Speech-Driven Conversational Agents using Conditional Flow-VAEs

Sarah Taylor
S.L.Taylor@uea.ac.uk
University of East Anglia
Norwich, UK

Jonathan Windle
J.Windle@uea.ac.uk
University of East Anglia
Norwich, UK

David Greenwood
David.Greenwood@uea.ac.uk
University of East Anglia
Norwich, UK

Iain Matthews
Iain.Matthews@epicgames.com
Epic Games
Pittsburgh, USA

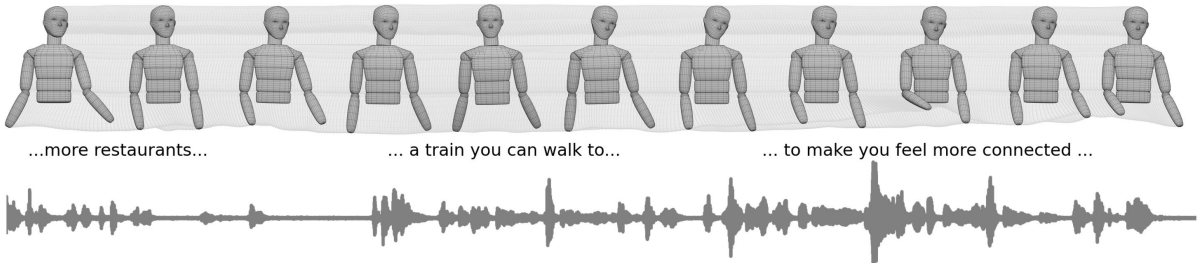

Figure 1: Our method uses Conditional Flow-VAEs to model the complex, many-to-many relationship between the speech signal and body gesture. Our approach works equally well for monologue *and* dyadic conversation, with a unified model providing compelling animation for both speaking and listening modalities.

ABSTRACT

Automatic control of conversational agents has applications from animation, through human-computer interaction, to robotics. In interactive communication, an agent must move to express its own discourse, and also react naturally to incoming speech. In this paper we propose a Flow Variational Autoencoder (Flow-VAE) deep learning architecture for transforming conversational speech to *body* gesture, during both speaking and listening. The model uses a normalising flow to perform variational inference in an autoencoder framework and is a more expressive distribution than the Gaussian approximation of conventional variational autoencoders. Our model is non-deterministic, so can produce variations of plausible gestures for the same speech. Our evaluation demonstrates that our approach produces expressive body motion that is close to the ground truth using a fraction of the trainable parameters compared with previous state of the art.

CVMP '21, December 6–7, 2021, London, United Kingdom

ACM ISBN 978-1-4503-9094-1/21/12…$15.00
https://doi.org/10.1145/3485441.3485647

CCS CONCEPTS

• **Computing methodologies** → **Continuous models**; **Supervised learning by regression**; **Neural networks**; **Motion processing**.

KEYWORDS

normalising flows, variational autoencoder, body animation, speech driven

ACM Reference Format:
Sarah Taylor, Jonathan Windle, David Greenwood, and Iain Matthews. 2021. Speech-Driven Conversational Agents using Conditional Flow-VAEs. In *European Conference on Visual Media Production (CVMP '21), December 6–7, 2021, London, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3485441.3485647

1 INTRODUCTION

Speech gesturing encompasses the motions of the body that accompany speech, including movements of the head, arms and torso. Gestures play a key role in human communication by conveying messages that are complementary to speech, providing information about the semantic content of utterances, emotion, and emphasis [De Ruiter et al. 2012; Kendon 1994; Mcneill 1985; McNeill 1992]. In conversational speech, gestures may facilitate speech understanding and are critical to natural interaction and turn-taking [Maatman et al. 2005]. They are used for expressing feedback during listening (eg. nodding) [Wagner et al. 2014] and are indicative of levels of understanding and agreement. Without natural speech gesturing,

the communicative extent of conversational agents is limited and perceived realism is reduced [Ennis et al. 2010].

In this work, we seek to model the relationship between speech and upper body gesture in dyadic conversation. We automatically generate realistic body animation from just an audio signal and an indication of who is talking. This work has applications in character animation, social robotics and driving conversational agents.

The relationship between speech and gesture is complex and many-to-many; the same phrase may be connected to any number of gestures, and similar gestures may occur during different utterances. Speech and body events may occur asynchronously, and the onset of a gesture very often appears prior to the realisation of the speech [Kendon 1972; Wagner et al. 2014]. Gesture is also idiosyncratic, and we each express ourselves uniquely. These factors make speech-driven gesture generation particularly challenging.

We pose the problem of speech gesture estimation as: given an input sequence of *interactive* conversational speech, automatically generate the corresponding body motion for *one* of the speakers, whilst speaking *and* listening. Probabilistic generative models are appealing for this task, as they provide a statistical model of the data, usually in the form of a probability density. This allows for a non-deterministic output for any input, which is well suited to the complex relationship between speech and gesture.

