EUROGRAPHICS 2020 / U. Assarsson and D. Panozzo
(Guest Editors)

Volume 39 (2020), Number 2

Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows

Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko and Jonas Beskow

Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

Abstract

Automatic synthesis of realistic gestures promises to transform the fields of animation, avatars and communicative agents. In off-line applications, novel tools can alter the role of an animator to that of a director, who provides only high-level input for the desired animation; a learned network then translates these instructions into an appropriate sequence of body poses. In interactive scenarios, systems for generating natural animations on the fly are key to achieving believable and relatable characters. In this paper we address some of the core issues towards these ends. By adapting a deep learning-based motion synthesis method called MoGlow, we propose a new generative model for generating state-of-the-art realistic speech-driven gesticulation. Owing to the probabilistic nature of the approach, our model can produce a battery of different, yet plausible, gestures given the same input speech signal. Just like humans, this gives a rich natural variation of motion. We additionally demonstrate the ability to exert directorial control over the output style, such as gesture level, speed, symmetry and spacial extent. Such control can be leveraged to convey a desired character personality or mood. We achieve all this without any manual annotation of the data. User studies evaluating upper-body gesticulation confirm that the generated motions are natural and well match the input speech. Our method scores above all prior systems and baselines on these measures, and comes close to the ratings of the original recorded motions. We furthermore find that we can accurately control gesticulation styles without unnecessarily compromising perceived naturalness. Finally, we also demonstrate an application of the same method to full-body gesticulation, including the synthesis of stepping motion and stance.

CCS Concepts

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

Keywords: Gestures, Motion capture, Data-driven animation, Character control, Probabilistic models

1. Introduction

The ability to automatically synthesise gestures is a key endeavour to provide compelling and relatable characters for many applications including animation, crowd simulation, virtual agents and social robots. This has however proved to be a particularly difficult problem. A major challenge is the lack of coherence in gesture production – the same speech utterance is usually accompanied by different gestures from speaker to speaker and time to time. Previous rule-based or deterministic methods fail to model this massive variation. Data-driven regression techniques minimising a mean square error instead lead to "average" gestures that are unlikely to be seen in real life. In order to model realistic motion, we need to move from *deterministic* to *generative* models that are capable of modelling the full space of plausible motion.

In this paper, we present a probabilistic generative model for speech-driven gesture synthesis that builds upon recent work on normalising flows with autoregression, especially MoGlow [HAB19]. The model can be trained on large sets of unstructured motion data without any need for manual labelling. Instead of directly regressing motion from speech, we train our system to model the *conditional probability distribution* of the motion given speech as input. Novel gestures can be sampled repeatedly from the probability distribution, yielding different but plausible gestures every time. This is not only consistent with human behaviour, but also provides great benefits for applications in virtual agents and animation. For virtual agents, the non-deterministic nature gives a rich set of gestures making the interaction more varied. For off-line applications, animators can effortlessly generate several gesture examples and then pick the one that best suits their scenario.

For any motion synthesis it is desirable to control or modify the style of the output motion. In gesture synthesis, use cases include artistic control over gesturing style to match a desired personality or mood, or automatic control over, e.g., gesture- or gaze direction. Research has found that motion statistics like average gesture velocity, spacial extent and height are correlated with the perception of personality traits [SN17, CN19, KG10]. Such statistics are easily extracted from unstructured motion data in data-driven scenar-

ios. In this paper, we show that our probabilistic model is ideally suited for this kind of style control. As a demonstration, we automatically extract control parameters for average gesture speed, height, spacial extent and lateral symmetry, and then generate output gestures conditioned on this control. Evaluations show that the synthetic gestures obey the style control without unnecessarily sacrificing naturalness. Finally, we show that our method not only is capable of generating upper body gestures, but also generalises to the full body, including stance shifts, stepping and giving full-body emphasis to prominent words. For this scenario, we additionally demonstrate control over character location and direction.

Our proposed method has the benefits of requiring no manual labelling, being non-deterministic (yielding unlimited gesture variation), and being able to output full-body gestures. The contributions of our paper include 1) adapting MoGlow to speech-driven gesture synthesis, 2) adding a framework for high-level control over gesturing style, and 3) evaluating the use of these methods for probabilistic gesture synthesis. Videos can be found in the supplement as well as with the code at github.com/simonalexanderson/StyleGestures.

2. Related work

Gestures are essential to human non-verbal communication. McNeill [McN92] categorises co-speech gestures into *iconics*, *metaphorics*, *beats*, *deictics* and *emblems*. Out of these categories, we focus on beat gestures, since our model only takes acoustic features as input, while appropriately triggering gestures in the other categories requires a higher degree of speech understanding.

Synthesis of body motion and, in particular, gestures has recently shifted from rule-based systems – comprehensively reviewed in [WMK14] – towards data-driven approaches. Below, we discuss only data-driven methods, since we continue this line of research.

2.1. Data-driven human body-motion generation

Several recent works have used neural networks to generate body-motion aspects such as *locomotion* [HHS*17, HKS17, HAB19], *lip movements* [SSKS17] and *head motion* [GLM17, SB18]. A challenge in these domains is the large variation in the output given the same control. Different approaches have been employed to overcome this issue. For locomotion synthesis, studies have leveraged constraints from foot contacts to simplify the problem [HSK16, HKS17, HHS*17]. Unfortunately, this is not applicable to speech-driven gestures. Closer to our domain is speech-driven head-motion synthesis, where Greenwood et al. [GLM17] apply a conditional variational autoencoder (CVAE) while Sadoughi & Busso [SB18] use conditional generative adversarial networks, but these methods have not been evaluated for gesture synthesis.

2.2. Deterministic and probabilistic gesture generation

Like body motion in general, data-driven methods are on the rise in gesture generation. Levine et al. [LKTK10] used an intermediate state between speech and gestures and a hidden Markov model to learn the mapping. They selected motions from a fixed library, which limits the range of gestures their approach can generate. Our model, in contrast, is capable of generating unseen gestures.

Recently, Hasegawa et al. [HKS*18] designed a speech-driven neural network capable of producing 3D motion sequences. Kucherenko et al. [KHH*19] extended this work to incorporate representation learning for the motion, achieving smoother gestures as a result. Yoon et al. [YKJ*19] meanwhile used neural-network sequence-to-sequence models on TED-talk data to map text transcriptions to 2D gestures. Some recent works used adversarial loss terms in their training to avoid mean-collapse, while still remaining deterministic [FNM19, GBK*19]. In another recent work, Ahuja et al. [AMMS19] conditioned pose prediction not only on the audio of the agent, but also on the audio and pose of the interlocutor. All these methods produce the same gesticulation every time for a given input, while our method is probabilistic and can produce different gestures for the same input through random sampling.

Several researchers have applied probabilistic methods to gesture generation. For example, Bergmann & Kopp [BK09] applied a Bayesian decision network to learn a model for generating iconic gestures. Their approach is a hybrid between data-driven and rule-based methods because they have rules, but they learn them from data. Chiu & Marsella [CM11] took a regression approach: a network based on restricted Boltzmann machines (RBMs) was used to learn representations of arm gesture motion, and these representations were subsequently predicted based on prosodic speech-feature inputs by another network also based on RBMs. Later, Chiu et al. [CMM15] proposed a method to predict co-verbal gestures using a machine learning model which is a combination of a feed-forward neural networks and Conditional Random Fields (CRFs). They limited themselves to a set of 12 discrete, pre-defined gestures. Sadoughi & Busso [SB19] used a probabilistic graphical model for mapping speech to gestures, but only experimented on three hand gestures and two head motions. We believe that methods that learn and predict arbitrary movements, like the one proposed herein, represent a more flexible and scalable approach than the use of discrete and pre-defined gestures.

