focused on automatic gesture generation (or synthesis) through deep learning. Such systems are able to animate embodied agents much faster and less time-demanding than traditional techniques such as hand-crafted animations or motion capture. Additionally, these techniques may not be suited for applications whose speech content is unknown beforehand, such as an avatar being controlled by a human or an embodied agent powered by a language model.

Most research on gesture generation has a cross-modal mapping approach to this problem, similar to a translation between different behaviour modalities [4]. Also, gestures are correlated with prosody and may be associated with semantics [21]. Thus, most systems use speech audio, speech text, or both to guide gesture generation [23].

However, synthetic data still struggles to appear human-like and appropriate to speech if compared to real human data [33]. More challenging scenarios could widen the gap between synthetic and real data. For example, in dyadic interactions, people are expected to take turns being the active speaker for brief or long moments. Most research has not addressed such situations. We propose a monadic gesture generation model that considers the voice activity for better alignment and responsiveness of gestures given speech audio. The model is based on a composition of the DiffuseStyleGesture [32], a speech-driven diffusion model and the Motion Diffusion Model (MDM) [26], which is text-driven. The main contributions of this paper to the aforementioned models are:

- the integration of voice activity information to improve turn-taking and speech audio synchrony while using only monadic inputs;
- the employment of aligned speech text as input through a pre-trained CLIP model, thus supporting the generation of gestures semantically related to speech;
- the use of speech audio representations suited for content-related tasks from a pre-trained WavLM model.

Our code can be accessed via https://github.com/AI-Unicamp/ggvad-genea2023.

This article is structured as follows: Section 2 presents related works on gesture generation and diffusion; the data processing is

[*] Both authors contributed equally to this research.
[†] Also with Artificial Intelligence Lab., Recod.ai, Institute of Computing, University of Campinas, SP, Brazil..

ICMI '23 Companion, October 9–13, 2023, Paris, France

ACM ISBN 979-8-4007-0321-8/23/10…$15.00
https://doi.org/10.1145/3610661.3616554

detailed in Section 3; Section 4 describes the proposed model and qualitative evaluations of our model are presented in Section 5; the results of the proposed model compared to other entries to the GENEA Challenge 2023 are detailed in Section 6; and Section 7 presents the conclusion and final remarks.

2 BACKGROUND AND PRIOR WORK

Generative models enable the capture of the one-to-many nature of gestures. Studies using VAEs [9], GANs [8], and Normalizing Flows [10] show that such models surpass deterministic ones. However, these approaches still suffer from generalized problems such as mean pose convergence and training instability. Recently, diffusion models arose as a new class of promising generative models achieving state-of-the-art results across a wide range of multimodal tasks validated by perceptual evaluations without the same pitfalls as the generative models mentioned before. Additionally, these models were shown to be capable of handling data with special structures, efficient sampling and providing improved likelihood estimation [31].

Denoising Diffusion Probabilistic Models (DDPMs) [12] are a type of generative model that synthesize new samples from an underlying data distribution by learning how to reconstruct information. During the training process, the model takes one noisy data point (x_t), obtained by applying t Gaussian noise addition steps to the original data (x), with $0 < t \leq T$, as T is the size of the complete diffusion noise-adding chain, and is set to equivalently predict either a one-step denoised sample (x_{t-1}), a fully reconstructed data point (x_0), or the noise contained (ϵ). On inference, the process is started from a pure Gaussian noise distribution and the reconstruction is performed iteratively T times, generating a new sample [12].

Diffusion models exhibited state-of-the-art performance in several different tasks. On image synthesis, diffusion models achieved superior performance to the at the time GAN-based state-of-the-art synthesis [7], and were also proven to be able to generate and edit hyper-realistic images [22, 25]. In the audio domain, diffusion models have been successfully exploited for audio generation [15] and text-to-audio [19] tasks, obtaining higher performance when compared to other current staple models. Recently, diffusion models have also been explored on the task of video generation, which were demonstrated to synthesize high-fidelity videos with a high degree of controllability and world knowledge [11].