One such model is the generative variant of an autoencoder, the Variational Autoencoder (VAE), which learns a distribution that can be sampled to generate new examples. VAEs typically approximate the latent space with a Gaussian distribution for its simplicity and efficiency. However, a simple Gaussian may lack the expressiveness to accurately capture the complex true latent distribution, which can impact on the generative capacity of the model [Cremer et al. 2018; Mescheder et al. 2017]. Flow Variational Autoencoders (Flow-VAEs) [Bhattacharyya et al. 2019] overcome this limitation by using normalising flows [Kingma and Dhariwal 2018] to model the distribution of the variational latent space.

Normalising flows transform a complex distribution to a simpler, typically Gaussian, distribution through a chain of bijective and differentiable transformations, and have recently been successfully applied to speech gesture estimation [Alexanderson et al. 2020]. However, normalising flows can be computationally demanding. By embedding flows in an autoencoder framework, we are able to speed up training time and reduce model complexity while achieving human-like gesturing.

We extend monologue approaches and introduce conditional Flow-VAEs for estimating speech gestures from *dyadic* speech by augmenting the control signal with an indication of conversational role. Our model predicts body gestures relating to *speaking* and *listening*, as well as *cross-talk* and the ambient moments between speech activity. Our main contributions can be summarised as: 1) We introduce a speech body gesture model for conversational agents which estimates upper body motion for both talking and listening; 2) We introduce a speech activity indicator which allows for control of conversational role; 3) We propose a new Flow-VAE architecture with a novel loss term for speech gesture generation; and 4) We propose a set of summary statistics for qualitatively evaluating the behaviour of conversational agents. Our code is available at https://github.com/UEA-digital-human-group/udh-flowvae.

2 RELATED WORK

2.1 Audio-Driven Body Pose Estimation

The first approaches for driving body motion from speech were rule-based [Cassell et al. 1994, 2004; Hartmann et al. 2005; Marsella et al. 2013]. Rule-based techniques were mostly concerned with semantic aspects of human gesturing. With limited fidelity, these approaches tended to lack realism.

Early data-driven approaches were based on probabilistic modelling. Neff *et al.* [Neff et al. 2008] computed the probability that a body gesture from a library of gestures was to be generated, conditioned on context. Chiu and Marsella [Chiu and Marsella 2014] used Gaussian process latent variable models to learn a mapping from speech to hand gestures through an intermediate representation of gesture annotation. Levine *et al.* [Levine et al. 2009] trained a hidden Markov model on prosody features. This idea was later integrated into a reinforcement learning framework [Levine et al. 2010]. Exposing the underlying probability distribution of body motion conditioned on speech is desirable as sampling this distribution generates non-deterministic output. However, the Gaussian assumptions of these prior works are limiting.

Naturally, a new wave of solutions arrived as deep learning architectures and algorithms were developed. For example, Long Short Term Memory (LSTM) models were trained to animate a skeleton to play along with piano and violin music in [Shlizerman et al. 2018], and a recurrent network with an encoder-decoder structure was used for gesture generation in [Ferstl and McDonnell 2018]. Generative Adversarial Networks (GANs) have been used to train co-speech gesture models in a variety of ways. Pang *et al.* [Pang et al. 2020] trained a GAN using an autoregressive generator and a sinusoidal activation function to mimic periodic behaviour [Sitzmann et al. 2020]. A Recurrent Neural Network (RNN) generator was trained using multiple adversaries in work by Ferstl *et al.* [Ferstl et al. 2019]. They additionally classified gesture phase, which was subsequently used to train one of the discriminators. The adversarial training paradigm by Ginosaur *et al.* [Ginosar et al. 2019] learned to automatically control 2D keypoints. They used data taken from *in-the-wild* videos. Yoon *et al.* [Yoon et al. 2020] used an adversarial training scheme to learn body motion given inputs corresponding to speech text, audio, and speaker identity.

The recent GENEA Gesture Generation Challenge [Kucherenko et al. 2020] delivered a competitive array of methods for predicting body gesture from audio speech, and evaluated them against one-another in a fixed evaluation framework. One of the best performing techniques [Korzun et al. 2020] used both audio and text input to train an attention-based sequence-to-sequence translation model, and was based on the work of [Kucherenko et al. 2019; Yoon et al. 2019]. Another of the best-performing methods was based on normalising flows [Alexanderson 2020; Alexanderson et al. 2020]. This was an auto-regressive technique for estimating the probability distribution of the next pose in a sequence conditioned on characteristics of the motion (eg. hand height, speed and radius). Our approach is inspired by this work, and we discuss more details in Section 3.2.

Most work has been limited to estimating co-speech gestures for a single speaker. However, recent work by Ahuja *et al.* [Ahuja et al. 2020] instead learned a gesture space and a per-speaker style

embedding. To estimate body pose, the gesture space was sampled conditioned on some speech and a style embedding. We instead focus on modelling body motion during conversational interaction between two speakers.