2.3. Style control

Control over animated motion can be exerted at different levels of abstraction. while animators and actors have explicit control over motion, it is often of interest to control higher-level properties that relate to how they are perceived. The relation between low-level motion and these properties has been extensively studied. Studies have uncovered a significant correlations between statistical properties of the motion (such as gesticulation height, velocity and spatial extent) and the perception of personality along the Big Five personality traits [Lip98, KG10, SN17] and emotion [NLK*13, CN19]. In particular, Smith & Neff [SN17] modify statistical properties of existing gestures and demonstrate that these modifications create distinctly perceived personalities. Normoyle et al. [NLK*13] used motion editing to identify links between motion statistics and the emotions and emotion intensities recognised by human observers.

Another line of research considers how to use machine learning to modify motion expression, based not on emotional categories or low-level statistics but on transferring stylistic properties from other recordings onto the target motion [HPP05, XWCH15, HHKK17, SCNW19]. This is known as *style transfer*. Style can also be controlled in some underlying parameter space. Aristidou et

al. [AZS*17] present a system to modify emotional expression (valence and arousal) of a given dance motion, while Brand & Hertzmann [BH00] jointly synthesise both style and choreography without motion as an explicit input. In our work, we similarly pursue the synthesis of novel motion with continuous and instantaneous control of expression. Our approach is agnostic to the level of abstraction of the desired control space and we refer to this broadly as *style control* although our experiments are limited to controlling mid-level statistical correlates of the motion.

2.4. Probabilistic generative sequence models

This sub-section reviews probabilistic models of complex sequence data, especially multimedia, to connect our method of choice – MoGlow [HAB19] – to related methodologies and applied work.

Early works on probabilistic human locomotion modelling investigated Gaussian process dynamical models [WFH08], along with their predecessors GP-LVMs [GMHP04, LWH*12], as approaches that combined autoregressive aspects with a continuous-valued hidden state. In this work, we will model pose sequences using a similarly autoregressive approach that incorporates recurrent neural networks (RNNs) for the hidden state. Unlike approaches like [GH00], where dynamics are linear if the hidden state is fixed, we use so-called "deep autoregression" [WTY18], which has produced impressive results in diverse problems such as generating intonation [WTY18], locomotion [HAB19] and video [KBE*20].

To escape inflexible distributional assumptions, variational autoencoders (VAEs) [RMW14, KW14] can generate samples from more complex distributions by incorporating an unobservable (latent) variable. Lately, generative adversarial networks (GANs) [GPAM*14, Goo16] – another deep-learning method using a latent variable – have been the state-of-the-art in, e.g., natural image generation [BDS19]. Especially notable for this paper are applications of GANs to synthesising speech-driven head motion [SB18] and video of talking faces [VPP19,PAM*18,PWP18]. While GANs have been found to be capable of producing highly convincing random samples, they are notoriously difficult to train [LKM*18].

In this work, we will use *normalising flows* [KD18,PNR*19] for speech-driven gesture generation. Flows have gained interest since they have the same advantage as GANs of generating output by non-linearly transforming a latent noise variable, but by using a reversible neural network to do this it becomes possible to compute and maximise the likelihood of the training data, just like in classical probabilistic models like GMMs. Recent work has shown that normalising flows successfully can generate complex data such as natural images [KD18, CBDJ19], audio waveforms [PVC19] and motion data [HAB19] with impressive quality. This paper builds on the latter work by adapting it to gesture generation.

3. Method

This section introduces normalising flows and how they can be used to model speech-driven gesticulation. We will use underline to signify sequences, bold type for vectors, and non-bold type for scalars, including vector elements. Random variables and limits of summation are written in upper case, with lower case denoting specific distribution outcomes or indexing operations.

3.1. Normalising flows and Glow

The idea of the motion models in this paper is to learn the multidimensional next-step distribution of poses $\boldsymbol{X}$ in a stationary autoregressive model of pose sequences $\underline{\boldsymbol{x}} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]$ using normalising flows [PNR*19]. The latter are a general technique for representing a large variety of continuous-valued distributions $p(\boldsymbol{x})$ in a manner that allows both efficient inference (probability computation) and efficient sampling from the distribution. The idea is to describe a complicated distribution $\boldsymbol{X}$ on $\mathbb{R}^D$ as an invertible nonlinear transformation $\boldsymbol{f} : \mathbb{R}^D \to \mathbb{R}^D$ of a simple latent input distribution $\boldsymbol{Z}$, here a standard normal distribution $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}; \boldsymbol{I})$, a setup that resembles the generator structure used in many contemporary GANs. Normalising flows then construct the transformation $\boldsymbol{f}$ by chaining together a number of simpler invertible sub-transformations $\boldsymbol{f}_n : \mathbb{R}^D \to \mathbb{R}^D$, colloquially called "flows", such that the overall transformation and its intermediate results can be written

$$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{f}_1(\boldsymbol{f}_2(\ldots \boldsymbol{f}_N(\boldsymbol{z}))) \tag{1}$$

$$\boldsymbol{z}_n(\boldsymbol{x}) = \boldsymbol{f}_n^{-1}\left(\ldots \boldsymbol{f}_1^{-1}(\boldsymbol{x})\right), \tag{2}$$

where $\boldsymbol{z}_0(\boldsymbol{x}) = \boldsymbol{x}$ and $\boldsymbol{z}_N(\boldsymbol{x}) = \boldsymbol{z}$. The probability of any given datapoint $\boldsymbol{x}$ under the full distribution $\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{Z})$ can then be using the chain rule, and depends on the prior probability $p_{\boldsymbol{Z}}\left(\boldsymbol{f}^{-1}(\boldsymbol{x})\right)$ and the log-determinants of the Jacobian matrices $\partial \boldsymbol{z}_n / \partial \boldsymbol{z}_{n-1}$ of the sub-transformations $\boldsymbol{f}_n^{-1}$ at $\boldsymbol{z}_n(\boldsymbol{x})$. One can use this straightforward computation to tune the transformations $\boldsymbol{f}^n$ to maximise the exact log-likelihood of the training data using gradient-based methods.

The central design challenge of normalising flows is to devise a parametric family of $\boldsymbol{f}_n$-transformations that are flexible yet invertible, differentiable and has fast-to-compute Jacobian determinants. Recently, Kingma & Dhariwal [KD18] introduced a particular choice of $\boldsymbol{f}_n^{-1}$ called *Glow*, and demonstrated impressive results for synthesising facial images. Each flow in Glow consists of three sub-steps, of which two are learned affine transformations while the third step, called an affine coupling, is a an invertible nonlinear transformation whose parameters are determined by a neural network. Each sub-step has a Jacobian log determinant that is a simple sum of D terms that readily arise during the computations.