In the context of human motion generation, text-based models aim to control the movements via natural language semantically. The MotionDiffuse model [35] is the first model to exploit DDPMs for this task, combining these models with a cross-modal Transformer based architecture. In another approach, denominated Motion Diffusion Model (MDM) [26], textual representations extracted from a pre-trained CLIP [24] are combined with a Transformer model in a classifier-free guidance diffusion training process [13]. Other works tackle the dance generation task, which intends to generate dances given music as audio input. The EDGE [27] method pairs a diffusion model with Jukebox, a generative model for music, whereas the Listen, Denoise and Action! [1] model adapts DiffWave [15] to generate poses and synthesize dances in various styles.

More recently, diffusion models have also been applied to the gesture generation task. DiffMotion [34] is the first approach that applies DDPMs to generate gestures. It leverages an autoregressive temporal encoder based on an LSTM that processes context represented by spectral audio features and previous poses to condition a diffusion process, generating each pose individually.

The DiffGesture [37] model uses a convolutional audio encoder to extract representations directly from the raw audio. A Transformer model then uses these representations that undergoes an implicit classifier-free guidance diffusion training.

The GestureDiffuCLIP [2] model introduces a multimodal (text, motion or video) prompt-conditioned style-controlled gesture generation via mode-specific pre-trained CLIP encoders. Also, they use a contrastive learning strategy to learn semantic correspondences between textual transcripts of the input speech and gestures, allowing for the generation of semantically-aware gestures. These contributions, along with a denoiser network based on Transformers, attention, and AdaIN layers [14] to incorporate style guidance, compose a latent diffusion training process [25].

Finally, the DiffuseStyleGesture [32] model combines layers of cross-local and global attention to better capture the localized aspects of gestures. With representations extracted from the self-supervised WavLM model [6], the authors perform a diffusion training process and are able to generate and control gestures based on a style label.

Although the increasing interest in the field, the synthesized motions from most models are still far from being indistinguishable from real human motion [33]. Moreover, research often concentrates on monadic scenarios in which only one participant actively communicates. Consequently, crucial behaviours of real-life interactions, such as listening, reciprocal expression, and interruptions, are disregarded during development and evaluation.

3 DATA AND DATA PROCESSING

The dataset used by the 2023 GENEA Challenge is an adaptation of the Talking With Hands 16.2M (TWH) data [18]. Pre-processing, data augmentation, and selection are described in the challenge's main paper [17]. The available dataset presents a dyadic scenario, i.e., it is composed of data from two people having a conversation, referred to as the main agent and interlocutor. Entries to the challenge should only generate movements for the main agent, and using the interlocutor's data was optional. Available data includes motion, speech audio, speech text (audio transcripts with timestamps), and speaker label. We only used data from the main agent; thus, our model depends on monadic information alone despite the dyadic scenario. Speaker labels were also ignored.

The dataset motions are BVH files with movements composed of 30 poses per second represented by Euler angles. We extracted each pose and composed a feature vector $\boldsymbol{g} = [\boldsymbol{\rho}_p, \dot{\boldsymbol{\rho}}_p, \boldsymbol{\rho}_r, \dot{\boldsymbol{\rho}}_r]$ where $\boldsymbol{\rho}_p \in \mathbb{R}^{3j}$ and $\dot{\boldsymbol{\rho}}_p \in \mathbb{R}^{3j}$ are the global 3D joint positions and positional velocities, $\boldsymbol{\rho}_r \in \mathbb{R}^{6j}$ and $\dot{\boldsymbol{\rho}}_r \in \mathbb{R}^{3j}$ are the local 6D joint rotations [36] and the local 3D joint rotational velocities, j represents the number of joints. The 30 frames per second rate of the original data and all 83 joints of the skeleton were preserved, thus $\boldsymbol{g} \in \mathbb{R}^{1245}$ for each pose. Each dimension of motion data is normalized to zero mean and unit standard deviation over the challenge training set. Audio files were resampled from 44.1 kHz to 16 kHz.