There has been previous work on predicting gesture in dyadic settings. For example a bi-directional LSTM was trained to predict head pose in [Greenwood et al. 2017], and [Jonell et al. 2020] used normalising flows for estimating facial motion in conversational speech. For body pose, [Yang et al. 2020] constructed a motion graph that was searched based on characteristics of the target audio speech and [Ahuja et al. 2019] proposed an attention-based model for switching between monadic and dyadic functions. Instead, our method augments the control signal with an indication of conversational role.

2.2 Normalising Flows and VAEs

Normalising Flows and VAEs have previously been combined in different ways. For example, Gritsenko *et al.* [Gritsenko et al. 2019] proposed a variational denoising autoencoder in which the encoder takes the form of a normalising flow. [Ziegler and Rush 2019] used normalising flow-based priors in the latent space of unconditional variational autoencoders for discrete distributions. [Mahajan et al. 2020] used normalising flows for modelling joint complex latent distributions for image captioning. The conditional flow variational autoencoder developed by [Bhattacharyya et al. 2019] learns conditional priors based on normalising flows to model distributions in the latent space of Conditional Variational Autoencoders (CVAEs). Our model is most similar to this architecture, and we provide details of the key differences in Section 3.

3 METHOD

In this section we introduce a conditional Flow Variational Autoencoder (Flow-VAE) for speech-driven gesture animation. We include related work to provide technical background.

3.1 Variational Autoencoders

Variational Autoencoders (VAEs) are generative variants of an autoencoder that describe observations in the latent space in a probabilistic manner. Given inputs $\mathbf{x}$, the encoder, $q_\phi(\mathbf{h}|\mathbf{x})$, learns the parameters of a Gaussian distribution that is used to approximate the posterior distribution $p(\mathbf{h}|\mathbf{x})$; where ϕ are the encoder network weights and biases, and $\mathbf{h}$ is the latent space. The decoder, $d_\theta(\mathbf{x}|\mathbf{h})$, maps a sample h from the variational distribution back to a distribution on the input domain, with network weights and biases θ.

VAEs are powerful generative models. However, a major limitation is that it is not possible to realise the true posterior distribution using a Gaussian variational distribution [Hoffman 2017; Rezende and Mohamed 2015]. Instead, Flow-VAEs use normalising flows for variational inference in an autoencoder framework.

3.2 Normalising Flows

Normalising flows provide a highly flexible method for transforming a simple distribution (E.g. Gaussian) to a more complex distribution through a series of invertible and differentiable transformations [Bhattacharyya et al. 2019; Kingma and Dhariwal 2018; Rezende and Mohamed 2015]. The goal is to learn a transformation $f : \mathbb{R}^D \rightarrow \mathbb{R}^D$, of a complex distribution $\mathbf{H}$ to a simple base distribution $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, that is typically Gaussian. The transformation f is non-linear and bijective, and is constructed by chaining together a *flow* of K simpler sub-transformations: $f_k : \mathbb{R}^D \rightarrow \mathbb{R}^D$, where $k \in 1, \ldots, K$.

$$\mathbf{z} \approx \mathbf{z}_K = f_K\left(f_{K-1}\left(\ldots f_1\left(\mathbf{z}_0\right)\right)\right) \tag{1}$$

$$\mathbf{h} = f^{-1}\left(\mathbf{z}\right) = f_1^{-1}\left(f_2^{-1}\left(\ldots f_K^{-1}\left(\mathbf{z}\right)\right)\right) \tag{2}$$

In Equation 1, $\mathbf{z}_0 = \mathbf{h}$, the latent distribution. One can efficiently sample from the normal distribution, $\mathbf{Z}$, and transform to the domain of $\mathbf{H}$ using Equation 2. Since f is invertible, the exact log-likelihood of a data sample $\mathbf{h}$ may be computed using the change-of-variables formula [Henter et al. 2020]:

$$\ln p_\psi(\mathbf{h}) = \ln p_\mathcal{N}(\mathbf{z}_K) + \sum_{k=1}^{K} \ln \left| \det \frac{\delta \mathbf{z}_k}{\delta \mathbf{z}_{k-1}} \right|, \tag{3}$$

where $p_\mathcal{N}$ is the probability density of the base distribution, ψ are the model parameters, and $\mathbf{z}_k$ is the result of the sub-transformation f_k. The parameters of the flow can be trained using a gradient-based optimisation framework by maximising the log-likelihood of the training data.

The set of transformations used in this work are called Motion Glow (MoGlow), which were developed by [Kingma and Dhariwal 2018] and extended to an autoregressive architecture by [Henter et al. 2020]. Each flow step, f_k, has three sub-steps. The first two, *Actnorm* and *Linear*, are parametric affine transformations, and the third, *Affine Coupling*, is a non-linear transformation of which the parameters are learned using a neural network (Figure 3). Alexanderson *et al.* [Alexanderson et al. 2020] opted for LSTMs in this final sub-step, which we also adopt in this work.

