Zero-Shot Stylized Gesture Generation from Speech: An Entry to the GENEA challenge 2022

ANONYMOUS AUTHOR(S)

We present a neural network framework for gesture generation from speech audio input with zero-shot style control by example. Style can hereby be controlled via only a short example motion clip, even for motion styles unseen during training. We use a variational framework to learn a disentangled style embedding.The probabilistic nature of our variational framework furthermore enables the generation of a variety of outputs given the same input, addressing the stochastic nature of gesture motion. The GENEA challenge evaluation showed that our model produces full-body motion with highly competitive levels of human-likeness.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: gesture generation, machine learning, style transfer, computer animation

ACM Reference Format:
Anonymous Author(s). 2022. Zero-Shot Stylized Gesture Generation from Speech: An Entry to the GENEA challenge 2022. In . ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Non-verbal behavior is an essential part of human interaction, carrying additional information to semantic speech content [8, 18]. Much research has focused on modelling this non-verbal behavior in order to increase virtual agents' life-likeness and appeal, among others, co-speech gesture has received much attention. A key challenge of generating gestures from speech input is the non-deterministic nature of the speech-gesture relationship, where a speech line may be associated with many different motions. Consequently, there is not one true gesture match, but numerous equally appropriate gestures, posing a one-to-many mapping problem. Moreover, each person has their own style of movement, implying that a generative model should capture a wide range motion variation, preserving characters' idiosyncrasies.

Previous works address speaker and gesture variety by modelling and generating gestures for specific speakers individually [2, 7, 17, 25], or by manipulating gesture motion through general statistics such as hand height and velocity [3, 26]. A shortcoming of these approaches is their dependence on the data for creating motion variation, requiring examples of each speaker and style for training the generative model and having trouble generalizing to motion styles beyond the training set.

In this work, we propose a zero-shot style-controllable gesture generation system that can generalize to motion styles beyond training data and alleviates previous limitations to generating gesture styles in a training set-dependent manner. Styles are encoded from short example motion clips that do not have to be part of the training set. Style representation and conditioning in this manner does not require any explicit labels, circumventing the difficulty of

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2022 Association for Computing Machinery.
Manuscript submitted to ACM

naming gesture styles. Our system is probabilistic, enabling generation of many variations for the same speech input simply by re-sampling, and therefore addressing the stochastic nature of co-speech gesture.

2 RELATED WORK

Data-driven approaches to generating co-speech gesture motion aim to learn a mapping from a speech input signal to the corresponding output motion. As this is a one-to-many mapping problem, as discussed above, directly minimizing a difference between target and output motion of a deterministic model can lead to mean pose convergence with small-range motion (e.g. [5] and [13]). Alternatively, the use of adversarial losses have been proposed [6, 7, 19], relying on a second neural network to judge output realism. However, generated motion still fails to match real motion in terms of naturalness and frequently ties with mismatched real motion on appropriateness measures [7, 14].

Employing a probabilistic framework, [3] predict the next pose distribution instead of a fixed pose and by re-sampling from the distribution, a variety of output motion can be generated. For further output variation control, the authors propose style manipulation via four motion parameters, namely gesture speed, height, spacial extent and lateral symmetry. In a similar manner, [26] enable style control via gesture speed, spacial extent, and handedness. [21] instead propose the use of the Laban Effort and Shape parameters to drive gesture style based on the desired personality of a character. Common to all three above approaches is a reliance handcrafted control parameters; these are often too limited to encode a wide range of distinct styles to be learned by the model.

Alternative to hand-crafted style features, a number of works from similar domains have proposed example-based style control. For human locomotion, Aberman et al. [1] perform style transfer via example clips and enable generalization to unseen styles, however, training requires labeled style examples. For dance motion generation, Valle-Pérez et al. [22] present a model conditioned on music and a short style example motion, though the model could not be shown to generalize beyond the motion styles of the training data.

For this work, we looked to recent advances in speech synthesis research. To enable speech stylization by example style generalization beyond the training data, previous works augment a text-to-speech model with a style encoder capturing the general style and prosody of a line, then used to condition the speech synthesis [11, 23, 28]. In addition to their style generalization capabilities, these models produce style latent space that allows for meaningful interpolations. Extending the style generalization, [28] enabled speech synthesis even for unseen speakers.