4 METHOD

Our approach consists of a combination of the MDM [26] and the DiffuseStyleGestures [32] models, with modifications aiming for improved responsiveness of gestures given speech audio. The architecture is shown in Figure 1. Our model generates sequences of 120 poses simultaneously, corresponding to 4 seconds. We consider inputs to be divided into global and fine-grained information. The first corresponds to information relevant to the 4-second sequence as a whole, which includes the words spoken (text), seed poses, and timestep embedding. On the other hand, fine-grained information is considered to be relevant at the frame level; thus, it includes audio and speech activity.

4.1 Global Information

Since gestures can be semantically related to speech, providing text information could improve gesture appropriateness. As textual features, we use spoken words within a motion sequence. Words timestamps from the audio transcript are used for extracting the corresponding words. As in the MDM [26] model, the speech text contained in the sequence of poses passes by a pre-trained CLIP [24] model[1] and then processed from the clip output dimension of 512 to a dimension of 64 by a fully connected layer.

For the motion between consecutive generated sequences to have cohesion, 10 previous seed poses are used as conditional input. These poses are flattened and then projected to a dimension of 192, and then concatenated with the textual information, forming a vector with the defined latent dimension of 256. Additionally, the timestep embedding of the diffusion process, which indicates which denoising step is being performed, is a sinusoidal positional embedding that is passed through two fully connected layers with a Sigmoid Linear Unit (SiLU) activation layer in between and projected to latent dimension. With this, the embedding that represents global conditioning information (the one that is invariant to the pose sequence) is obtained by summing the time-step embedding with the concatenation of the textual and seed poses embedding.

4.2 Fine-grained Information

We work with chunks of sequences of 120 poses corresponding to 4 seconds of motion. The noisy poses for the diffusion process are obtained by adding t steps of Gaussian noise on a sequence. These poses are then projected via a linear layer from the pose dimension of 1245 to the latent space dimension. For the audio information, we use the resampled audio data and pass it through the WavLM [6] model[2]. Differently from the DiffuseStyleGestures [32], we use the representations extracted from the 11th layer instead of the 12th. The 11th layer is reported to perform better at content-related tasks, such as phoneme recognition and automatic speech recognition. These representations are first interpolated to match the length of the corresponding pose sequence and then projected to a dimension of 64 by a linear layer.

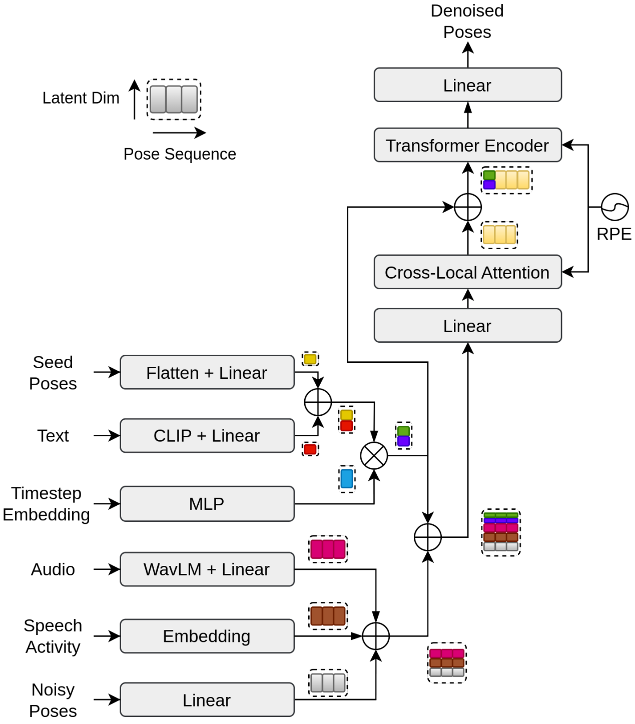

Figure 1: Model architecture.

