CGVU: Semantics-guided 3D Body Gesture Synthesis

Kunkun Pang
University of Edinburgh
k.pang@ed.ac.uk

Taku Komura
University of Edinburgh
t.komura@ed.ac.uk

Hanbyul Joo
Facebook AI Research
hjoo@fb.com

Takaaki Shiratori
Facebook Reality Labs
tshiratori@fb.com

ABSTRACT

In this paper, we explore a novel data-driven approach which is to synthesise continuous body gesture from speech with a fully connected neural network and a periodic activation function. To produce realistic gestures that follow the context of the conversation, we make use of both low-level and high-level semantic features obtained from the speech, namely the mel spectrogram and BERT features. We participate in the Generation and Evaluation of Non-verbal Behaviour for Embodied Agents (GENEA) challenge 2020 which provide the dataset to train the proposed system and perform crowdsourced evaluation to compare different gesture generation approches performance.

CCS CONCEPTS

• **Computer methodologies → Artificial intelligence**; *Machine learning*; Computer graphics.

KEYWORDS

3D Motion Synthesis, Neural Neworks

1 INTRODUCTION

Automatic synthesis of the 3D body gesture from the speech is a challenging problem that researchers in psychology, computer graphics and computer vision have been tackling. Most classic approaches are either rule-based approaches where the corresponding gesture for each context are carefully designed based on observation, or make use of low-level features of speech such as prosody to produce movements that are well synchronized with the speech simply. We wish to go beyond such carefully designed architectures and learn a mapping from speech to the body gesture automatically from a large amount of data.

However, there are various difficulties to learn the task of body gestures synthesis from the speech. First of all, this is a cross modality learning problem that requires a significant amount of training data for producing a proper mapping. Secondly, the correlation between the speech and the gesture is rather weak. Simply regressing the low-level speech features to the gesture may easily fail due to the ambiguity of the mapping.

In this paper, we investigate a novel deep learning approach to produce the speaker's 3D body gesture from speech. For coping with the difficulty of cross modality learning, our idea is to encode each modality by low-level and high-level representations and enable the system to learn a mapping between the provided feature representation and the desired motion. More specifically, we use the mel spectrogram to describe the low-level information and depict the high-level information with contextualised BERT feature [8]. The proposed approach learns the mapping from the given speech to body gesture in an end-to-end manner, so that the model is capable to produce realistic human motion. Moreover, we propose a novel module to address the weak correlation between the speech and gesture which is to perform random sampling in latent space to produce a different gesture.

We participate in the GENEA 2020 challenge which aims to have a better understanding and compare methods for gesture generation and evaluation since the variant dataset embodiment, and evaluation methodology may lead to a different conclusion [18]. Besides, both our proposed system and other gesture generation approaches were evaluated in a large user study during GENEA 2020 challenge.

2 RELATED WORKS

Speech-to-Gesture. The correlation of speech and gesture has been a long term interest in the area of psychology [3, 25, 29]. Kendon [17] analyzes the synchronization of the speech and gesture, and find that the gesture appears even earlier than speech. McNeill [25] insists that gesture and speech are occuring from a common source. Reiter et al. [7] claims that gesture and speech are complementary to each other to convey the speaker's intention.

It has also been an area of interest in the computer animation community, where researchers are interested in animating the gesture of virtual characters during their speech. Earlier, researchers used to define rules that produce the head and body motion according to the acoustic information and the text contents [4, 5]. These further evolved into probabilistic models [26] where the motion type was sampled among a number of potential movements according to the probability. Levine et al. [20] propose a model that selects a motion unit by an HMM using the speech prosody as the feature. This idea is later enhanced to use reinforcement learning to select an optimal series of gestures [19]. Marsella et al. [23] combines the prosody with the text information for deciding which gestures to animate. Chiu et al. [6] precompute a motion manifold using GPLVM and produce a mapping from the speech to the latent space of gestures by conditional random fields. Our work is similar in sense we also produce a latent space, but we do this in an end-to-end fashion. Among these works, many researches emphasize the strong correlation of the semantic context of the speech to the body gestures [4, 5, 23, 26], which we also pursue in this paper, but in a machine learning context. Ferstl et al. [10] use deep neural networks with adversarial loss for training a network that maps the speech to gesture, but the features that are used are limited to low level prosody.