3 SYSTEM OVERVIEW

Our gesture generation system is composed of three components, (1) a Speech Encoder producing a speech embedding sequence S, (2) a Style Encoder summarizing and example motion clip into a style embedding vector $\mathbf{e}$, and (3) a Gesture Generator, an autoregressive decoder that generates a pose sequence Y from the speech and style embeddings. An overview of our system is shown in Fig. 1.

3.1 Speech Encoder

The Speech Encoder, shown in Fig. 2, produces a sequence of speech embedding vectors from the audio input. First, audio features are extracted from the raw audio input as described in Sec. 5. Next, extracted features are passed through 1D convolution layers followed by non-linear operators, followed by a frame-wise linear layer. The kernels of the convolutional layers are set to create a receptive field of about 1 second of speech. The convolutional layers have 64 channels and a kernel size of 3 and 31, respectively, and are followed by a 0.2 dropout rate and ELU activations.

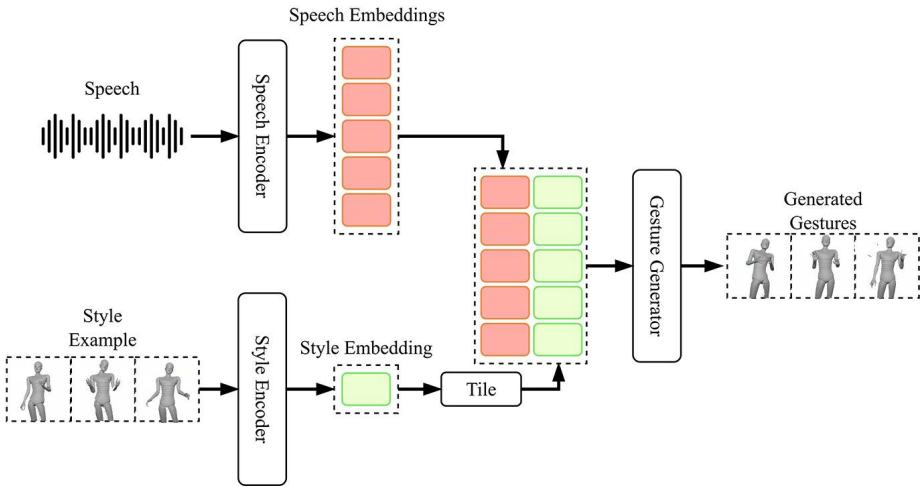


Fig. 1. Overview of our gesture generation system. The Speech Encoder takes the audio input and produces a speech embedding vector. The Style Encoder takes an example motion clip and produces a style embedding vector. The speech and style embedding vectors are fed to the Gesture Generator which generates the output pose sequence.

The Speech Encoder yields a sequence of embedding vectors $S = [\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_{T-1}]$ where T is the number of frames in the sequence and $\mathbf{s} \in \mathbb{R}^{D_S}$. The dimensionality of the speech embedding (D_S) is 64.