4.2.1 *Speech Activity Information.*

Due to the dyadic nature of the dataset, some sections of the data are composed of moments in which the main agent is not the active speaker, such as listening and turn-taking moments. Gestures performed in active or non-active moments may play different roles in human interaction and, thus, differ from those performed in other moments. For example, *beat* gestures occur during articulations of speech and may serve to emphasize what is being said [21]; differently, mimicry, often performed automatically, may enhance helpfulness and strengthen social bonds [28]. Although our model only uses monadic data, we introduce the use of speech activity information. This information, otherwise embedded in audio representations such as spectrograms and MFCCs, may be lost in the abstract WavLM representations. Furthermore, the interpolation of representations to match the pose sequence can blend moments with and without speech activity. Thus, the contribution of such inclusion is believed to be two-fold. First, it provides more straightforward access to fine-grained speech energy. Second, it helps to stress, during training, the difference between gestures in the aforementioned moments, not in terms of functionality, but dynamics.

Speech activity can be inferred through analytical approaches such as energy and F0. However, the dataset audios contain noise that could affect computing these parameters: various speakers, different speech volumes, and background noise such as speech from the interlocutor and breathing. Thus, we consider two scenarios for acquiring speech activity information. The first is based on a pre-trained Voice Activity Detector (VAD)[3] that consists in

[1] Version 'ViT-B/32' obtained from https://github.com/openai/CLIP
[2] Version 'Base+' obtained from https://github.com/microsoft/unilm/tree/master/wavlm
[3] Obtained from https://huggingface.co/speechbrain/vad-crdnn-libriparty

a small CRDNN (A combination of convolutional, recurrent and deep neural network) trained on the Libriparty dataset[4], which is a synthetic cocktail-party scenario derived from the Librispeech dataset. When speech is detected, the model outputs a 1 and otherwise a 0. The second approach is taken from the annotated speech text timestamps provided in the dataset. When there is any text, we consider the respective timestamps as 1 and otherwise as 0.

The major difference between these approaches is that the pre-trained model can detect intra-text pauses, whereas audio transcripts provide word-level timestamps granularity. A comparison of both is shown in Figure 2. From the figure, it is noticeable that VAD provides closer alignment with speech energy. Besides, the pre-trained VAD removes the need for audio-aligned annotated speech text, which is sensitive to human perception or error.

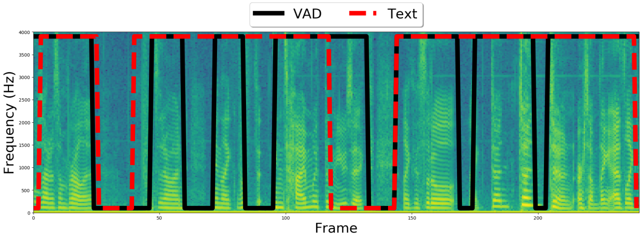

Figure 2: Scaled speech activities from timed audio transcripts (red) and from the VAD (black) overlapped with a spectrogram of an eight-second audio sample in the background.

The speech information sequence extracted from the VAD is used to select two embeddings with latent dimensions representing the presence of speech or no speech for each pose. This sequence of embeddings is then concatenated with the noisy poses and the audio embeddings forming the fine-grained information.

4.3 Training

The fine-grained information is concatenated with the global information along the latent dimension. Then, all the input information is projected back to the latent dimension by an input linear layer and fed to the cross-local attention layer to capture local relations between the features. Then, we concatenate the global information embedding one more time with the output along the sequence dimension before passing the sequence to the transformer encoder to capture the global context. Then, we ignore the first token of the output sequence and project the outputs to the pose dimension, which finally represents the denoised pose (x_0) itself. We use positional embeddings to add sequence information on both the cross-local attention and the transformer encoder.

On inference, a sequence at a time is generated. The model outputs a vector $G = [g_1, g_2, \cdots, g_{120}]$. The last 10 poses from the previously generated sequence are used to condition the generation of the next sequence; mean poses are used for conditioning the first sequence.