MoGlow is an autoregressive model used for estimating the next-step distribution of a sequence, $\mathbf{h} = [\mathbf{h}_0, \ldots, \mathbf{h}_T]$. Autoregression is achieved by feeding previous poses, $\mathbf{h}_{T-\tau:T-1}$ as additional inputs to the LSTM in each step of the flow. Other conditioning variables can also be fed into the LSTM for controlling properties of the output.

3.3 Conditional Flow-VAEs

Conditional Flow-VAEs *combine* VAEs and conditional normalising flows. Variational inference is achieved by directly maximising the log-likelihood in the latent space using a normalising flow. Since the encoder is trained with the flow, it generates a distribution that is more easily transformed to a Gaussian distribution, allowing for a simpler and shorter flow (with faster training) to achieve realistic results. Figure 2 shows the Flow-VAE framework during training. Given input $\mathbf{x}$, the encoder, $q_\phi(\mathbf{h}|\mathbf{x})$, learns a compressed latent representation, $\mathbf{h}$, from which the decoder, $d_\theta(\mathbf{x}|\mathbf{h})$, maps back to the input space. Simultaneously, a normalising flow, f_ψ, is trained to transform the distribution of the latent space, $\mathbf{H}$, to a normal distribution $\mathbf{Z} = \mathcal{N}(0, \mathbf{I})$. Conditioning is performed by feeding a control signal to the affine coupling sub-transformation in each step of the flow, following the implementation of MoGlow [Alexanderson 2020; Henter et al. 2020]. Figure 3 illustrates this process for flow step f_k. Note that, unlike regular conditional VAEs and previous work on Flow-VAEs [Bhattacharyya et al. 2019], there is

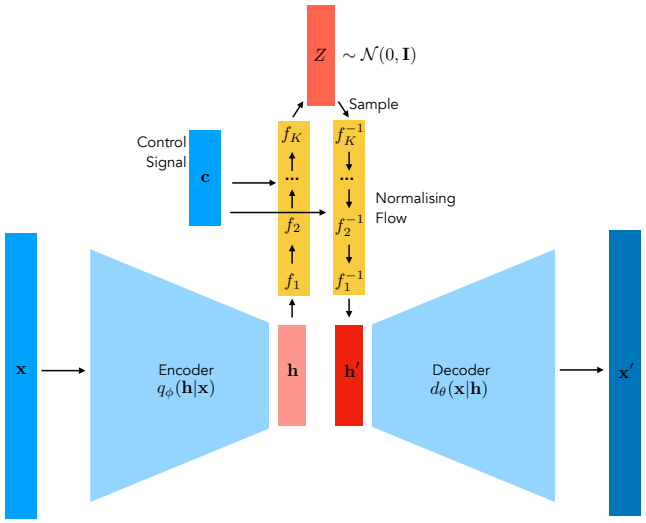

Figure 2: Flow-VAE at training time. The encoder generates a latent representation, h, of body pose x. A normalising flow transforms H to a Gaussian distribution, Z, conditioned on the control signal, c. The sampled latent variables h′ are fed to the decoder.

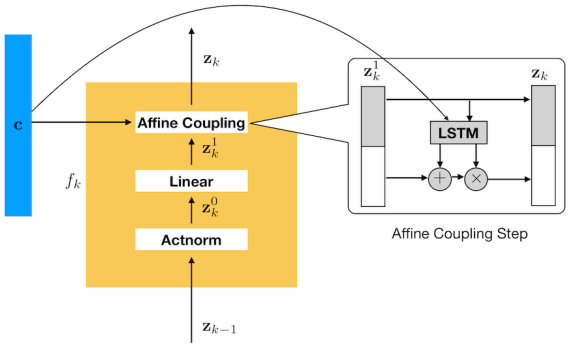

Figure 3: A forward flow step. The input $\mathbf{z}_{k-1}$ is transformed to $\mathbf{z}_k$ via three sub-transformations. The control signal, c, is fed to the final, affine coupling sub-transformation, illustrated on the right. The input is split and one half is used for computing a translation and scaling of the other half through an LSTM [Henter et al. 2020].

no conditioning input to the autoencoder, and all conditioning is performed in the flow.