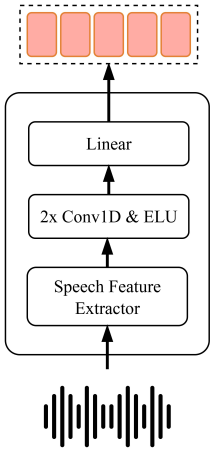
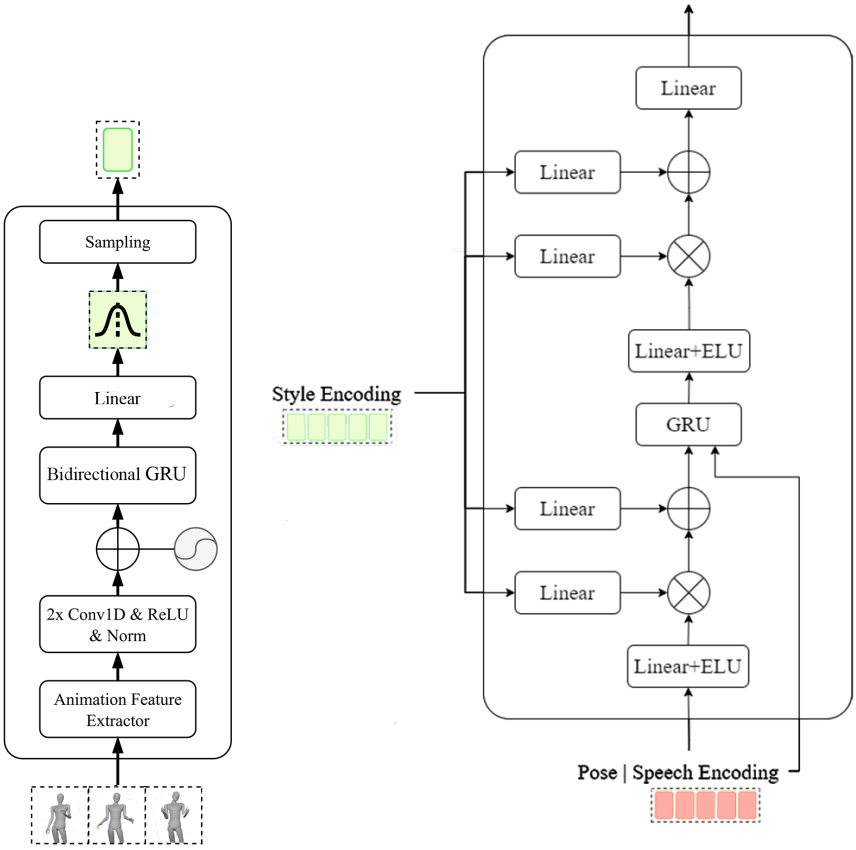


Fig. 2. Architecture of our Speech Encoder.

3.2 Style Encoder

The Style Encoder produces a fixed size low dimensional embedding vector representing the motion style of an example clip. Fig. 3a shows an overview of its architecture. This example clip is represented by a sequence of M feature frames, $A = [\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_{M-1}]$. Feature extraction from the example clip is described in Sec. 5. We apply two 1D convolution layers each with a kernel size of 3 and 512 output channels, followed by dropout layers with a drop rate of 0.2. The



(a) Style Encoder (b) Gesture Generator

Fig. 3. Architectures of the Style Encoder and the Recurrent Decoder

convolutional layers are followed by a ReLU activation and layer normalization, before the output is passed to a bidirectional GRU of size 512 and a linear projection layer. This yields the the parameters μ and σ of the D_e-dimensional multivariate Gaussian distribution from which we sample the final style embedding vector $\mathbf{e} \in \mathbb{R}^{D_e}$. The dimensionality of the style embedding (D_e) is 64. This sampling procedure is described in the Variational Auto-Encoder (VAE) framework [12]. VAEs are known for learning disentangled latent spaces [20] that allow for interpolation [4]. Moreover, they allow us to sample variations at inference time.

3.3 Gesture Generator

Our Gesture Generator takes the speech embedding sequence S and the reference style embedding vector $\mathbf{e}$ and produces the final gesture sequence $Y = \{\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{T-1}\}$. The Gesture Generator is depicted in Fig. 3b. The speech embedding is passed through two layers of Gated Recurrent Units (GRU) of size 1024, with the style embedding used as feature-wise linear modulation (FiLM). The Gesture Generator yields the output pose, parameterized similarly to the Style Encoder as $\mathbf{y} = [\boldsymbol{\rho}_p, \boldsymbol{\rho}_r, \dot{\boldsymbol{\rho}}_p, \dot{\boldsymbol{\rho}}_r, \mathbf{r}_p, \mathbf{r}_r, \dot{\mathbf{r}}_p, \dot{\mathbf{r}}_r]$, where $\boldsymbol{\rho}_p, \boldsymbol{\rho}_r, \dot{\boldsymbol{\rho}}_p, \dot{\boldsymbol{\rho}}_r$ are joint local translations and rotations along with their

velocities, and $\dot{\mathbf{r}}_p$ and $\dot{\mathbf{r}}_r$ are the character root translational and rotational velocity local to the character root transform. $\mathbf{r}_p \in \mathbb{R}^3$ and $\mathbf{r}_r \in \mathbb{R}^4$ are the position and orientation of the root (represented as quaternions), respectively, which are updated using the root translational and rotational velocities at each frame.