3.2. MoGlow for gesture generation

MoGlow [HAB19] extends Glow to the problem of modelling and generating motion, by using Glow to describe the next-step distribution in an autoregressive model. It also adds control over the output and uses recurrent neural networks for long-term memory across time. To make the Glow transformations conditional on other information, such as the previous poses $\boldsymbol{x}_{t-\tau:t-1}$ and a current control signal $\boldsymbol{c}_t$, MoGlow simply feeds this additional conditioning information into all neural networks in the system (i.e., the affine coupling layers), similar to [PVC19]. The resulting autoregressive sequence-to-sequence model can be written

$$p_{\underline{\boldsymbol{X}}\,|\,\underline{\boldsymbol{C}}}(\underline{\boldsymbol{x}}\,|\,\underline{\boldsymbol{c}}) = p_{\boldsymbol{X}_{1:\tau}}(\boldsymbol{x}_{1:\tau}) \cdot \prod_{t=\tau+1}^{T} p_{\boldsymbol{X}_t\,|\,\boldsymbol{X}_{t-\tau:t-1},\boldsymbol{C}_t}(\boldsymbol{x}_t\,|\,\boldsymbol{x}_{t-\tau:t-1}, \boldsymbol{c}_t, \boldsymbol{h}_t) \tag{3}$$

$$\boldsymbol{h}_{t+1} = \boldsymbol{g}(\boldsymbol{x}_{t-\tau:t-1}, \boldsymbol{c}_t, \boldsymbol{h}_t). \tag{4}$$

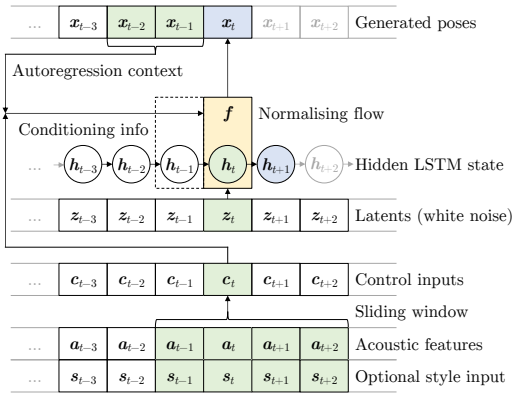


Figure 1: *Autoregressive speech-driven gesture generation. The yellow box is the MoGlow-based next-step distribution $p_{\boldsymbol{X}_t|\boldsymbol{X}_{t-\tau:t-1},\boldsymbol{C}_t}$. Inputs to the synthesis are coloured green while outputs (the next pose and the unobserved LSTM state) are blue.*

We assume stationarity, meaning that $p_{\boldsymbol{X}_t|\boldsymbol{X}_{t-\tau:t-1}\boldsymbol{C}_t}$ and $\boldsymbol{g}$ do not depend on t. Eq. 4 represents the (hidden) LSTM-state evolution. In this work, we will use a sequence of neutral (mean) poses for the initial motion $\boldsymbol{x}_{1:\tau}$, although many other choices are possible.

For speech-driven gesture generation, the control information $\boldsymbol{c}_t$ will be a sub-sequence excerpted from an acoustic feature sequence $\underline{\boldsymbol{a}} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_T]$ time-aligned with $\underline{\boldsymbol{x}}$. [HAB19] found it necessary to apply data dropout to the poses in the autoregressive inputs $\boldsymbol{x}_{t-\tau:t-1}$ to the next-step distribution, as models learned without such dropout were found not to respect the other control inputs $\boldsymbol{c}_t$.

While the original MoGlow focussed on locomotion control with zero algorithmic latency, this is not a good match for speech-driven gesture generation. Human gestures that co-occur with speech are segmented into preparation, stroke, and a retraction phase. In order to synchronise gestures with speech (e.g., perform beat-gestures concurrently with prosodic emphasis in the acoustic features), the gestures must be prepared in advance. For this reason, we let the control inputs $\boldsymbol{c}_t$ at time instance t contain not only the current speech features $\boldsymbol{a}_t$, but also a window of surrounding speech features $\boldsymbol{a}_{t-\tau:t+r}$, where the lookahead r is set so that a sufficient amount of future information can be taken into account. Subjectively, we found one second to be sufficient, but not 0.5 s. The full motion-generation procedure is visualised in Fig. 1.

In addition to letting gestures depend on speech, one may wish to exert further control over the style or other properties of the gesticulation. We propose to add such style-control input values $\boldsymbol{s}_t$ alongside the speech-feature inputs $\boldsymbol{a}_t$, as seen in Fig. 1, in order to train a style-controllable gesture-generation system. By appending control vectors to each time frame, we allow control inputs to change over time with the same granularity as the output motion. In Sec. 4.2 we explore a few scalar control schemes that modify meaningful properties of the gesticulation such as gesture radius and height.

4. System setup and training

4.1. Training-data processing

For the experiments, we trained and tested our system on the Trinity Gesture Dataset (available at trinityspeechgesture.scss.tcd.ie), which is a large database of joint speech and gestures collected by Ferstl et al. [FM18]. The data consists of 244 minutes of motion capture and audio of one male actor speaking spontaneously on different topics. The actor's movements were captured with a 20-camera Vicon system and solved to a skeleton with 69 joints. The actor moved freely around the capture area, so gestures were generally performed while shifting stance or taking a few steps back and forth. The spontaneous setting caused a large number of speech disfluencies and fillers, but there are remarkably few silent pauses in the data despite its spontaneous nature.

To process the motion data, we initially synchronised audio and video and downsampled all recordings to a consistent rate of 60 fps. We then rotated the motion-capture joint angles to be expressed relative to a T-pose and transformed them to an exponential map representation, to obtain features without discontinuities. We then removed all the root- and lower-body motion, keeping only 15 upper-body joints, from the first spine joint up to and including the hands and head. Finger motion was removed due to poor data quality.

The audio signal was transformed to 27-channel mel-frequency power spectrograms. (We also experimented with MFCC features, but did not find any notable differences in training loss or subjective quality.) To obtain inputs and outputs, we further downsampled the data to three times as much material at 20 fps (using frames $t = 0, 3, 6, \ldots$, and $t = 1, 4, 7, \ldots$, and $t = 2, 5, 8, \ldots$) and sliced it into 80 frame-long (4 s) time-series excerpts with 50% (2 s) overlap. This resulted in 20,665 samples of data, each with 80×27 speech features as input and 80×45 joint angle features as output. One session, *NaturalTalking_007*, was held out from training and cut into two parts: the first 4000 frames (200 s) for validation and network-tuning, and the last 8000 frames (400 s) for system evaluation, cut into 19 non-overlapping segments of equal length.

Finally, we augmented the data – but only for the proposed systems – with a mirrored version of the joint angles together with the unaltered speech. This is because the proposed systems are autoregressive and thus take past poses as part of the input when generating a new pose, the other systems in Sec. 4.4 utilise speech input exclusively. For these systems we found that our data augmentation resulted in only perfectly symmetric gestures being generated.

4.2. Style-control data

In order to demonstrate style control, we decided to focus on style-correlated aspects of the gesticulation that can be computed from pose sequences alone, without manual annotation. Specifically, since hand motion is central to speech-driven gestures, we studied control over various aspects of the motion of the wrist joints (whose positions we computed, in hip-centric coordinates, using forward kinematics). This joint position data was then used to calculate the hand height (right hand only), the hand speed (sum of left and right hands) and the gesticulation radius (the sum of the hand distances to the up-axis through the root node). Each of these three quantities were then averaged using a four-second sliding window and the resulting, smoothed time-series used as an additional input $\boldsymbol{s}_t$ to train style-controllable model as in Sec. 4.4. In addition, we also computed the correlation between right and left hand movements (mirrored along the x-axis) across 4 s sliding windows, to enable learning of control over the symmetry of generated gestures.