To produce temporally cohesive gestures, the Flow-VAE is autoregressive. The model is trained to estimate the next pose given speech, conditioning and pose information from the previous τ frames (together with speech and conditioning from the future γ frames). During training, the control signal at frame t, $\mathbf{c}_t$, is composed of the previous τ frames of gesture data, $\mathbf{x}_{t-\tau,\ldots,t-1}$, a window of speech, $\mathbf{s}_{t-\tau,\ldots,t+\gamma}$, and a window of conditioning variables $\mathbf{a}_{t-\tau,\ldots,t+\gamma}$ (see Section 4.3). The control signal is fed into the affine coupling sub-step in each flow step (Figure 3). Since

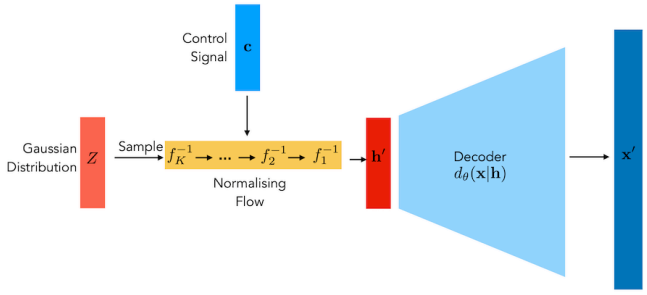

Figure 4: Flow-VAE at test time. Z is sampled to generate a latent vector h′ conditioned by the input speech which is fed to the decoder and converted to x′.

speech and gesture production may be asynchronous [Butterworth and Hadar 1989], the model must take a window of audio speech that contains γ frames of look-ahead to trigger motions that occur prior to their related audio.

The Flow-VAE is trained by minimising the objective:

$$L(\mathbf{x}, \mathbf{c}|\phi, \theta, \psi) = \alpha \cdot \mathrm{NLL}(\mathbf{h}) + D_{\mathrm{rec}}(\mathbf{h}) + D_{\mathrm{rec}}(\mathbf{h}') \quad (4)$$

$$\mathrm{NLL}(\mathbf{h}) = -\ln p_\psi(q_\phi(\mathbf{h}|\mathbf{x}, \mathbf{c})) \quad (5)$$

$$D_{\mathrm{rec}}(\mathbf{h}) = \mathrm{MSE}(\mathbf{x}, d_\theta(\mathbf{x}|\mathbf{h})) \quad (6)$$

$$D_{\mathrm{rec}}(\mathbf{h}') = \mathrm{MSE}(\mathbf{x}, d_\theta(\mathbf{x}|\mathbf{h}')) \quad (7)$$

The first term (Equation 5) is the negative log-likelihood of Z given the encoded latent variables $\mathbf{h}$ and is computed using Equation 3. This term is weighted by α to account for the difference in scale to the other loss terms. The second term (Equation 6) is the reconstruction loss of the autoencoder, computed as the Mean Squared Error (MSE) between input $\mathbf{x}$ and the decoded $\mathbf{h}$, $\mathbf{x}'$. The final term (Equation 7) is another reconstruction loss, this time between input $\mathbf{x}$ and decoded $\mathbf{h}'$, which is sampled from the flow. This final term is novel to our approach, and we observe that it improves the stability of training and acts a form of regularisation.

Figure 4 shows the model at test time. A sample, $\mathbf{z_K}$, is randomly generated from $\mathbf{Z}$. Together with the control signal, $\mathbf{c}$, the sample is transformed through the inverse flow, f^{-1}, to $\mathbf{h}'$, which is fed to the decoder to get a prediction, $\mathbf{x}'$. At test time, the control signal is composed of the previous τ frames of *estimated* gesture data, $\mathbf{x}'_{t-\tau,\ldots,t-1}$, along with the speech and conditioning information (Section 4.3).

4 DATA AND PREPROCESSING

There are no conversational speech and 3D motion datasets that are publicly available for training our model. Existing speech and body datasets contain monologue [Ferstl and McDonnell 2018]. We collected a rich speech and gesture dataset, with humble non-specialist hardware and a setup that is easy to replicate for future collaborative growth.

4.1 Dataset

A male speaker (Speaker A) was filmed conversing with a female speaker (Speaker B) who was off-camera. Speaker A was filmed before a green backdrop from three synchronised views (see Figure 5).

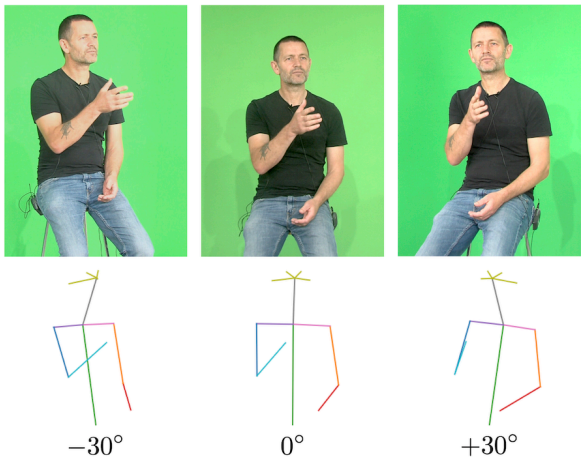

Figure 5: A frame from each camera view (top), and the corresponding pose at 0 and ±30 degrees from frontal pose (bottom) shown on a reference skeleton.

The video was recorded at 25fps and 1080p resolution with 48kHz audio.