4 TRAINING & LOSSES

Our network is trained end-to-end with the Rectified Adam optimizer [16] with a learning rate of $1e-4$ and a decay factor of 0.995 applied at every 1k iterations. We use a batch size of 32 sequences of length $T = 256$ frames (4.26 seconds) and train the network for 160k iterations based on visual result quality. The model is trained using its own predictions, letting the model learn to recover from its own errors and therefore producing more robust results.

The input style example is created in a speaker-specific manner by extracting a style encoding from one sample of each speaker, with the length M randomly spanning between 256 and 512 frames, ensuring de-correlation between styles and clip length.

Framing our model as a conditional VAE, the objective is to maximize the evidence lower bound (ELBO) of the marginal log likelihood of gesture motion given a speech sequence. The training loss can then be expresssed as the negative ELBO:

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{q(\mathbf{z}|\mathbf{e})}\left[-\log p\left(Y \mid S, \mathbf{z}\right)\right] + D_{KL}\left(q\left(\mathbf{z} \mid \mathbf{e}\right) \| p\left(\mathbf{z}\right)\right) \\ &= \mathcal{L}_{recon} + D_{KL}\left(q\left(\mathbf{z} \mid \mathbf{e}\right) \| p\left(\mathbf{z}\right)\right)\end{aligned} \tag{1}$$

The first term represents the expected negative log-likelihood of the gesture motion, i.e. the reconstruction loss. The second term is the regularization term in the form of the Kullback–Leibler divergence between the posterior distribution $q\left(\mathbf{z} \mid \mathbf{e}\right)$ predicted by the style encoder and the prior distribution $p\left(\mathbf{z}\right)$, which is a standard multivariate Gaussian distribution. For weighting the regularization term, we use cost annealing as proposed by [4].

Our reconstruction loss divides into the following terms:

$$\begin{aligned}\mathcal{L}_{recon} =& \lambda_p \mathcal{L}_p + \lambda_r \mathcal{L}_r + \lambda_{vp} \mathcal{L}_{vp} + \lambda_{vr} \mathcal{L}_{vr} + \\ & \lambda_{dp} \mathcal{L}_{dp} + \lambda_{dr} \mathcal{L}_{dr} + \lambda_f \mathcal{L}_f\end{aligned} \tag{2}$$

where $\mathcal{L}_p$, $\mathcal{L}_r$, $\mathcal{L}_{vp}$, and $\mathcal{L}_{vr}$, are the mean absolute error (MAE) between predicted and target joint positions, rotations, translational velocities, and rotational velocities, respectively, in both local and world spaces. In addition to direct velocity predictions, $\mathcal{L}_{dp}$ and $\mathcal{L}_{dr}$ penalize the velocity MAE by computing the translational and rotational velocities in the local and world spaces on-the-fly via finite-difference. We took inspiration from the reconstruction loss for Learned Motion Matching proposed by [10]. Lastly, $\mathcal{L}_f$ penalizes the MAE for the facing direction in the world space to prevent any character rotational drift. Loss terms are empirically weighted to be on a similar scale.

5 DATA PREPARATION

The GENEA challenge provided full-body motion files with synchronized audio and transcriptions from the *Talking With Hands 16.2M* conversational dataset [15]. The data contains long periods of silence, low volume audio, cross-speaker audio bleeding, speaker data with little to no hand gesture motion and noisy finger motion. All of which could hinder model training. To address this, we filter and process the data as follows:

We remove long periods of silence (> 0.5 seconds) in the recordings based on the speech transcripts. This process also removes longer sections in which an interlocutor is speaking causing audio bleeding into the recording. We normalize the relative loudness of all recordings to address low and varying volume of some audio sections. We identify sections

with low average hand velocity to remove data lacking in gesture motion. Finally, we remove finger motion data and apply normalization to the animation data.