For post-processing, we use linear interpolation to impose continuity between successive sequences. To smooth motion artifacts in the output, we also apply a Savitzky-Golay [20] filter with a window length of 9 and polynomial order of 3.

The model was trained for 290k steps, with a batch size of 64, in a single NVIDIA Titan Xp GPU, which took about 2.5 days.

5 EVALUATION

There still is no objective metric to measure gesture perception reliably. Moreover, previous research has found that object metrics differ from subjective ones [16]. Therefore, the research team empirically evaluated the proposed model, its variations, and the reference models through visual inspection of their outputs.

We trained the MDM [26] and the DiffuseStyleGestures [32] and used them as references for comparison, i.e., a starting point for development. Although providing reasonable human-like motion, in terms of appropriateness to speech, we found the results unsatisfactory. The outputs seemed unaware of moments such as brief pauses, turn-taking, and listening moments. That is, the agent would frequently make gestures in those moments that appeared inadequate and similar to behaviours performed when it was the active speaker. So, our main focus in developing the model for the GENEA Challenge 2023 was to overcome those issues of disregard for no-speak moments. Thus, a VAD was employed to leverage speech activity information.

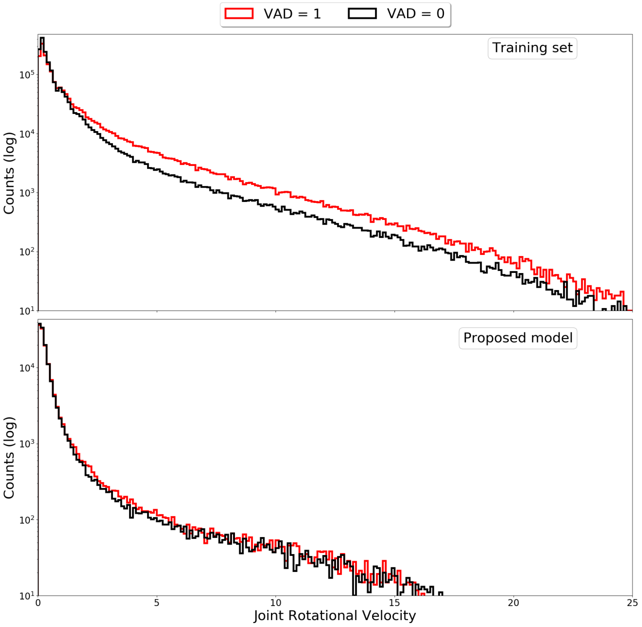

Figure 3: Histograms of the rotational velocities from the main agent's left and right forearm joints from the training set of the dataset (top), and the output of the proposed model with (bottom). Red and black indicate velocities extracted when the main agent was the active speaker and when it was not.

[4] https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriParty/generate_dataset

In order to examine the effectiveness of the VAD, we present histograms of the rotational velocities of the forearms, a joint that is very active when gesturing with the arms, on Figure 3, for the real training set (top), and the output of the proposed model (bottom). The figure splits each set considered in two distributions: when the VAD indicates that there is an occurrence of speech, VAD output equals one, and when the VAD indicates that there is no speech, its output is zero.

For the training set, the histograms reveal distinct patterns associated with speech activity during gesticulation. Speakers in the dataset exhibit increased forearm movements while talking versus silence periods. These insights support our underlying assumption that people tend to perform more gestures - or at least more abrupt gestures - when they are speaking.

The proposed model could reproduce, to some extent, the overall behaviour of the training set. However, it was unable to synthesize motion that reproduced the differences seen in the training set given speech activity, that is, a larger concentration of higher velocities when the agent is speaking. We did an ablation study with the proposed model without the VAD module. Its histogram was similar to the one with VAD. However, visual inspections of the outputs by the research team favored outputs by the proposed model with VAD in terms of speech and gesture alignment.