The dataset contains ≈3.5 hours of dialogue and is made up of three parts: Part 1 (1 hour) contains unscripted conversation between the two speakers. Part 2 (1 hour) is a debate on a topic that was chosen from a list by Speaker A. Speaker B argued the opposing view to Speaker A to incite a heated discussion. Part 3 (1.5 hours) is a performance of scripted emotional monologue vignettes, which were included to provoke a broader range of affective states.

4.2 Body Pose Representation

We locate 2D keypoints independently in each of the three camera views using the monocular body pose detection system OpenPose [Cao et al. 2019]. We calibrate the cameras using a checkerboard target, and project the 2D keypoints from each view into 3D world space by triangulation. If a keypoint does not appear in all three views, it is omitted. We remove lower body keypoints, and we discard the hand keypoints as these were not reliably tracked. The head is considered a rigid object, so we reduce the rotations of the eyes and ears to a single rotation. We represent body gesture using the 9 remaining keypoints as illustrated in Figure 5: sternum; left and right shoulders, elbows and wrists; nose; and head. Poses are translated so that the base of the spine rests at the world origin, $(0, 0, 0)$. Since the data were recorded over multiple sessions (on the same day), the speaker's global orientation varies over time. We frontalise the pose by using orthogonal procrustes alignment [Schönemann 1966] to compute the rotation, $\mathbf{R}$, that minimises the distance between the hip and shoulder landmarks of a clip, to the corresponding landmarks of a frontal reference skeleton. $\mathbf{R}$ is computed and applied once per clip, and a clip is defined as a natural break in the capture or a 13 minute segment, whichever is shortest.

For representing pose, Cartesian coordinates [Ahuja et al. 2020; Ginosar et al. 2019], or a transformation of these (eg. using PCA [Shlizerman et al. 2018]), are often used. However, keypoints explicitly encode the body proportions of the speaker. Whilst we believe it valuable to model the gestural *style* of a speaker, we do not wish to encode physical attributes. Instead, we define pose as the angle of each joint with respect to a reference neutral pose.

At frame t, we represent pose by 9 joint rotations that represent the shortest angle between each limb and the corresponding limb on a reference skeleton in a T-pose. Euler angles suffer from gimbal lock, and quaternions are discontinuous [Zhou et al. 2019], so we use the 6DoF rotations used by [Pang et al. 2020]. In practice, the 6 elements are the first two rows of the 3×3 rotation matrix. We stack these elements to form a pose vector, $\mathbf{x}_t^{\text{raw}} = [x_{0,0}, \ldots, x_{8,5}]$ which is of dimension 9 (joints) $\times$ 6 (elements) = 54. To reconstruct the skeleton for visualisation, the final three elements of each joint's rotation matrix can be recomputed as the cross product of the two rows.

Not all keypoints are visible in all three camera views in all frames. We postprocess the joints, $\mathbf{x}_{\text{raw}}$, using a denoising autoencoder [Lu et al. 2013], which both temporally smooths the joint trajectories and imputes these missing rotations. At training time we ignore frames with missing joints and only use those with all keypoints visible in all views. The autoencoder takes a stacked window of 5 frames of rotations as input. 10% of these rotations are set to zero and the autoencoder is trained to reconstruct the complete set of rotations by minimising the MSE between the estimated and the original rotations. We slide the 5-frame window, shifting by one frame at a time. The encoder is a feed-forward network with three layers containing 70, 50 and 30 nodes respectively, and the decoder mirrors this structure. We use batch normalisation after the first layer, and Rectified Linear Unit (ReLU) activations between layers. The model was trained for 250 epochs using the Adam optimiser and a learning rate of 0.001. The trained model is used for processing all of the raw joint rotations to generate a complete set of rotations, $\mathbf{x} = [\mathbf{x}_0, \ldots, \mathbf{x}_T]$, where T is the number of frames in the dataset, which forms the training data for all subsequent models. Two 300s segments are held out for testing and validation.

4.3 Control Signal

To generate co-speech gestures that are synchronous with Speaker A's speech and that exhibit natural behaviour throughout the interaction with Speaker B we must provide the system with both rich acoustic information and an indication of the speaker's conversational role. Poses from previous frames must also be included to enable the model to estimate temporally consistent motion. Thus, the control signal is a concatenation of the previous gesture data, speech and speaker activity features, $\mathbf{c} = [\mathbf{x}; \mathbf{s}; \mathbf{a}]$, which are defined below.

4.3.1 Speech Audio. We extract 27-channel mel-frequency spectrograms from 40ms windows of speech to match the frame rate of the video. This gives a set of speech vectors $\mathbf{s} = [\mathbf{s_0}, \ldots, \mathbf{s_T}]^T$, of dimension $T \times 27$ that align with the gesture data $\mathbf{x}$.

4.3.2 Speech Activity Indicator. We provide the system with a single audio track that contains the dialogue between Speakers A and B. The speech activity indicator specifies, for each frame, which speaker the audio belongs to. This information is represented as a 4D one hot encoding with channels corresponding to:

Speaker A; Speaker B; Both or; None, and results in a set of vectors $\mathbf{a} = [\mathbf{a_0}, \ldots, \mathbf{a_T}]^T$ with dimension $T \times 4$.