We extract low-level features from the raw audio data. We extract spectrograms using an FFT Hanning-window of 50 ms and a hop length of 12.5 ms. We project the spectrogram into the mel frequency scale, and use the log amplitude of each of the 80 channels. We additionally extract the total energy per frame as a secondary feature. Finally we re-sample the speech features to 30 frames per second (fps) to match the frame rate of the provided motion files.

For motion data, we represent each frame by a feature vector $\mathbf{a} = [\boldsymbol{\rho}_p, \boldsymbol{\rho}_r, \dot{\boldsymbol{\rho}}_p, \dot{\boldsymbol{\rho}}_r, \dot{\mathbf{r}}_p, \dot{\mathbf{r}}_r]$ where $\boldsymbol{\rho}_p \in \mathbb{R}^{3j}, \boldsymbol{\rho}_r \in \mathbb{R}^{6j}$ are the joint local translations and rotations, $\dot{\boldsymbol{\rho}}_p \in \mathbb{R}^{3j}$ and $\dot{\boldsymbol{\rho}}_r \in \mathbb{R}^{3j}$, are the joint local translational and rotational velocities, and $\dot{\mathbf{r}}_p \in \mathbb{R}^3$ and $\dot{\mathbf{r}}_r \in \mathbb{R}^3$ are the character root translational and rotational velocity local to the character root transform. j corresponds to the number of joints. Joint rotations are represented by 2-axis rotation matrix and joint and root rotational velocities are specified using the scaled angle axis representation as in [29]. The root position of the character is computed by projecting the position of the second spine joint on the ground. The root rotation is obtained by projecting the *z-axis* of the hip joint onto the ground.

6 RESULTS

Our model produced highly competitive results in the evaluation of the GENEA Challenge 2022 [27]. For human-likeness, the full-body motion produced by our model ranked second, placing only behind ground truth and the entry ranking on-par with ground truth. Regarding appropriateness for speech, our system performed similarly to most other entries, with all entries falling significantly behind ground truth and obtaining similar scores. (refer to GENEA challenge main paper [27]). The GENEA challenge also included a number of objective, numeric measures of performance, namely average joint jerk, positional speed profile, and canonical correlation. However, scores on these measures did not correlate well with perceptual results and therefore do not appear to be informative of output quality.

7 DISCUSSION

Our system performed overall well in the challenge and showed highly competitive performance. Our method was designed for zero-shot style control by example, meaning that our model can generate gestures for unseen speakers' styles, without retraining on the new styles. There is a trade-off between increased generalization capabilities and tailoring a model for a specific task. Nonetheless our system performed overall well and showed highly competitive performance. Moreover, our model is probabilistic, allowing generation of many output variations given the same input. While not evaluated in the challenge, we believe that strong generalization capabilities and being able to generate diverse variations for the same audio clip are two highly desirable qualities for a generator model.

While ranking second and third in the human-likeness subjective evaluations, this performance was not reflected in the objective evaluations. However, by comparing the overall challenge results, it appears that these measures did not strongly correlate with perceptual evaluations across all methods. This led the organizers to argue that objective evaluation is currently not particularly meaningful for assessing gesture generation quality, and that subjective evaluations are the best option for such assessment. This is in line with the conclusions of [24]. We believe that distance metric, in their current form, will always fail to accurately capture perceptual gesture quality; as distance is measured with respect to a ground truth sequence, it is implicitly assumed that anything that deviates from it is of lesser quality, however, several gesture sequences may be equally appropriate for the same speech clip, but only those closer to a particular ground truth realization will be correctly scored.

Based on our experience participating in the challenge, we highlight the importance of data preparation. We observed significant improvement in our model performance after applying the data preparation operations described in Sec. 5.

In future work, we would like to iterate upon our data preparation methods to further improve result quality. It is likely that some non-informative segments remain despite our filtering operation. Moreover, some data augmentation techniques could be applied to audio files as well as animation clips (e.g. mirroring, speed manipulations). Given that numeric measures do not strongly correlate with perceptual quality we will investigate the effect of adversarial losses on subjective results. For instance, as done in traditional generative adversarial networks [9], a second network trained to distinguish between generated and ground truth animation could potentially guide our system in generating more realistic animations. We could also implement cycle consistency loss similarly to [30] to ensure that generated gesture accurately reproduce the style of the conditioning example.

