The StyleGestures entry to the GENEA Challenge 2020

Simon Alexanderson
simonal@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

ABSTRACT

This paper describes the StyleGestures entry to the GENEA (Generation and Evaluation of Non-verbal Behaviour for Embodied Agents) challenge 2020. The GENEA challenge is a recent initiative designed to benchmark and compare systems for gesture generation by providing a common dataset and evaluation. For this first edition of the challenge, we submitted our recently published probabilistic gesture synthesis system based on normalising flows. Only minor adjustments were made to the published system. The method takes speech audio as input and generates new gesture poses in a continuous fashion. As the method is probabilistic, a large variety of gestures can be sampled from the same speech audio. The system can be trained end-to-end and requires no manual annotation. We were pleased to see that our system ranked as one of the top two systems in both challenge evaluations, even though we only used audio as input and did not exploit the text transcriptions.

CCS CONCEPTS

• **Computing methodologies** → **Motion capture**; **Animation**; *Neural networks.*

KEYWORDS

Gestures synthesis, Data-driven animation, Probabilistic neural networks

1 INTRODUCTION

In the GENEA 2020 challenge we participated with a system called StyleGestures, based on our recent publication [2]. StyleGestures was developed at the division of Speech, Music and Hearing at KTH Royal Institute of Technology, with a long tradition of research in speech technology and multimodal communication.

One of the difficulties for speech driven gesture generation comes from the weak coupling between speech and gestures. During natural speech, any same utterance will typically be accompanied by very different gestures from time to time. Deterministic methods fail to capture this variation and tends to collapse to some average gesture. StyleGestures tackles this challenge by adapting a novel probabilistic modelling technique called MoGlow [5]. MoGlow is an autogregressive sequence model that uses normalizing flows to model the probability distribution of the next pose in a pose-sequence. Once initialised, new poses are generated sequentially by random sampling from the model. While MoGlow is completely general and makes no assumption of the nature of the motion or control, StyleGestures was developed to investigate the specific case of speech-driven gesture synthesis. As described in [2], StyleGestures allows for generation of full-body gestures (including stepping motion and stance shifts) and also supports an optional control over

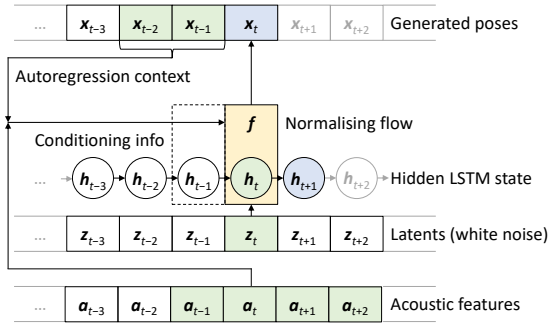

Figure 1: System overview. Green elements are inputs, blue outputs.

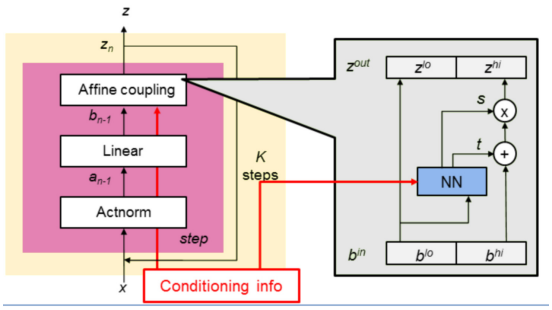

Figure 2: Structure of the normalising flow during inference. During synthesis the arrows are reversed.

gesturing *style* (involving e.g. gesture speed, spatial extent or symmetry). In this entry, we restricted the method to synthesising upper body motion without style control.

2 SYSTEM OVERVIEW

The architecture behind our GENEA entry follows closely that of [2]. In this section we give a brief overview of the architecture and its general concepts. We encourage interested readers to delve into the details in the original paper, as well as the underlying MoGlow publication [5].