We also compared outputs from models with and without text input. However, we did not find a significant amount of semantically related gestures in their output. Further investigation should be carried out to indicate if there is a sufficient amount of such gestures in the dataset for models to be able to learn from. Still, we kept texts as input as motion quality was not impaired.

Compared to our reference models, the output of the proposed model seems better, especially in terms of speech audio and gesture alignment. However, we notice that some artifacts are still present in the motions. Motions occasionally converge to an unusual or odd-looking pose, absurd rotations still take place, and jittering is sometimes noticeable.

6 RESULTS AND DISCUSSION

The results of the shared evaluations of the GENEA Challenge 2023 indicated that our model (condition SH) is competitive with most conditions in terms of human-likeness but obtained relatively poor results for appropriateness to speech [17].

Figure 4 presents human-likeness ratings. Subjects participants gave their ratings based on how human-like the motions appeared, from 0 (worst) to 100 (best). Real motion data (NA) achieved a median rating of 71, the baselines 46 (BD) and 43 (BM), while our condition scored 46. We believe that the module that contributed the most to the human-likeness of generated gestures is the attention mechanism. As Yang et al. [32] showed in their ablation studies, the cross-local attention module played a significant role in terms of human-likeness ratings.

Two evaluations were performed to assess gesture appropriateness to speech: appropriateness for agent speech and for the interlocutor speech. The first contains mainly moments where the main agent is the active speaker, while the roles are reversed in the latter.

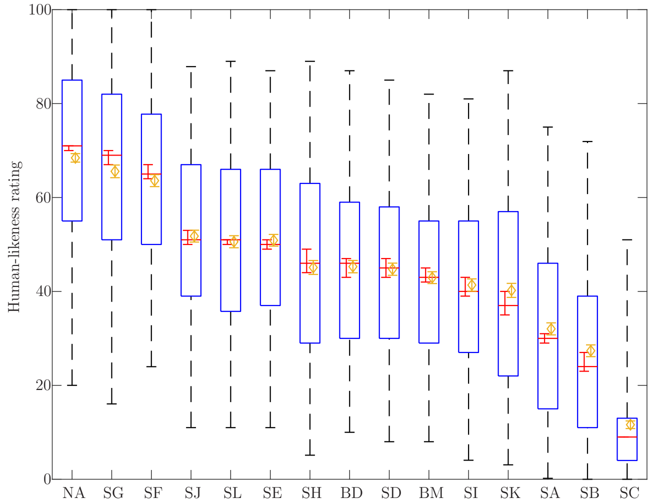

Figure 4: Shared human-likeness rating study. Red bars are median ratings; yellow diamonds are mean ratings. Entries to the challenge are labeled SA-SL (ours is SH), BD and BM are the baselines [5], and NA is real data. Extracted from Kucherenko et al. [17].

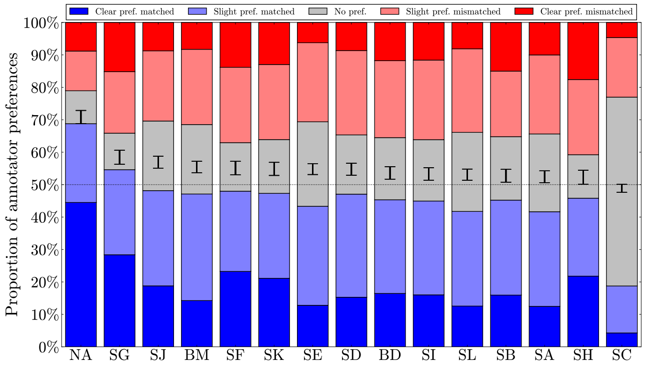

Figure 5: Shared appropriateness for agent speech responses. Entries to the challenge are labeled SA-SL (ours is SH), BD and BM are the baselines [5], and NA is real data. Extracted from Kucherenko et al. [17].