We note that our style-control approach is highly general: If it is possible to associate each frame in the data with a feature or style vector (which may vary for each time t or be constant per speaker, recording, etc.), this can be used to train a system with style input to the synthesis; the four lower-level style attributes discussed here are only intended as examples.

4.3. Network tuning and training

Starting from the hyperparameters of MoGlow for locomotion [HAB19], we first tuned model complexity (i.e., the number of flow-steps K and units H in the affine coupling LSTM layers), followed by the data-dropout probability and finally the learning rate. Model-complexity parameters were tuned with grid search, where $K = 16$ and $H = 800$ were chosen based on training-data likelihood and speed of computation. (We were not able to tune the model using subjective impressions, since the differences between similar systems were too small to be noticeable). We used the Adam optimizer [KB15] with Noam learning rate decay and tuned the maximum and minimum learning rate by incrementally scaling up the original MoGlow values by 1.5 until no improvement was found. The final values used were $\mathrm{lr}_{\max} = 10 \cdot \mathrm{lr}_{\min} = 1.5 \cdot 10^{-3}$. All proposed models (during hyperparameter tuning and in the final evaluation) were trained for 160,000 optimisation steps.

Unlike the network in [HAB19], our proposed systems took both past and future conditioning information into account. Specifically, our models took $\tau = 5$ historic frames (0.25 s) of concatenated joint poses and speech features, and 20 future frames (1 s) of speech as input when generating the next frame. The short context history is possible since older information can be propagated forward through the RNN. As described in Sec. 3.2, the 20-frame acoustic lookahead was necessary to for the model to prepare gestures so that they could be executed in synchrony with the speech.

As stated in Sec. 3.2, the use of data dropout prevents information from past poses from overriding other input signals. To tune the dropout rate for the poses in the autoregressive context $\boldsymbol{x}_{t-\tau:t-1}$, we exploited the fact that the accuracy of our style control can be evaluated objectively, since the realised control-parameter trajectories of any given gesture can be computed from sampled motion. Using the height of the right hand as our control parameter, we trained five separate networks with data-dropout rates from 0.0 to 0.8 increasing in steps of 0.2. By then providing the trained systems with a constant control input and evaluating the resulting (four-second average) right-hand height in sampled gestures, we picked the lowest dropout rate – 0.4 – where the sampled gestures obeyed the control over many random samples. This dropout rate was applied for all subsequent MoGlow-based systems, since we observed that not having any data dropout diminished the impact of speech control.

4.4. Proposed systems and baselines

Following parameter tuning, we trained a total of five different MoGlow-based systems: one system, denoted MG, conditioned only on speech, along with four systems that also allowed style control. Based on Sec. 4.2, these latter systems enabled control over the four-second average of either the right-hand hand height (system MG-H), the hand speed (MG-V, for velocity), the gesture radius (MG-R), or the degree of gesture symmetry (MG-S). All these systems used the same hyperparameters identified in Sec. 4.3.

To assess the quality of our approach, we compared our proposed systems against a number of toplines and baselines. As a topline, we used held out ground-truth gestures from the motion captured database (condition GT). We also evaluated the same ground-truth gestures but with mismatched speech audio taken from elsewhere in the database (condition MM). This condition should also exhibit fully natural motion, but should rate relatively lower on appropriateness of the gesticulation for the speech audio. We also trained and compared three baseline motion-generation systems taking the same speech-feature representation as input: A simple unidirectional LSTM network (LSTM) [HS97] with 1 layer and 350 nodes; an implementation of the conditional variation autoencoder (CVAE) for head motion presented in [GLM17] (1 layer and 350 nodes in each BLSTM, 2 latent dimensions); and the audio-to-representation-to-pose system (ARP) recently proposed in [KHH*19]. The latter maps speech audio to a (here) 40D motion representation space learned using denoising autoencoders, and then decodes the predicted representations to poses. Unlike MG, output from the three synthetic baselines required post-processing for smoothness. Details on the different systems trained in this paper can be found in Table 1.

As a bottom line, we created a small set of obviously unnatural gesticulation videos (condition BL), by synthesising output from the systems at an early stage of training. These videos (available, with all other videos seen by raters, in the supplement) served as lower anchors for the rating scale, and also provided a quality measure for filtering out spammers and highly inattentive raters.

4.5. Full-body synthesis

For simultaneous synthesis of full-body gesture and stance, we included the lower-body and hip joints and expressed the motion in a floor-level coordinate system that followed the character's position and direction. Following [HSK16, HAB19], we extracted three features for the root translation and rotation, namely the frame-wise delta x and z-translations together with the delta y-rotation of the floor-projected, smoothed hip pose. The smoothing is essential for control, and was set to 0.25 s for translation and 0.5 s for rotation. In this setting we retargeted the data to a slightly different skeleton with fewer spine and neck joints and also re-tuned the network parameters, yielding the following values: $K = 16$, $H = 512$, $\mathrm{lr}_{\max} = 2 \cdot 10^{-3}$, $\mathrm{lr}_{\min} = 5 \cdot 10^{-4}$. To speed up training times we also discarded intermediate frames from downsampling and trained the network for 80,000 steps.

Two full-body systems were trained: one (FB-U, for *uncontrolled*) in which all motion (joint angles and root translation/rotation) was synthesised from speech, and one (FB-C, for *controlled*) synthesising only body poses while treating the three root-motion features as additional control inputs. While the former system replicates stepping movements and pose shifts from the original data in an uncontrolled manner, the latter gives explicit control over character location and direction. This may be important in many scenarios, such as facing different interlocutors, portraying restlessness, or simply making the character stand still.

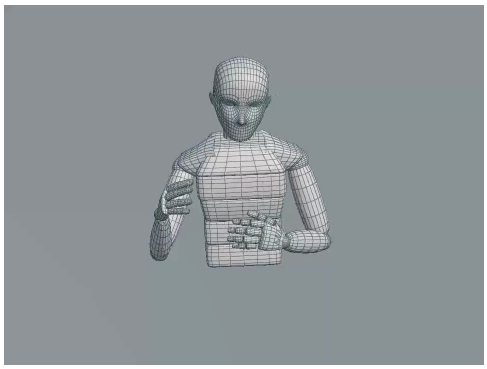


Figure 2: *Still image from video used in the subjective evaluation.*

5. Evaluation

In this section we describe the experiments used to evaluate our proposed approach to motion generation. We stress that objective evaluation of gesture synthesis is generally difficult – many plausible motion candidates exist for a fixed speech input, so a direct comparison against held-out natural motion recordings is not guaranteed to be meaningful. Instead, we base our evaluation on an extensive subjective evaluation against the toplines and baselines described in Sec. 4.4. We have performed three perceptual evaluation studies, looking at *human-likeness and appropriateness of motion*, *effects of style control* and *full-body synthesis* (simultaneous gesturing and walking). To enable meaningful cross-comparisons, the human-likeness, appropriateness and style-control conditions were evaluated in the same user study, while the full-body synthesis was evaluated in a separate study.