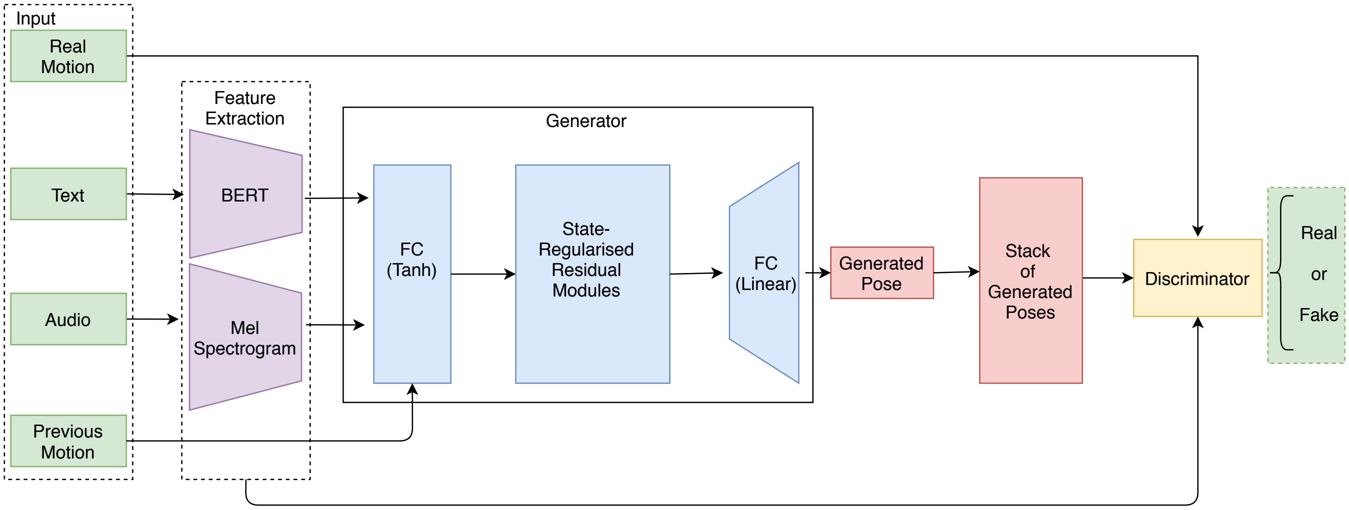

Figure 1: System Overview

The topic has recently attracted researchers in the computer vision community due to the availability of pose estimation tools with high precision. Shlizerman et al. [27] trains an LSTM to map the sound to the body gesture. Ginosar et al. [11] proposes an adversarial training model for mapping the speech features to the gestures using the in-the-wild videos. The model is trained using 144-hour person-specific video dataset of 10 speakers, where the 2D poses are automatically predicted by OpenPose [2]. There has been little deep learning based research that focuses on the synthesis of 3D human movements from the speech data. The main difficulty is in the access to a data that contains both the speech and the 3D human motion.

Generative Models. Despite both LSTM and Temporal Convolutional Network (TCN) achieve reasonable performance in time-series problem, they are still the deterministic model which principally lack the ability to solve the ambiguity between speech and gesture. Since the determinisitc model learns the one-to-one mapping but speech to gesture synthesis problem suppose to be many-to-many. One possible approch is to improve the determinisitc model to a generative model. Recently, one of the successful and powerful generative approach is generative adversarial networks (GANs) [12]. GANs have already produced impressive results in the computer vision, such as image generation, image editing, and video future prediction [16, 21, 31]. The key idea of using GANs is the adversarial loss which is to encourage the generated sample to be hardly distinguished with the real one.

In this paper, we make use of state-of-the-art features in natural language processing, speech processing and human motion synthesis for achieving this task. In addition, we use the generative adversarial network to enable the generated motion could be diverse which won't have the one-to-one mapping restriction.