For the appropriateness of agent speech evaluation, subjects were presented with speech audio and two motions generated by the model. One motion is the output generated with the speech audio presented as input, and the other is the output from another segment of speech audio. For our condition, subjects preferred the matching motion 52.9% of the time, slightly above chance. Although one of the lowest mean appropriateness scores, there is no statically significant differences in the scores of ours and another ten conditions (conditions BM to SA, in Figure 5).

Our condition had the lowest score in the appropriateness for the interlocutor evaluation. This means that subjects found the mismatched stimuli more appropriate. However, our model does not use any interlocutor information as input. Thus, from the model's

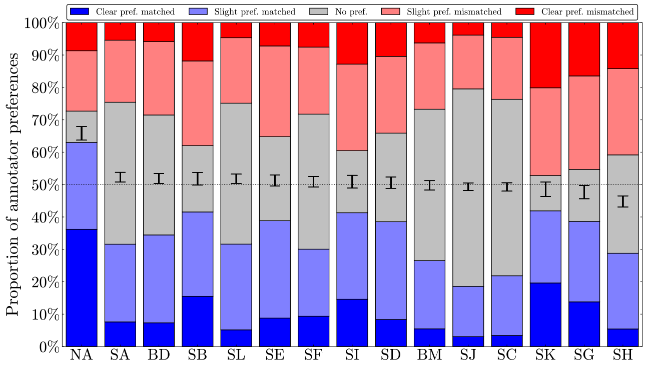

Figure 6: Shared appropriateness for the interlocutor responses. Entries to the challenge are labeled SA-SL (ours is SH), BD and BM are the baselines, and NA is real data. Extracted from Kucherenko et al. [17].

perspective, the output of both matched and mismatched stimuli for this evaluation was generated using the same inputs. Evidently, it is not expected that the outputs be exactly the same due to the probabilistic nature of the model. But both outputs are expected to be equivalent in terms of human-likeness and appropriateness, thus scoring similarly to chance (50%).

We also noticed from all three evaluations that our model had a wide range of scores. For instance, whiskers from the box plot visualization of Figure 4 span almost the entire y-axis; our condition, along with condition SK, had the highest confidence intervals of median and mean ratings. In the appropriateness for agent speech, our condition had the third highest number of clear preferences for matched stimuli, the highest for mismatched, and the second lowest for no preferences when compared to other entries to the challenge. Thus, we argue that the proposed model is indeed capable of generating gestures that are competitive in terms of human-likeness and appropriateness for the main agent. However, the artifacts mentioned in the previous section hinder gesture perception and should be addressed before any conclusion regarding the proposed architecture and individual modules.

7 CONCLUSION

This paper describes the proposed diffusion-based model for gesture generation that uses pre-trained VAD. Incorporating speech activity information in such models could improve responsiveness during rapid back-and-forth interactions. Also, a VAD can explicitly provide this information without needing human-annotated transcripts, thus potentially suited for real-time dialogue. Our model has been compared with others in the GENEA Challenge 2023, a crowd-sourced evaluation that directly compares different methods while controlling factors such as data and evaluation methodology. The evaluation showed that our model is compatible with other entries to the challenge in terms of human-likeness, but appropriateness to speech is still unimpressive despite our efforts.

Our experiments revealed mixed results regarding the effectiveness of the proposed implementation improvements to the gesture generation system. While convergences to undesired poses, extreme joint rotations, and jittering were not frequent, they nonetheless occurred. Besides, output motion was unstable, i.e., when generating motions given the same inputs, the resulting motion quality varied greatly. These issues may have contributed to subpar performance in evaluations and compromised the responsiveness of generated gestures to speaking moments. Although our adaptations hold potential value for gesture generation tasks, further improvements are needed to leverage their benefits fully. Especially the explicit use of speech activity information that could be leveraged to address turn-taking moments

We intend to focus primarily on improving speech and gesture alignment for future work. An interesting approach is adapting an external framework for alignment as the one proposed by Badlani et al. [3]. Another obvious path is to incorporate data from the interlocutor to capture the aspects of dyadic scenarios.