All perceptual studies were carried out using online experiments on the Figure Eight crowdworker platform, with the highest-quality contributor setting (allowing only the most highest-accuracy contributors) and country origin set to English-speaking countries (US, Canada, UK, Ireland, Australia and New Zealand). In all experiments, raters were instructed to watch and listen to 18-second video clips of a gesticulating figure as in Fig. 2, and rate them on five-point scales according to given scoring criteria. Since finger motion was not included in the study due to insufficient capture accuracy in the training data, the figure was provided with lightly cupped hands, static after the wrist, in the generated videos.

5.1. Evaluation of human-likeness and appropriateness

In this experiment the goal was to compare two aspects of the systems: 1) to what degree the generated motion looked like the motion of a real human, and 2) to what degree the gestures matched the speech. Six conditions were included: the proposed MG systems, the three baseline systems LSTM, CVAE and ARP, and finally GT (ground truth recordings).16 BL (bottom-line) examples were also included, to filter out unreliable raters. Raters were instructed to wear headphones and sit in a quiet environment. Prior to the start of the rating, subjects were trained by viewing example motion videos from the different conditions evaluated, as well as some of the bottom line examples.

Each stimulus was assessed by 40 independent crowdworkers, who were asked to rate the videos on a scale from 1 to 5 with respect to human-likeness (*"to what extent does the motion of the character look like the motion of a real human being"*) and appropriateness (*"to what extent does the motion match the audio"*), 5 being best. Other studies have found that many crowdworkers do not give gesture-rating tasks the attention they require [YKJ*19, JKEB19, KJvW*20]. As quality assurance, we employed strict rejection criteria: (a) any rater that had given any of the bottom-line (BL) examples a human-likeness rating above 3 *or* (b) had given a GT stimulus a rating of 1 on either human-likeness or appropriateness were excluded from the study. Also, (c) any ratings where the total time taken was less than 22 s (length of video + 4 s) or greater than 1200 s were discarded. Together, this removed 63% of the judgements. Out of these, 80% matched criterion (a), 29% matched (b) and 19% matched (c) (some judgements matched multiple rejection criteria). With (a) being the dominant rejection criterion, we can take a closer look at the distribution of responses for the BL examples: 1 (37%), 2 (13%), 3 (12%), 4 (21%), 5 (15%). This distribution has two peaks: the most common response was clearly 1, indicating that BL stimuli are indeed perceived as unnatural, but the high number of 4 and 5 ratings indicate that some subjects most likely are not making a sincere effort or are not understanding the task correctly, and should be discarded.

Mean ratings from the study are shown in Fig. 3 and Table 1. The human-likeness for GT and MG were 4.08±0.12 and 3.58±0.14, respectively, and for appropriateness 4.18±0.12 and 3.53±0.13. A one-way ANOVA revealed main effects of *human-likeness* and *appropriateness*, and a post-hoc Tukey multiple comparison test identified a significant difference between GT and all other conditions. MG was rated significantly above CVAE ($p < 0.001$) and LSTM ($p < 0.005$) on *human-likeness* and above CVAE ($p < 0.001$), LSTM ($p < 0.005$) and ARP ($p < 0.02$) on *appropriateness*.

5.2. Evaluation of style control

We now turn to evaluate the style control, both subjectively and objectively. The subjective evaluation was carried out in the same experiment described in the previous section, using the same number of raters and rater-exclusion criteria. Five different systems from Sec. 4.4 were assessed, namely the proposed MG system without style control, MG-H (hand height control), MG-V (velocity control), MG-R (gesture radius control) and MG-S (gesture symmetry control). For each of the four style-control systems, three groups of five animations were generated, where each group had a constant low, mid or high value of the control-input, defined by the 15th, 50th and 85th percentile of the control signal values in the training data. This yielded a total of 60 controlled video stimuli.

Mean values for the *human-likeness* rating for the different systems can be seen in the third plot in Fig. 3 and in Table 1. We see that style control at different levels had a minor effect on the perceived naturalness of the systems. The only significant difference between MG and the style controlled variants was for MG-S in the 85% setting ($p < 0.05$).

Fig. 4 illustrates the effect of style control on the motion generated by our systems, with one style-controlled system in each column (M-H, MG-V, MG-R and MG-S). The first two columns visualise the effect of low (first column) and high (second column)

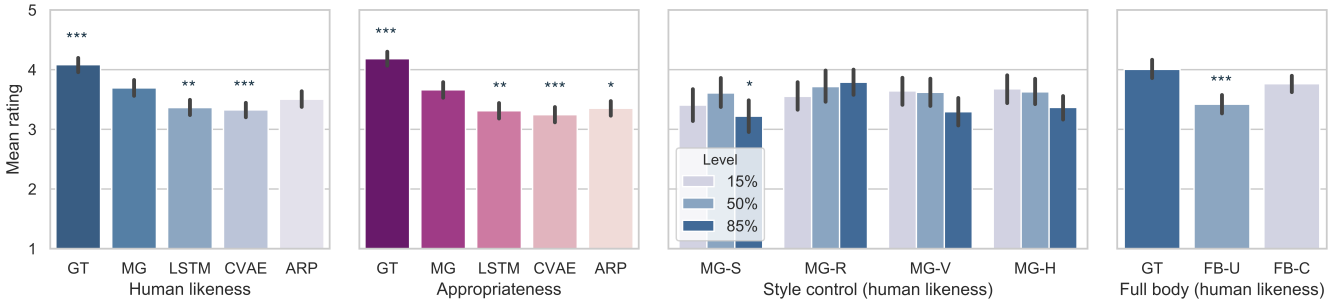


Figure 3: *Mean ratings from the perceptual experiments with 95% confidence intervals. Asterisks indicate significant effects ($*$: $p < 0.05$, $**$: $p < 0.01$, $***$: $p < 0.001$); for the three leftmost graphs comparisons are made against MG and for the rightmost graph against GT.*

System	Proba-bilistic?	Context frames	Hidden state	Pose params.	Training loss	Training epochs	Time (GPUs)	Log-like.	Mean rating	
									Human-likeness	Appropriateness
LSTM	✗	-	LSTM	1M	MSE	50	1 h (4)	N/A	3.36±0.13	3.31±0.13
CVAE	Partially	-	BLSTM	4M	MSE+KLD	50	12 h (4)	N/A	3.33±0.12	3.25±0.13
ARP	✗	61	GRU	0.5M	MSE	300	10 h (1)	N/A	3.51±0.14	3.35±0.13
MG	✓	26	LSTM	172M	Log-likel.	387	37 h (1)	266	3.69±0.13	3.66±0.14
MG-H	✓	26	LSTM	173M	Log-likel.	387	38 h (1)	267	3.63±0.23	N/A
MG-V	✓	26	LSTM	173M	Log-likel.	387	38 h (1)	264	3.62±0.23	N/A
MG-R	✓	26	LSTM	173M	Log-likel.	387	38 h (1)	270	3.72±0.26	N/A
MG-S	✓	26	LSTM	173M	Log-likel.	387	38 h (1)	307	3.61±0.25	N/A
FB-U	✓	26	LSTM	86M	Log-likel.	702	16 h (1)	320	3.42±0.16	N/A
FB-C	✓	26	LSTM	88M	Log-likel.	702	16 h (1)	303	3.76±0.14	N/A

Table 1: *Overview of the automatic gesture-generation systems (baselines, proposed and full-body) evaluated in this paper. Perceptual ratings of the style-controlled systems refer to the mid (i.e., 50%-level) control-input setting.*

control by superimposing motion frames from short excerpts of the generated output. The constant control-values for MG-H and MG-R are shown in red. The images suggest that the control input in all cases has affected the generated motion in the desired direction, with the effect being most visually obvious for MG-H and MG-R. The final column in Fig. 4 visualises – over time, and statistically – how the sampled output motion from the four models adheres to the given control signal for three control-signal input levels: low (in orange), mid (in cyan) and high (in green). The left plot shows time series ranging over 3700 frames of sampled motion and indicates instantaneous values, four-second smoothed values (extracted the same way as the control signal) and the control (line). The right boxplot shows the distribution of the residual between the input signal and the corresponding realised control and uses the same y-axis scale as time series to facilitate comparison.