2.1 Network architecture

An overview of our system is shown in Figure 1. Green elements are inputs and blue outputs. The idea is to treat motion as a stochastic sequence of poses and to model the next-step probability distribution conditioned on an autoregressive context of past poses and an external control signal. In our case, the control signal is a time slice of past and future speech, and the probability distribution

is modelled with a normalising flow. During synthesis, the conditioning information is concatenated to a single vector and fed into the normalising flow together with a drawn random sample from a latent Gaussian distribution. These inputs are transformed by the flow into the next-step pose, and synthesis proceeds to the next time-step.

Normalising flows are invertible neural networks capable of transforming a complex distribution $\boldsymbol{x}$ to a simple base distribution $\boldsymbol{z}$ through a series of invertible non-linear transformations, so called flow-steps, see Figure 2. As the transformations are invertible, samples from the complex distribution can be generated by reversing the flow. Each flow-step consists of three sub-steps: activation normalisation, a linear transform and an affine coupling layer. While the first two operations are linear and acts to normalise and permute variables between flows, the affine coupling layer is non-linear and passes variables through a neural network (in our case an LSTM). When multiple flows are stacked, these operations can yield very powerful transformations. The trick to make the coupling layer invertible is to affinely transform only half of the variables based on the second half, which are passed unchanged (Figure 2 right). Upon inversion, the unchanged variables can be used to reverse the transformation. The conditioning information is fed into the affine coupling layer. During training, the normalizing flow is optimised by maximising the likelihood of z.

3 DATA PREPARATION AND TRAINING

We used the supplied challenge data for model training and synthesis [3]. The data contains motion capture and audio recordings of a single actor talking spontaneously on different topics. The actor moves freely around the motion capture space, taking small steps back and forth and changing stance. We used the data in its original form without manual annotation or cleaning. Although allowed by the challenge, we did not explore the possibilities of pre-training our model on external data.

We extracted 27 mel-spectrogram features from the audio as input to our system, and 45 joint angle features from the motion data as output. The joint angles were expressed as exponential maps [4].

3.1 Networks and training

In addition to our base StyleGestures implementation[1], we prepared two systems for comparison and possible improvements. In the first system, we replaced the LSTM in the affine coupling layer with a GRU. This was made in attempt to speed up training time and reduce the amount of parameters. In the second system, we changed the latent $\boldsymbol{z}$-distribution from a Gaussian to a t-distribution, turning the flow into a *studentising flow* [1]. This was in attempt to make the model training robust to outliers arising from noisy motion capture data or uncommon movements. For this system, we omitted gradient clipping, and set the degrees of freedom of the t-distribution to $\nu = 50$.

All systems were trained using the same hyperparameters. We used $K = 16$ flows and $H = 512$ units in the recurrent networks (each containing 2 layers). The models were trained using the Adam optimizer [6] and a Noam learning rate scheduler [9] with 3k steps of warm-up and peak learning rate 15^{-3}. The number of past context frames was set to 5 (0.25s) for both motion and speech, and the number of future frames to 20 (1s) for the speech. These hyperparameters differs slightly to those in [2]. Most notably, H was decreased from 800 to 512, which is consistent to the full-body (FB) systems in the paper. We have found that the lower value of H leads to a more diverse gesture behaviour. We held out one of the sessions ("Recording 1") for network tuning.

We subjectively assessed the output of the three systems by inspecting random samples from the three models. Unfortunately, we did not have time to perform a full-scale perceptional study. As we could not identify any salient differences between the systems, we chose to use the base model as our final system. We re-trained it using all sessions (including the one previously held out) and employed it to generate gestures from the challenge test data. During synthesis we initialized the poses in the autoregressive context with a static mean pose, and padded the audio features with one second of silence in the end. This was done to be able to generate the last second of gestures. Note that these actions causes the beginning and end of each generated gesture sequence to be of lesser quality. We did not apply any smoothing or other post-processing to the generated motion.