3 PROPOSED SYSTEM OVERVIEW

Here, we describe the details of the proposed speech to gesture system. As illustrated in the fig 1, the entire system consists of feature extraction, motion generation, and motion discrimination. For simplicity, we synchronise pose, audio, and word for every frame in 20 fps so that the length of cross-modal data could be consistent. In training time, given all available inputs, we firstly extract the audio features with log scale mel-spectrogram and the contextualised word feature from BERT [8]. Then, the generator is modelled as an auto-regressive fully connected neural network which takes a sequence of future audio, word, and 3D previous motion as the input and output of the next frame of the pose. Finally, the discriminator receives a stack of poses to classify whether the poses are the real or generated sequences. Ideally, the generated sequences are supposed to be hardly distinguished by a human.

More specifically, we represent the generator as G and the sequences of audio, words, and pose as $\boldsymbol{a}_{t\to t+\tau}, \boldsymbol{w}_{t\to t+\tau}, \boldsymbol{x}_{t-\tau\to t}$ respectively, where the feature of every frame is defined as $\boldsymbol{a}_t \in \mathbb{R}^{D_a}$, $\boldsymbol{w}_t \in \mathbb{R}^{D_w}$, $\boldsymbol{y}_t \in \mathbb{R}^{D_y}$. Then, the generator will take the vectorised features as the input and output the next frame of pose $\boldsymbol{y}_{t+1} = G(\boldsymbol{a}_{t\to t+\tau}, \boldsymbol{w}_{t\to t+\tau}, \boldsymbol{y}_{t-\tau\to t}; \omega)$. Thus, we are learning the dynamics of predicting the future frame by conditioning on the future audio, word signal and the past motion.

3.1 Pre-processing

To extract the speech features, we use the librosa library to convert the raw audio to mel frequency power spectrogram with 27 channels [1, 24] and transform the extracted spectrogram feature with logarithm function. For the contextualised word features, we merged every single word into a sentence and extract the contextualised word feature from the BERT model's hidden space. Since the original size of the BERT representation is relatively larger than the current mel-spectrogram feature, we perform dimensionality reduction with PCA to reduce the word feature size from 512 to 32. Some frames do not cover by a word would be padded a zero vector.

To process the motion data, we use a 6 degrees of freedom representation which can contain the joint rotation information. We

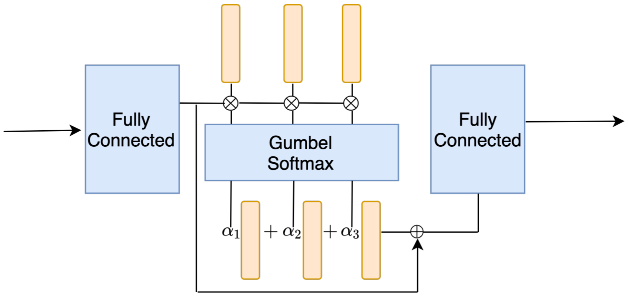

Figure 2: State-Regularized Residual Module. Orange: State.⊕ is addtion operation. ⊗: is element-wise multiplication operation.

first compute the 3×3 rotation matrix and use the first two-row as the representation for every joint individually $\boldsymbol{y}_t \in \mathbb{R}^{6D_y}$. The third vector could be obtained via the cross product.

During the training and inference, our system also learns to predict the finger joints since we believe the correlation between the fingers and wrist representation is strong. In the workshop challenge, we remove the fingers' motion during the visualisation since the challenge is mainly to evaluate and analyse the quality of the torso and body gestures.

3.2 Sinusoidal Activation Function

Similar to the previous research [28], we investigate the performance of the sinusoidal activation function with fully connected neural network in the time-series problem. We define a series of non-linear transformation as

$$\Phi(X) = \omega_n(\phi_{n-1} \circ \phi_{n-2} \ldots \circ \phi_0) + b_n, \tag{1}$$

$$x_i \longrightarrow \phi(x_i) = \sin(\omega_i x_i + b_i) \tag{2}$$

where $\phi_i : \mathbb{R}^{M_i} \mapsto \mathbb{R}^{N_i}$ consist of the affine transform with the parameter matrix $\omega_i \in \mathbb{R}^{N_i \times M_i}$, biases $b_i \in \mathbb{R}^{N_i}$ and a sinusodial activation function sin.