Looking at the plots in the figure, we see that the curves generally are ordered orange, cyan, green (bottom to top), as expected. As an indication of control precision, the boxes showing the interquartile ranges of the realised control are mostly narrower than the separation between the constant control levels. The control of gesture radius is particularly distinct in this regard, with narrow boxes compared to the offset between the control levels. Both hand height and symmetry control demonstrate an intriguing behaviour where the observed variance around the style control input value is significantly greater for the low control input than at the other two levels. We hypothesise this might be due to discrepancies between the control input and the contexts in which that control input value occurs in the training data. For instance, long stretches of low hand height are rare in the data, as low hand heights often are associated with wide swinging motions. Sustained periods of negative correlation between the left and right hand are similarly uncommon in the training material. As a consequence, feeding in a low control input produces motion with inherently greater variability, inflating the boxes in the box plots.

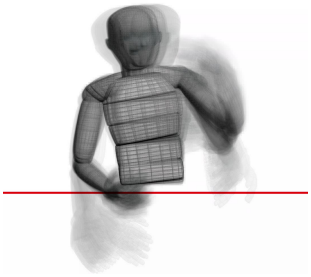
5.3. Evaluation of full-body gestures

The subjective evaluation of the full-body synthesis contained three conditions, GT and the two MoGlow systems from Sec. 4.5: FB-U (full-body motion from speech only) and FB-C (full-body motion also with controlled location and direction). 19 animations (cf. Fig. 5) from each condition were used in the evaluation, together with 16 BL (bottom line) animations used for quality control purposes like before. Subjects were asked to rate the animations on a scale from 1 to 5 for *human-likeness* (*"to what extent does the motion of the talking character look like the motion of a real human being?"*).

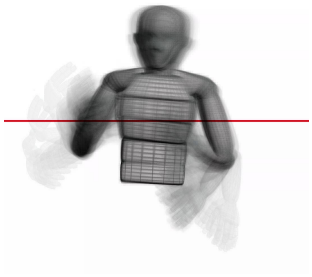
20 ratings were obtained for each stimulus. Raters who scored any of the BL animations a above 3 were excluded from the study, removing 44% of the judgements. Results can be seen in the rightmost pane of Fig. 3. Full-body GT received a mean rating of 4.005, FB-C 3.764 and FB-U 3.421. One-way ANOVA and a post-hoc Tukey multiple comparison test found a significant difference between GT and FB-U ($p < 0.001$), but not between GT and FB-C.

5.4. Discussion

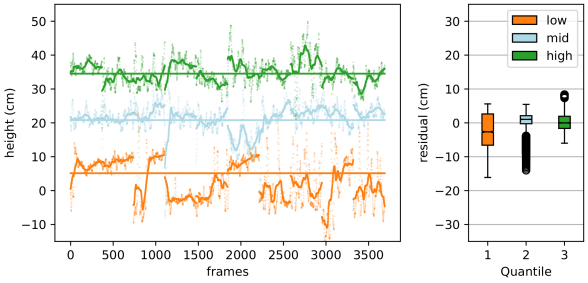
The results confirm that we have successfully achieved our goal of enabling probabilistic speech-driven gesture generation that permits optional style control and compares favourably against previous methods in the literature. However, while evaluations found MG gesticulation to be quite human-like and a reasonable match for the speech, it is our subjective impression (reinforced by the user study) that the generated gestures are not as vivid or diverse



(a) Low right hand, left hand unconstrained.

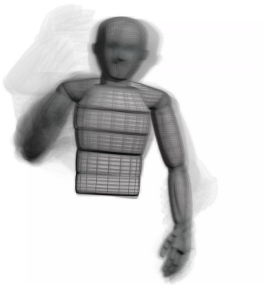


(b) High right hand, left hand unconstrained.

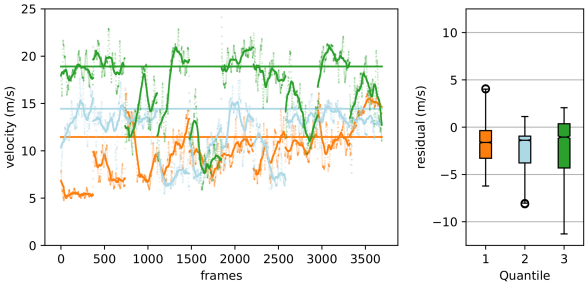


(c) Time series and residual statistics of right-hand height control.

(d) Low average speed.

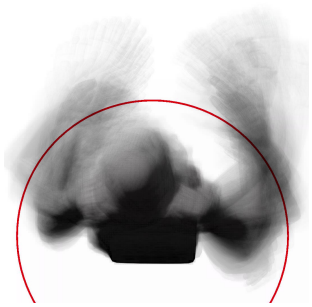


(e) High average speed.

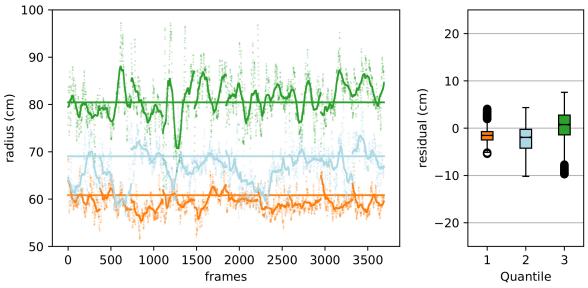


(f) Time series and residual statistics of speed control.

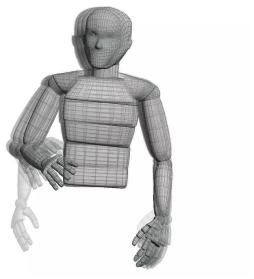
(g) Low average radius.



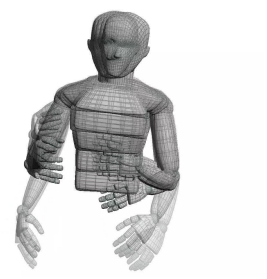
(h) High average radius.



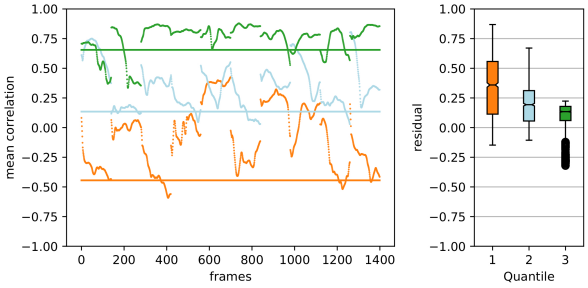
(i) Time series and residual statistics of gesture radius control.