4.4 Data Augmentation

We augment our dataset by mirroring the gesture data along the vertical y-axis. The control signal is duplicated for the mirrored joint rotations.

5 FLOW-VAE DESIGN AND OPTIMISATION

We optimised the Flow-VAE by evaluating performance on the validation set. We first tuned the hyperparameters of the autoencoder. This was performed independently of the flow by removing the variational inference step and finding an autoencoder architecture that accurately reconstructs the input gesture data. A symmetric feed-forward network with two layers, respectively containing 105 and 35 ReLUs, was sufficient.

The autoencoder architecture was fixed and a grid search was used to tune the flow. The flow was initialised on the implementation of MoGlow by [Alexanderson et al. 2020], which contained $K = 16$ flow steps and $N = 800$ units in the LSTM of the affine-coupling transformation in each sub-step. We varied K and N and reviewed both the loss and animation quality of the validation set, and observed good results with $K = 8$ and $N = 400$. This is half the size of the MoGlow model, yet we achieved no clear improvement using a larger model.

The flow is autoregressive and is fed a 5 frame (200ms) window of gesture data that leads up to but does not include the current frame. This is concatenated with a 30 frame window of the control signal, spanning 5 frames (200ms) of the past and 25 frames (1s) of the future. Previous work found that 1s of future audio context was necessary for capturing the asynchronous production of speech and gesture [Alexanderson et al. 2020]. Data dropout was applied to the gesture data elements of the control signal at a rate of 0.4.

We trained all components of the Flow-VAE simultaneously using the Adam optimiser with Noam learning rate decay. Maximum and minimum learning rates were set to 1.5×10^{-3} and 1.5×10^{-4} respectively. Models were trained for 300 epochs and α was set to 5×10^{-4} when computing the loss (Equation 4).

6 EXPERIMENTS

We make an evaluation by comparing our method against contemporary approaches on a 300*s* held out sequence. The sequence is a continuous conversational exchange that contains Speaker A speaking and Speaker B speaking (while Speaker A is listening). There are also periods of cross-talk, where both speakers are speaking, and periods of silence.

It is difficult to quantitatively evaluate speech-driven gesture estimation since the output must display the characteristics of the speaker's motion rather than match a ground truth sequence frame-by-frame. We make a qualitative comparison and present summary statistics in Section 6.2.

6.1 Baseline Systems

6.1.1 MoGlow. We compare against the normalising flow approach taken by MoGlow [Alexanderson et al. 2020]. Although originally developed for monologue, the authors demonstrated that their system can be controlled by various conditioning signals such as speed, symmetry, and spatial coverage. We replace these conditioning signals with speech activity indicators, matching the inputs to our system defined in Section 4.3. The model was trained for 300 epochs, using the original MoGlow hyperparameters and optimisation schedule.

6.1.2 CVAE. We also compare against a CVAE. The encoder and decoder both have 5 LSTM layers, each with 200 hidden units with 20% dropout. The conditioning signal is the combined audio features and the speech activity described in Section 4.3. The reconstruction loss is MSE, and the total loss adds the Kullback?Leibler Divergence (D_{KL}). The CVAE network is trained using Adam, with a learning rate of 10^{-4}. Training continues until no further loss reduction with a patience of 10 epochs, a total of 50 epochs.

6.2 Qualitative Evaluation

Making judgements on the quality of an animation is difficult by comparing time series alone. One would not expect a model to predict a motion close in value to a ground truth sequence. Rather, one would expect the *characteristics* of the motion to be faithful to the ground truth. We present a qualitative comparison of the time series and a set of summary statistics to evaluate the performance of the systems against ground truth motion (GT).

We evaluate motion of the 3D joint keypoints since it is arguably more intuitive to consider the distance between points than rotation angles. We find the mean pose for all ground truth data and calculate the Euclidean distance from each landmark to its mean, and report that distance in real world centimetres. Figure 6 shows the time series plots for each of GT (a), our method Flow-VAE (d), MoGlow (g), and finally CVAE (j). In the same order, we show waisted box plots for each comparative method to display summary statistics for each joint involved in the motion. Finally, in the rightmost column of Figure 6, we render a perspective projection of the skeleton hierarchy at every second in the test sequence, that is, 300 frames of motion each overlaid to give a visual impression of the physical space that the motion occupies.

By examining Figure 6, we can make some observations regarding the character of the motion. Our first observation is that CVAE does not appear to be as expressive as the ground truth sequence, or the other comparative methods. Comparing Figures 6(c) and 6(l) clearly shows the body as a whole is less active, whereas the comparison with our method in Figure 6(f) shows we are able to model motion that occupies a very similar space.