3.3 State-Regularised Residual Module

Inspired by the recent research to regularise the latent state transition in recurrent networks [30], we propose a stochastic state-regularised residual module for the fully connected network (SRRFC). The SRRFC is a module to combine vanilla fully connected layer, stochastic state transition, and residual operation. The vanilla fully connected layer learns a fixed mapping between the input and output. As stated above, we also use the sinusoidal activation function for every fully connected layer in the module. Regarding the stochastic state transition, we defined a set of K states as learnable parameters $\boldsymbol{s}_i \in d \times K$ where d is the size of the $i-$th hidden layer. The stochastic state transition will take the output $\boldsymbol{x}_i$ from the last fully connected layer and compute the probability $\boldsymbol{\alpha}$ for every state.

$$\boldsymbol{\alpha}_i = \psi(\boldsymbol{s}_i, \boldsymbol{x}_i) \tag{3}$$

To select the state randomly, we apply the Gumbel softmax [15] to be a differentiable approximation of the arg max operation which enables the module and network could be trained in an end-to-end manner,

$$\alpha_i^k = \frac{\exp((\boldsymbol{x}_i \cdot \boldsymbol{s}_i^k + g_k)\backslash\mathcal{T})}{\sum_k \exp(\boldsymbol{x}_i \cdot \boldsymbol{s}_i^k + g_k)\backslash\mathcal{T})} \tag{4}$$

where $g_1, g_2, \ldots, g_K$ are i.i.d drawn from the Gumbel(0,1) distrbution and $\mathcal{T}$ is the temperature of the softmax distribution. Here, we define the new states $\boldsymbol{s}_i^{\text{new}}$ as the probability mixtures of the states:

$$\boldsymbol{s}_i^{\text{new}} = \sum_k \alpha_i^k \boldsymbol{s}_i^k \tag{5}$$

After we get the new states, we are using the residual operation [13] to get the new output

$$\boldsymbol{x}_{i+1}^{new} = \phi_{i+1}(\boldsymbol{x}_i + \boldsymbol{s}_i^{\text{new}}) \tag{6}$$

The whole state-regularised residual fully connected module is illustrated at the fig 2. The intuition of the stochastic state-regularised residual module is regularising latent space should be partially similar to the during the training time so that the generated gesture should also be similar to the motion in the training set.

3.4 Entire System

We now combine all techniques described above in the whole system. For the generator, we use 1 fully connected layer with tangent activation function to rescale the data range into $[-1, 1]$. Then, we stack 5 SRRFC modules and 1 fully connected layer so that we can map the multi-modal inputs to the desired size of the pose. Regarding the discriminator, there are 3 fully connected layers with sinusodial activation function which takes the vectorised stack of poses as the input and output the label whether the given sequence is real or fake. We denote the gesture distribution as p_{data} and the joint distribution of audio and word as p_{aw}. Since we are using the Gumbel softmax which already include noise from the uniform distribution p_z.

Adversarial loss. We use the adversarial loss for training both generator G and its discriminator D. The adversarial objective function is defined as

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{\boldsymbol{y} \sim p_{\text{data}}}[\log D(\boldsymbol{y})] + \\ \mathbb{E}_{\boldsymbol{a}, \boldsymbol{w} \sim p_{\text{aw}}, \boldsymbol{y} \sim p_{data}, \boldsymbol{z} \sim p_z}[1 - \log D(G(\boldsymbol{a}, \boldsymbol{w}, \boldsymbol{y}, z))]\end{aligned} \tag{7}$$

where G is expected to generate the sequences of motion which hardly distinguish between the real