(j) Low symmetry.



(k) High symmetry.



(l) Time series and residual statistics of gesture symmetry control.

Figure 4: *Effect and accuracy of style control. Each row is a system (MG-H, MG-V, MG-R and MG-S). Colours encode control-input values: orange for low (15th percentile), cyan for mid (50th), and green for high (85th). The first two columns show average images, each over 20 s excerpts with constant low or high control-input. (The bottom row instead uses a few onion-skinned snapshots for a better impression of symmetry.) The graphs on the right show the control input (flat line) and the corresponding instantaneous and smoothed control values of the output. For details, see Sec. 5.2.*

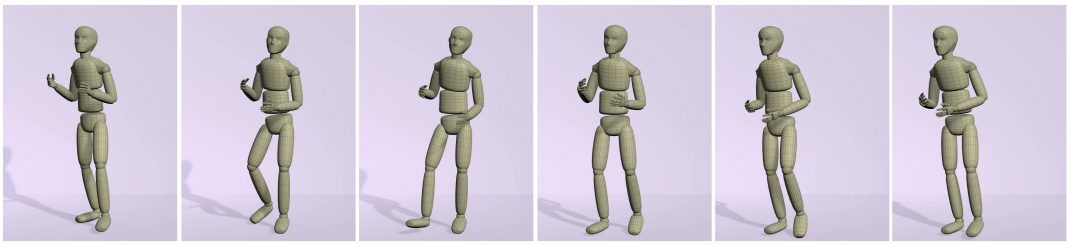


Figure 5: *Snapshots of full-body gestures and body motion.*

as the original motion-capture recordings. We believe this to indicate overfitting. During network tuning, we found that providing the right model complexity (especially K and H) was important for balancing stable gesture generation and perceived gesture quality. While underfitted models showed a great range of different behaviours (many of them unnatural) and sometimes got stuck in unnatural poses, overfitted models exhibited a reduced range of gestures, but those displayed were stable and followed the rhythm of the speech. We thus opted for a more stable gesture synthesis, with less vivid arm movements in our first experiment. The modified training scheme for the FB models was instituted as an attempt to strike a different balance between human-likeness and liveliness.

The speaker's high gesture rate and low amount of pauses may have affected the study in several ways. On the one hand, it may have been beneficial for learning, as it gave the systems a large number of gestures to train on. On the other hand, it may have complicated the evaluation, where a speaker with more pauses and slower speaking rate might have been easier to assess.

6. Conclusions and future work

We have presented a deep-learning-based system for automatic synthesis of co-speech gestures from speech input. The system is probabilistic, meaning it describes the entire distribution of likely gesture motions, and not only the mean pose. User studies find our system rated above several baselines from prior literature both in terms of human-likeness and on the appropriateness of the gestures for the given speech. We furthermore demonstrate that the approach can be extended to exert various kinds of directorial control over the style of the gesticulation without needlessly compromising human-likeness. Finally, we show that the method is capable of convincingly synthesising (controlled and uncontrolled) joint full-body posture, gesticulation and stance. This lifts the perspective from a focus on isolated body parts (e.g., hands or head) to holistically treating the entire human figure.

Future research goals include: 1) broadening the gesture repertoire, including towards gestures driven not only by acoustics but also by semantic content like in [KJvW*20]; 2) extending the model to cross-speaker synthesis, including developing speaker-independent features and assessing their transferability; and 3) developing and validating the style control for expressions at higher degrees of abstraction, such as emotion and character personality. (For goal 2, preliminary video examples of our trained models applied to audio from new speakers can be found in the supplement.) We additionally aim to unify the style models into a single model with optional inputs for controlling multiple styles simultaneously.

Acknowledgement

This work was supported by the Swedish Research Council proj. 2018-05409 (StyleBot), the Swedish Foundation for Strategic Research contract no. RIT15-0107 (EACare) and by the Wallenberg AI, Autonomous Systems and Software Program (WASP) of the Knut and Alice Wallenberg Foundation, Sweden.

References

[AMMS19] AHUJA C., MA S., MORENCY L.-P., SHEIKH Y.: To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *Proc. ICMI* (2019), pp. 74–84. 2

[AZS*17] ARISTIDOU A., ZENG Q., STAVRAKIS E., YIN K., COHEN-OR D., CHRYSANTHOU Y., CHEN B.: Emotion control of unstructured dance movements. In *Proc. SCA* (2017), p. 9. 3

[BDS19] BROCK A., DONAHUE J., SIMONYAN K.: Large scale GAN training for high fidelity natural image synthesis. In *Proc. ICLR* (2019). 3

[BH00] BRAND M., HERTZMANN A.: Style machines. In *Proc. SIGGRAPH* (2000), pp. 183–192. 3

[BK09] BERGMANN K., KOPP S.: GNetIc–using Bayesian decision networks for iconic gesture generation. In *Proc. IVA* (2009), pp. 76–89. 2

[CBDJ19] CHEN R. T. Q., BEHRMANN J., DUVENAUD D., JACOBSEN J.-H.: Residual flows for invertible generative modeling. In *Proc. NeurIPS* (2019), pp. 9913–9923. 3

[CM11] CHIU C.-C., MARSELLA S.: How to train your avatar: A data driven approach to gesture generation. In *Proc. IVA* (2011), pp. 127–140. 2

[CMM15] CHIU C.-C., MORENCY L.-P., MARSELLA S.: Predicting co-verbal gestures: A deep and temporal modeling approach. In *Proc. IVA* (2015). 2

[CN19] CASTILLO G., NEFF M.: What do we express without knowing?: Emotion in gesture. In *Proc. AAMAS* (2019), pp. 702–710. 1, 2

[FM18] FERSTL Y., MCDONNELL R.: Investigating the use of recurrent motion modelling for speech gesture generation. In *Proc. IVA* (2018), pp. 93–98. 4

[FNM19] FERSTL Y., NEFF M., MCDONNELL R.: Multi-objective adversarial gesture generation. In *Proc. MIG* (2019), pp. 3:1–3:10. 2

[GBK*19] GINOSAR S., BAR A., KOHAVI G., CHAN C., OWENS A., MALIK J.: Learning individual styles of conversational gesture. In *Proc. CVPR* (2019), pp. 3497–3506. 2

[GH00] GHAHRAMANI Z., HINTON G. E.: Variational learning for switching state-space models. *Neural Comput. 12*, 4 (2000), 831–864. 3

[GLM17] GREENWOOD D., LAYCOCK S., MATTHEWS I.: Predicting head pose from speech with a conditional variational autoencoder. In *Proc. Interspeech* (2017), pp. 3991–3995. 2, 5

[GMHP04] GROCHOW K., MARTIN S. L., HERTZMANN A., POPOVIĆ Z.: Style-based inverse kinematics. *ACM T. Graphic. 23*, 3 (2004), 522–531. 3

[Goo16] GOODFELLOW I.: NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint* (2016). arXiv:1701.00160. 3

[GPAM*14] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. In *Proc. NIPS* (2014), pp. 2672–2680. 3

[HAB19] HENTER G. E., ALEXANDERSON S., BESKOW J.: MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *arXiv preprint* (2019). arXiv:1905.06598. 1, 2, 3, 4, 5

[HHKK17] HOLDEN D., HABIBIE I., KUSAJIMA I., KOMURA T.: Fast neural style transfer for motion data. *IEEE Comput. Graph. 37*, 4 (2017), 42–49. 2