ACKNOWLEDGMENTS

This study was partially funded by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Finance Code 001. The first author is grateful to the Eldorado Research Institute.

REFERENCES

[1] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2022. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *arXiv preprint arXiv:2211.09707* (2022).

[2] Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. GestureDiffuCLIP: Gesture diffusion model with CLIP latents. *arXiv preprint arXiv:2303.14613* (2023).

[3] Rohan Badlani, Adrian Łańcucki, Kevin J Shih, Rafael Valle, Wei Ping, and Bryan Catanzaro. 2022. One TTS alignment to rule them all. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6092–6096.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (Feb. 2019), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607 Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[5] Che-Jui Chang, Sen Zhang, and Mubbasir Kapadia. 2022. The IVI Lab entry to the GENEA Challenge 2022–A Tacotron2 Based Method for Co-Speech Gesture Generation With Locality-Constraint Attention Mechanism. In *Proceedings of the 2022 International Conference on Multimodal Interaction (ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 784–789. https://doi.org/10.1145/3536221.3558060

[6] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518.

[7] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.

[8] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2019. Multi-Objective Adversarial Gesture Generation. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games* (Newcastle upon Tyne, United Kingdom) *(MIG '19)*. Association for Computing Machinery, New York, NY, USA, Article 3, 10 pages. https://doi.org/10.1145/3359566.3360053

[9] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. 2023. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 206–216.

[10] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. Moglow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–14.

[11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[13] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[14] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*. 1501–1510.

[15] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761* (2020).

[16] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021. A Large, Crowdsourced Evaluation of Gesture Generation Systems on Common Data: The GENEA Challenge 2020. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 11–21. https://doi.org/10.1145/3397481.3450692

[17] Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. The GENEA Challenge 2023: A large-scale evaluation of gesture generation models in monadic and dyadic settings. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '23)*. ACM.

[18] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S Srinivasa, and Yaser Sheikh. 2019. Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV '19)*. 763–772.

[19] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503* (2023).

[20] Jianwen Luo, Kui Ying, and Jing Bai. 2005. Savitzky–Golay smoothing and differentiation filter for even number data. *Signal processing* 85, 7 (2005), 1429–1434.

[21] David McNeill. 1992. *Hand and Mind: What Gestures Reveal About Thought.* University of Chigado Press. https://doi.org/10.1177/002383099403700208

[22] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[23] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 569–596.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[26] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=SJ1kSyO2jwu

[27] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 448–458.

[28] Rick B Van Baaren, Rob W Holland, Kerry Kawakami, and Ad Van Knippenberg. 2004. Mimicry and prosocial behavior. *Psychological science* 15, 1 (2004), 71–74.

[29] V. Vinayagamoorthy, M. Gillies, A. Steed, E. Tanguy, X. Pan, C. Loscos, and M. Slater. 2006. Building Expression into Virtual Characters. In *Eurographics 2006 - State of the Art Reports*, Brian Wyvill and Alexander Wilkie (Eds.). The Eurographics Association. https://doi.org/10.2312/egst.20061052

[30] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview. , 209–232 pages.

[31] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2022. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796* (2022).

[32] 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 preprint arXiv:2305.04919* (2023).

[33] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A Large Evaluation of Data-Driven Co-Speech Gesture Generation *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 736–747. https://doi.org/10.1145/3536221.3558058

[34] Fan Zhang, Naye Ji, Fuxing Gao, and Yongping Li. 2023. DiffMotion: Speech-Driven Gesture Synthesis Using Denoising Diffusion Model. In *MultiMedia Modeling*, Duc-Tien Dang-Nguyen, Cathal Gurrin, Martha Larson, Alan F. Smeaton, Stevan Rudinac, Minh-Son Dao, Christoph Trattner, and Phoebe Chen (Eds.). Springer International Publishing, Cham, 231–242.

[35] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001* (2022).

[36] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the Continuity of Rotation Representations in Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[37] 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 (CVPR)*. 10544–10553.