Gesture Loss. In theory, the learned mapping G is supposed to generate reasonable poses by conditioning on the speech. However, training such a mapping with adversarial loss alone cannot produce realistic result in practice [14, 31]. Here, we also introduce a mean square error to encourage the generator to match the data distribution and stabilise the training as well.

$$\mathcal{L}_{MSE}(G) = \mathbb{E}_{\boldsymbol{a}, \boldsymbol{w} \sim p_{\text{aw}}, \boldsymbol{y} \sim p_{data}, \boldsymbol{z} \sim p_z}[(\boldsymbol{y} - G(\boldsymbol{a}, \boldsymbol{w}, \boldsymbol{y}, z))^2] \tag{8}$$

Entropy Regularisation. Moreover, the generated output is supposed to vary when we are sampling from the generator with the same input. We introduce the entropy regularisation on the state selection probability $\boldsymbol{\alpha}$. Thus, we maximise the entropy regularisation to encourage the probability $\boldsymbol{\alpha}$ to be more uniform. This means that the selected state is supposed to vary when we are sampling from the state.

$$\mathcal{L}_{Entropy}(G) = -\sum_{k} \alpha_k \log(\alpha_k) \tag{9}$$

Finally, the final objective function used to train the GANs is:

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{MSE}} - \lambda_1 \mathcal{L}_{\text{Entropy}} + \lambda_2 \mathcal{L}_{\text{GAN}} \tag{10}$$

where the adversarial loss and entropy regularisation are balanced by $\lambda_1 = 0.01$ $\lambda_2 = 0.01$. The model is only trained with the Trinity College Dataset [9] using the GENEA challenge release.

3.5 Post-processing

Though we use the sequence discriminator to smoothen the generated motion, the generated motion may still have a rapid change from the last pose. We smooth every frame with fixed spacing

$$y_t = 0.5 * (y_{t-1} + y_{t+1}) + 0.5 * y_t, t = 2, 4, 6, 8, ...T \tag{11}$$

4 EXPERIMENT AND ANALYSIS

In this section, we provide the detail of the experiment setting and analyse the usefulness of the proposed system and illustrated the finding of proposed sinusoidal activation function, state regularised residual fully connected module and the adversarial training.

4.1 Experiment

Initialisation. We initialise both generator and discrminator parameters with Pytorch default setting and use 5 states for every state regularised residual modules in the generator. All of the centroids are initialised uniformly between the range $[-0.5, 0.5]$. During the training, we use the AdamW [22] optimiser to optimise both generator and discriminator with learnining rate 5e-5 and 1e-5 respectively. In addition, the weight decay is defined as $1e - 4$ to regularise the parameters and the centroids.

Training Details. We update the parameters of generator and discrminator jointly for every iteration. The relatively larger learning rate can provide a larger step of the generator so that the GANs training will be stable and converge eventually. Then, the temperature of the softmax distribution initialised as $\mathcal{T} = 3$ and using the schedule $\mathcal{T} = \max(0.5, 3 * \exp(0.1 * epochs))$ to anneal the temperature. These two networks are trained with 500 epochs and apply in the test set directly.

4.2 Analysis

Training Performance. We compared the training performance between sinusoidal and ReLU activation function with 3 different architectures: fully connected networks, state regularised and residual fully connected module, GAN.

As stated in [28], training a model with sinusoidal activation function could converge faster than the ReLU function. This phenomenon is verified in our experiment, as well. According to fig 4, the proposed sinusoidal activation function shows powerful learning ability which not only trains faster but also better on the training set. Apart from this, we also explore the state-regularised fully connected module without residual operation, but this kind of network tends to converge hardly. One of the reason is that the state regularisation may be too strong for the generalisation of the latent space, which leads the model to overfit the training set hardly. Thus, it is necessary to include the residual operation into the state regularisation fully connected module.