8 CONCLUSION

We described our entry to the GENEA 2022 Gesture Generation Challenge. Our system allows for zero-shot style-control via short example motion clips. This allows for operation in limited data settings and alleviate the difficulty of devising and fine-grained taxonomy for gesture styles. In addition, our system can generate many output variations for the same input speech segment. The GENEA challenge results showed that our method produces competitive natural-looking motions when compared to other challenge entries, while offering the aforementioned capabilities. Perhaps unsurprisingly, we found that data preparation is key in obtaining good performance, and that since current objective measures do not strongly correlate with subjective evaluations, losses used during training should be revisited to better align with human perception.

REFERENCES

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 64–1.

[2] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. 2020. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *European Conference on Computer Vision*. Springer, 248–265.

[3] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 487–496.

[4] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating Sentences from a Continuous Space. In *SIGNLL Conference on Computational Natural Language Learning (CONLL)*. arXiv:1511.06349 http://arxiv.org/abs/1511.06349

[5] Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 93–98.

[6] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2019. Multi-objective adversarial gesture generation. In *Motion, Interaction and Games*. 1–10.

[7] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3497–3506.

[8] Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in Cognitive Sciences* 3, 11 (1999), 419–429. https://doi.org/10.1016/S1364-6613(99)01397-2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[10] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. 2020. Learned motion matching. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 53–1.

[11] Wei-Ning Hsu, Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, Patrick Nguyen, and Ruoming Pang. 2019. Hierarchical Generative Modeling for Controllable Speech Synthesis. In *ICLR*. arXiv:1810.07217 http://arxiv.org/abs/1810.07217

[12] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[13] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 242–250.

[14] 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*. 11–21.

[15] 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*. 763–772.

[16] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265* (2019).

[17] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)* 27, 1 (2008), 1–24.

[18] Regina Pally. 2008. A Primary Role for Nonverbal Communication in Psychoanalysis. *Psychoanalytic Inquiry* 21, 1 (2008), 71–93. https://doi.org/10.1080/07351692109348924

[19] Manuel Rebol, Christian Güti, and Krzysztof Pietroszek. 2021. Passing a Non-verbal Turing Test: Evaluatina Gesture Animations Generated from Speech. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 573–581.

[20] Michal Rolinek, Dominik Zietlow, and Georg Martius. 2019. Variational Autoencoders Pursue PCA Directions (by Accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Sinan Sonlu, Uğur Güdükbay, and Funda Durupinar. 2021. A Conversational Agent Framework with Multi-modal Personality Expression. *ACM Transactions on Graphics (TOG)* 40, 1 (2021), 1–16.

[22] Guillermo Valle-Pérez, Gustav Eje Henter, Jonas Beskow, André Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. 2021. Transflower: probabilistic autoregressive dance generation with multimodal attention. *arXiv preprint arXiv:2106.13871* (2021).

[23] Yuxuan Wang, Daisy Stanton, Yu Zhang, Rj Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. 2018. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. (2018).

[24] 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 (2022), 379–389. https://doi.org/10.1109/THMS.2022.3149173

[25] 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 (TOG)* 39, 6 (2020), 1–16.

[26] Youngwoo Yoon, Keunwoo Park, Minsu Jang, Jaehong Kim, and Geehyuk Lee. 2021. Sgtoolkit: An interactive gesture authoring toolkit for embodied conversational agents. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 826–840.

[27] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. GENEA Challenge 2022 evaluation results description v. 1.0. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI '22)*. ACM.

[28] Julian Zaïdi, Hugo Seuté, Benjamin van Niekerk, and Marc-André Carbonneau. 2021. Daft-Exprt: Robust Prosody Transfer Across Speakers for Expressive Speech Synthesis. *arXiv preprint arXiv:2108.02271* (2021).

[29] He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. 2018. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–11.

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.