[HHS*17] HABIBIE I., HOLDEN D., SCHWARZ J., YEARSLEY J., KOMURA T.: A recurrent variational autoencoder for human motion synthesis. In *Proc. BMVC* (2017). 2

[HKS17] HOLDEN D., KOMURA T., SAITO J.: Phase-functioned neural networks for character control. *ACM T. Graphic. 36*, 4 (2017), 42:1–42:13. 2

[HKS*18] HASEGAWA D., KANEKO N., SHIRAKAWA S., SAKUTA H., SUMI K.: Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proc. IVA* (2018), pp. 79–86. 2

[HPP05] HSU E., PULLI K., POPOVIĆ J.: Style translation for human motion. In *ACM T. Graphic.* (2005), vol. 24, pp. 1082–1089. 2

[HS97] HOCHREITER S., SCHMIDHUBER J.: Long short-term memory. *Neural Comput. 9*, 8 (1997), 1735–1780. 5

[HSK16] HOLDEN D., SAITO J., KOMURA T.: A deep learning framework for character motion synthesis and editing. *ACM T. Graphic. 35*, 4 (2016), 138:1–138:11. 2, 5

[JKEB19] JONELL P., KUCHERENKO T., EKSTEDT E., BESKOW J.: Learning non-verbal behavior for a social robot from YouTube videos. In *Proc. ICDL-EPIROB Workshop Nat. Non-Verbal Affect. Hum.-Robot Interact.* (2019). 6

[KB15] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. In *Proc. ICLR* (2015). 5

[KBE*20] KUMAR M., BABAEIZADEH M., ERHAN D., FINN C., LEVINE S., DINH L., KINGMA D.: VideoFlow: A conditional flow-based model for stochastic video generation. In *Proc. ICLR* (2020). 3

[KD18] KINGMA D. P., DHARIWAL P.: Glow: Generative flow with invertible 1x1 convolutions. In *Proc. NeurIPS* (2018), pp. 10236–10245. 3

[KG10] KOPPENSTEINER M., GRAMMER K.: Motion patterns in political speech and their influence on personality ratings. *J. Res. Pers. 44*, 3 (2010), 374–379. 1, 2

[KHH*19] KUCHERENKO T., HASEGAWA D., HENTER G. E., KANEKO N., KJELLSTRÖM H.: Analyzing input and output representations for speech-driven gesture generation. In *Proc. IVA* (2019), pp. 97–104. 2, 5

[KJvW*20] KUCHERENKO T., JONELL P., VAN WAVEREN S., HENTER G. E., ALEXANDERSON S., LEITE I., KJELLSTRÖM H.: Gesticulator: A framework for semantically-aware speech-driven gesture generation. *arXiv preprint* (2020). arXiv:2001.09326. 6, 9

[KW14] KINGMA D. P., WELLING M.: Auto-encoding variational Bayes. In *Proc. ICLR* (2014). 3

[Lip98] LIPPA R.: The nonverbal display and judgment of extraversion, masculinity, femininity, and gender diagnosticity: A lens model analysis. *J. Res. Pers. 32*, 1 (1998), 80–107. 2

[LKM*18] LUCIC M., KURACH K., MICHALSKI M., GELLY S., BOUSQUET O.: Are GANs created equal? A large-scale study. In *Proc. NeurIPS* (2018), pp. 698–707. 3

[LKTK10] LEVINE S., KRÄHENBÜHL P., THRUN S., KOLTUN V.: Gesture controllers. *ACM T. Graphic. 29*, 4 (2010), 124. 2

[LWH*12] LEVINE S., WANG J. M., HARAUX A., POPOVIĆ Z., KOLTUN V.: Continuous character control with low-dimensional embeddings. *ACM T. Graphic. 31*, 4 (2012), 28. 3

[McN92] MCNEILL D.: *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992. 2

[NLK*13] NORMOYLE A., LIU F., KAPADIA M., BADLER N. I., JÖRG S.: The effect of posture and dynamics on the perception of emotion. In *Proc. SAP* (2013), pp. 91–98. 2

[PAM*18] PUMAROLA A., AGUDO A., MARTINEZ A. M., SANFELIU A., MORENO-NOGUER F.: GANimation: Anatomically-aware facial animation from a single image. In *Proc. ECCV* (2018), pp. 818–833. 3

[PNR*19] PAPAMAKARIOS G., NALISNICK E., REZENDE D. J., MOHAMED S., LAKSHMINARAYANAN B.: Normalizing flows for probabilistic modeling and inference. *arXiv preprint* (2019). arXiv:1912.02762. 3

[PVC19] PRENGER R., VALLE R., CATANZARO B.: WaveGlow: A flow-based generative network for speech synthesis. In *Proc. ICASSP* (2019), pp. 3617–3621. 3

[PWP18] PHAM H. X., WANG Y., PAVLOVIC V.: Generative adversarial talking head: Bringing portraits to life with a weakly supervised neural network. *arXiv preprint* (2018). arXiv:1803.07716. 3

[RMW14] REZENDE D. J., MOHAMED S., WIERSTRA D.: Stochastic backpropagation and approximate inference in deep generative models. In *Proc. ICML* (2014), pp. 1278–1286. 3

[SB18] SADOUGHI N., BUSSO C.: Novel realizations of speech-driven head movements with generative adversarial networks. In *Proc. ICASSP* (2018), pp. 6169–6173. 2, 3

[SB19] SADOUGHI N., BUSSO C.: Speech-driven animation with meaningful behaviors. *Speech Commun. 110* (2019), 90–100. 2

[SCNW19] SMITH H. J., CAO C., NEFF M., WANG Y.: Efficient neural networks for real-time motion style transfer. *ACM T. Graphic. 2*, 2 (2019), 13. 2

[SN17] SMITH H. J., NEFF M.: Understanding the impact of animated gesture performance on personality perceptions. *ACM T. Graphic. 36*, 4 (2017), 49. 1, 2

[SSKS17] SUWAJANAKORN S., SEITZ S. M., KEMELMACHER-SHLIZERMAN I.: Synthesizing Obama: learning lip sync from audio. *ACM T. Graphic. 36*, 4 (2017), 95. 2

[VPP19] VOUGIOUKAS K., PETRIDIS S., PANTIC M.: Realistic speech-driven facial animation with GANs. *Int. J. Comput. Vision* (2019), 1–16. 3

[WFH08] WANG J. M., FLEET D. J., HERTZMANN A.: Gaussian process dynamical models for human motion. *IEEE T. Pattern Anal. 30*, 2 (2008), 283–298. 3

[WMK14] WAGNER P., MALISZ Z., KOPP S.: Gesture and speech in interaction: An overview. *Speech Commun. 57* (2014), 209–232. 2

[WTY18] WANG X., TAKAKI S., YAMAGISHI J.: Autoregressive neural f0 model for statistical parametric speech synthesis. *IEEE/ACM T. Audio Speech 26*, 8 (2018), 1406–1419. 3

[XWCH15] XIA S., WANG C., CHAI J., HODGINS J.: Realtime style transfer for unlabeled heterogeneous human motion. *ACM T. Graphic. 34*, 4 (2015), 119. 2

[YKJ*19] YOON Y., KO W.-R., JANG M., LEE J., KIM J., LEE G.: Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *Proc. ICRA* (2019). 2, 6