Results Observation. During the challenge, we explored the generated motion from the model mentioned above. Firstly, we compare the generated motion of fully connected layers with ReLU and sinusoidal activation function. The generated motion from ReLU function tends to rotate the body more frequently than the one with a sinusoidal function. This frequent body rotation is less realistic since human may prefer facing the audience when they are speaking or presenting. Secondly, we investigated the usefulness of using discriminator. The models with using sequence discriminator tend to generate over smooth motion. It sometimes makes the generated motion less vivid since the motion may perform too slow.

Diverse Motion. Moreover, we also verify the stochasticity for our system with the proposed stochastic module. We randomly draw samples from the proposed model by giving the same sequences of speech; most of the motion trajectories are different. As illustrated in fig 3, the coverage area of the generated gestures are not precisely the same. This means that our system is able to generate different motion when the system receives the same speech as the input. However, the vanilla fully connected network without state-regularised module will still be a deterministic model which can only produce one sample during the training time. Even some of the recent researches using the GANs framework for their problem, they still suffer from too low stochasticity in their generated samples [31].

Crowdsource Evaluation. GENEA 2020 challenge conduct qualitative researches to evaluate different gesture-generation approaches and human gestures. The motion quality is evaluated into two aspects: human-likeness and the appropriateness. The human-likeness is to answer the question “how human-like does the gesture motion appear?" whereas the appropriateness is to respond "what extent the gestures are appropriate for the speech?" [18].

According to the crowdsourced evaluation, our current generated motion still has a large room to improve [18]. This could be raised by the torso body moves too frequent, and the body motion is not well aligned with the given speech. In future, we will focus on these two aspects and improve the generated motion quality.

Discussion. In this challenge, we insist on using the deep model architecture with fully connected layer rather than a convolutional layer or recurrent unit for such a time series task. One of the largest benefits of using a fully connected layer is run-time speed. Not only the training become faster but also the inference could be more efficient due to the parallel computing across the time dimension. This will be helpful if the system is required to be real-time in future. Besides, the recurrent unit may suffer from the initial and unseen hidden state during the test time. Those two may heavily affect the generated motion quality in an auto-regressive setting when the time horizon becomes longer. Using the state regularised module has the potential to prevent the appearance of an unseen state. This may make the generated motion to be more stable.

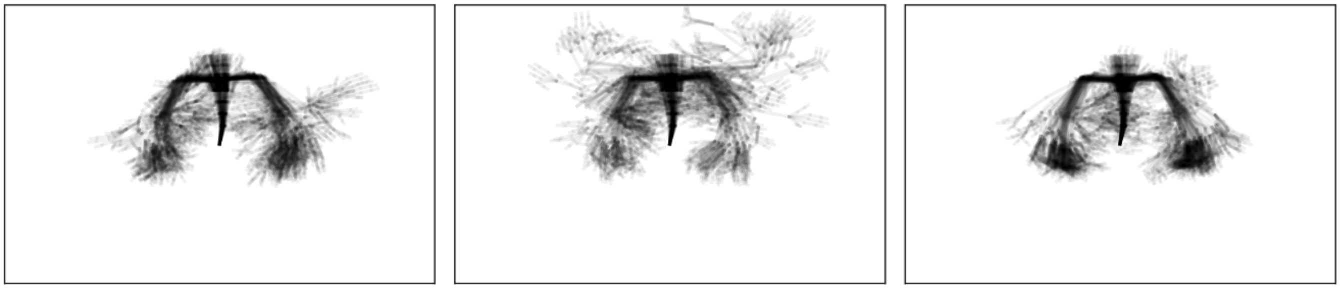

Figure 3: The generated motion trajectory with the same audio

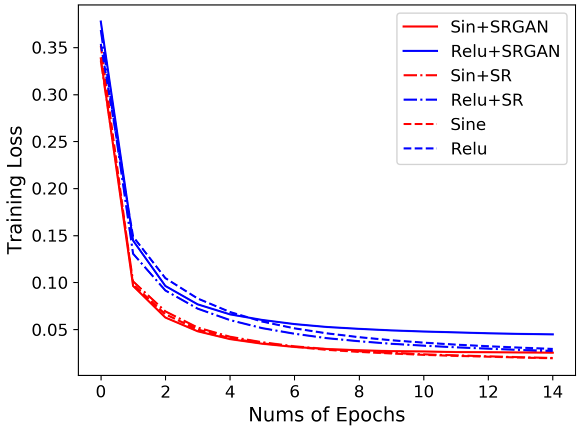

Figure 4: The comparison between the activation functions with various architectures. Sin: sinusoidal activation function. Sin+SR: the proposed framework without discriminator SRGAN: the proposed framework with discriminator