4 RESULTS

The challenge organization conducted a detailed subjective evaluation comparing all submitted systems. The evaluation comprised a human likeness study to assess motion quality, and an appropriateness study to assess how well the gestures match the speech. The evaluations were performed on an online crowd sourcing platform (Prolific), where the participants were asked to rate video stimuli on a 100-point scale. All systems showed the same 3D avatar. Submitted systems were presented page-wise side-by-side along with two baseline systems [7, 10] (labeled BA and BT) as well as a system with natural recordings (labeled N) and a system with miss-matched natural gesture and speech (labeled M). StyleGestures is labeled SC. For more details about the evaluation studies, please refer to the challenge paper [8].

4.1 Human likeness study

In the humans likeness study, the participants were presented with muted video stimuli and asked to consider the question “How human-like does the gesture motion appear?”. StyleGestures received a median score of 57 (mean 55.8), which ranked second among the participating systems. The difference to the higher ranked entry (SD) was not significant, nor were the differences to the two lower ranked systems (BT and SB). Bar plots and significance comparisons are shown in Figure 3.

4.2 Appropriateness study

In the appropriateness study, the participants were shown video stimuli accompanied with speech audio. Ratings were based on the question “How appropriate are the gestures for the speech?”. Unlike the human likeness study, this study contained the system with miss-aligned speech and gestures (labeled M).

[1] https://github.com/simonalexanderson/StyleGestures

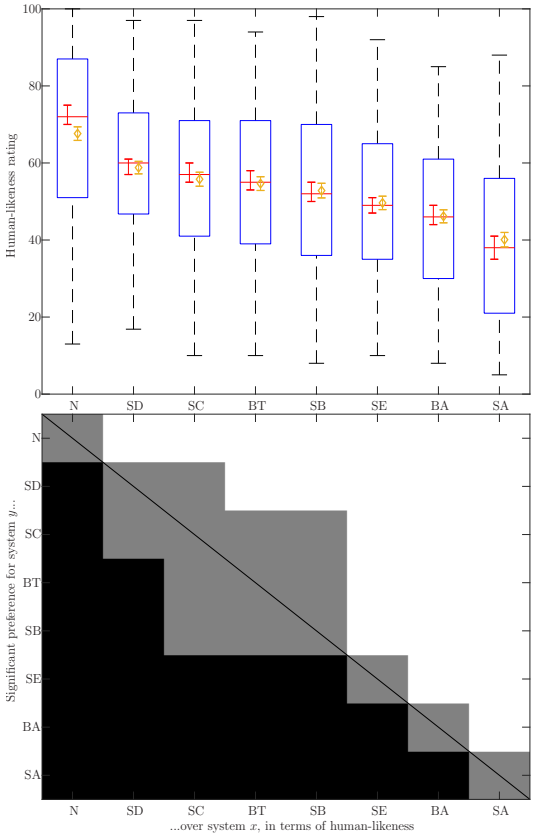

Figure 3: Boxplots (top) and significance of differences (bottom) for the human likeness study. In the latter, white means that system y rated significantly higher system x, black means the opposite, and grey means no significant difference (at the 0.01 level after Holm-Bonferroni correction).

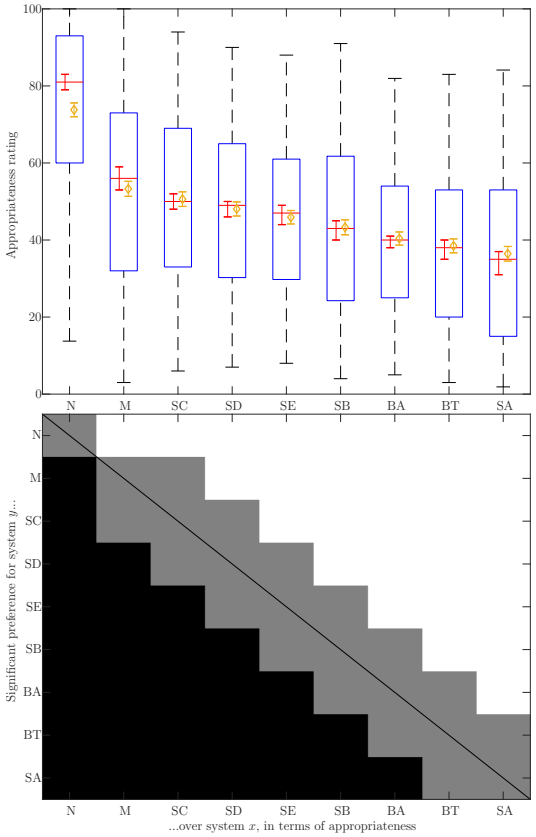

Figure 4: Boxplots (top) and significance of differences (bottom) for the appropriateness study. In the latter, white means that system y rated significantly higher system x, black means the opposite, and grey means no significant difference (at the 0.01 level after Holm-Bonferroni correction).