5 CONCLUSION AND FUTURE WORK

In this paper, we present a novel system to generate the motion from speech stochastically. The proposed state regularised residual fully connected module which consists of vanilla fully connected layers, state transition, and residual operation. This enables the vanilla fully connected layer to be stochastic during training and test time. However, according to the crowdsourced evaluation result, the current generated motion quality still has a large room to improve. We are planning to further improve the motion quality in terms of the human-likeness and the appropriateness since these are the primary goals for the task of motion synthesis from speech. One possible solution is adding timing information of the speech to the proposed fully connected network so that the generated motion can align with the speech frequency. Besides, we will also investigate the relation between the state transition and the generated motion. This could be another interesting angle for researchers to analyse the relationship between the synthesised motion and the speech.

REFERENCES

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

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299.

[3] Justine Cassell, David McNeill, and Karl-Erik McCullough. 1999. Speech-gesture mismatches: Evidence for one underlying representation of linguistic and non-linguistic information. *Pragmatics & cognition* 7, 1 (1999), 1–34.

[4] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. ACM, 413–420.

[5] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*. Springer, 163–185.

[6] Chung-Cheng Chiu and Stacy Marsella. 2014. Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 781–788.

[7] Jan P De Ruiter, Adrian Bangerter, and Paula Dings. 2012. The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science* 4, 2 (2012), 232–248.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *IVA '18 Proceedings of the 18th International Conference on Intelligent Virtual Agents*. https://trinityspeechgesture.scss.tcd.ie

[10] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2019. Multi-objective adversarial gesture generation. In *Motion, Interaction and Games*. ACM, 3.

[11] 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 Conference on Computer Vision and Pattern Recognition*. 3497–3506.

[12] 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 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[15] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparametrization with Gumbel-Softmax. In *Proceedings International Conference on Learning Representations 2017*. OpenReviews.net. https://openreview.net/pdf?id=rkE3y85ee

[16] T. Karras, S. Laine, and T. Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4396–4405.

[17] Adam Kendon. 1972. Some relationships between body motion and speech. *Studies in dyadic communication* 7, 177 (1972), 90.

[18] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2020. The GENEA Challenge 2020: Benchmarking gesture-generation systems on common data. In *Proceedings of the International Workshop on Generation and Evaluation of Non-Verbal Behaviour for Embodied Agents (GENEA '20)*. https://genea-workshop.github.io/2020/

[19] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture controllers. In *ACM Transactions on Graphics (TOG)*, Vol. 29. ACM, 124.

[20] Sergey Levine, Christian Theobalt, and Vladlen Koltun. 2009. Real-time prosody-driven synthesis of body language. In *ACM Transactions on Graphics (TOG)*, Vol. 28. ACM, 172.

[21] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P. Xing. 2017. Dual Motion GAN for Future-Flow Embedded Video Prediction. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 1762–1770. https://doi.org/10.1109/ICCV.2017.194

[22] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=Bkg6RiCqY7

[23] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. ACM, 25–35.

[24] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Vol. 8.

[25] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.

[26] 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), 5.

[27] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. 2018. Audio to body dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7574–7583.

[28] Vincent Sitzmann*, Julien N. P. Martel, Alexander Bergman, David B. Lindell, and Gordon Wetzstein. 2020. Implicit Neural Representations with Periodic Activation Functions. In *Proceedings of the CVPR*.

[29] Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and speech in interaction: An overview.

[30] Cheng Wang and Mathias Niepert. 2019. State-Regularized Recurrent Neural Networks *(Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, California, USA, 6596–6606. http://proceedings.mlr.press/v97/wang19j.html

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.