StyleGestures received a median score of 50 (mean 50.6) which was highest among the participating systems. The statistical analysis showed that the ratings were significantly above all participating systems but system SD. Interestingly, the miss-matched system M was higher ranked than all synthesis systems. This may be explained by the high gesture rate and low amount of pauses in the dataset in combination with the known fact that the temporal alignment between speech and gesture is not exact. Our system was the only one not rated significantly lower than M. Bar plots and significance comparisons are shown in Figure 4.

4.3 Joint visualisation

A comprehensive visualisation comparing all systems is shown in Figure 5. Here, each system is represented as an ellipse, and ordered according to median rating. Overlapping ellipses means that the conditions were not statistically significantly different at the 0.01 level after Holm-Bonferroni correction.

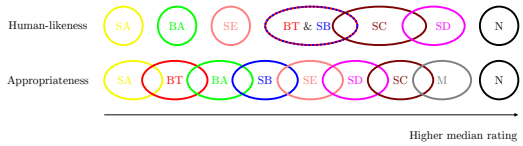

Figure 5: Significance of differences between conditions in the two studies.

5 DISCUSSION

Even though we did not make use of the text transcriptions provided with the challenge, we can observe that our system compares favourably to most of the others. Regarding that the challenge data consisted of spontaneous monologues with a high proportion of beat gestures, we think that our model in its current form is well suited. We stress that nothing in our model prevents the use of semantic features in future experiments.

A benefit that may contribute to our high naturalness ranking is our system's ability to generate other motion than hand-gestures, for example head motion and arm swinging. As our model does require any labels to differentiate gestures from other types of

motion, such movements are naturally replicated from the training data.

Taking a closer look at data quality, we think that improvements could be made by data cleanup. Especially, we found that the skeleton in the GENEA dataset has more raised shoulders in some sessions than others. We have previously found that inconsistencies in data severely affects model training and generation. For example, discontinuities in joint angles representation due to a poor choice of reference pose caused early models to generate very jerky motion. In this regard we hoped our experiments with robust model training using studentising flows would have given a clearer result. Although the results showed better likelihood scores on held-out validation data (similar to the training curves seen in [1]), it was hard to assess any improvements visually. In lack of a conclusive result, we chose to use our base model as our challenge entry.

6 CONCLUSIONS AND FUTURE WORK

We have described our entry to the GENEA challenge 2020, which closely followed our recent publication [2]. We look forward to improving the system for future GENEA challenges, for example by pre-training the model on external data, or exploiting text features and language models.

REFERENCES

[1] Simon Alexanderson and Gustav Eje Henter. 2020. Robust model training and generalisation with Studentising flows *(INNF+'20)*, Vol. 2. Article 15, 9 pages. https://arxiv.org/abs/2006.06599

[2] 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.

[3] 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

[4] F. Sebastian Grassia. 1998. Practical parameterization of rotations using the exponential map. 3, 3 (1998), 29–48. https://doi.org/10.1080/10867651.1998.10487493

[5] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2020. MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics* 39, 4 (2020), 236:1–236:14.

[6] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. ICLR*.

[7] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. [n.d.]. In *Proceedings of the ACM International Conference onIntelligent Virtual Agents (IVA '19)*.

[8] 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/

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need *(NIPS'17)*. Curran Associates, Inc., Red Hook, NY, USA, 5998–6008. https://papers.nips.cc/paper/7181-attention-is-all-you-need

[10] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '19)*. 4303–4309. https://doi.org/10.1109/ICRA.2019.8793720