It is not surprising that the largest space is occupied by the most active joints, the shoulders to wrists. Viewing the interquartile range in Figures 6(b) and 6(e) shows our method makes gestures at a similar scale to the ground truth. Interestingly, the ground truth shows a bias to left handed gesturing, yet our result reverses this. Recall from Section 4.4 that we augment the motion data by mirroring, so left right biases may exchange sides. Particularly large gestures are recognised as outliers in the box plots, and high points in the time series plots. We note that MoGlow does not appear to perform as well in this regard.

An important mode of dyadic visual speech is rigid head pose. The time series plots in Figures 6(a) and 6(d) show rigid head pose

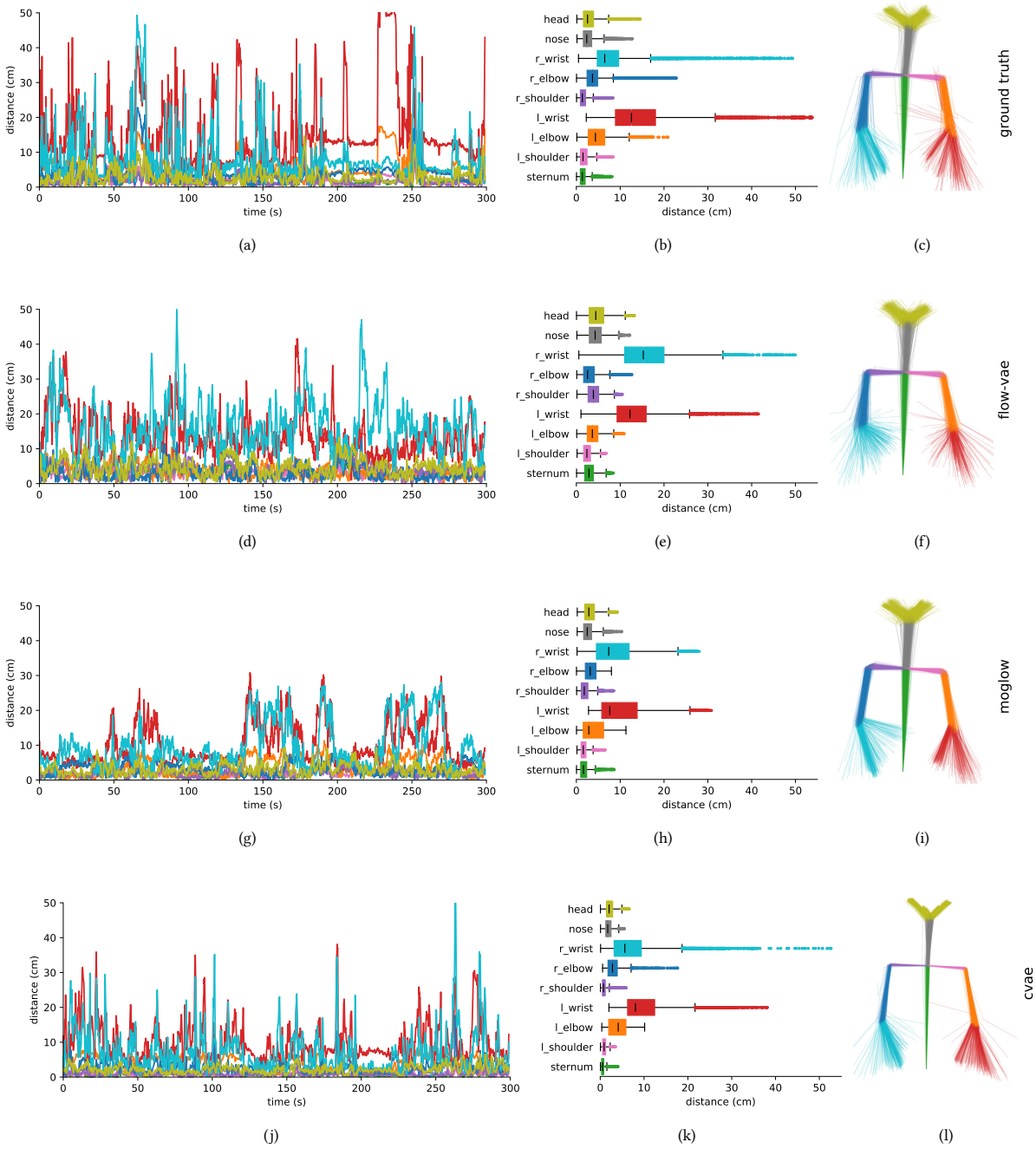

Figure 6: Qualitative assessment sequence. For each of ground truth, Flow-VAE, MoGlow and CVAE, we show the distance from the mean pose as a time series, the summary statistics box plot, and a perspective projection of the pose at every 1 second interval. These views give a good impression of the comparative approaches in our experiments. They show the scale of the motion, timing of events and frequency of activity. It is immediately apparent that the CVAE model is less active than the other models and the ground truth. We can also see our method produces animation close to the ground truth, particularly for the most active lower arms.

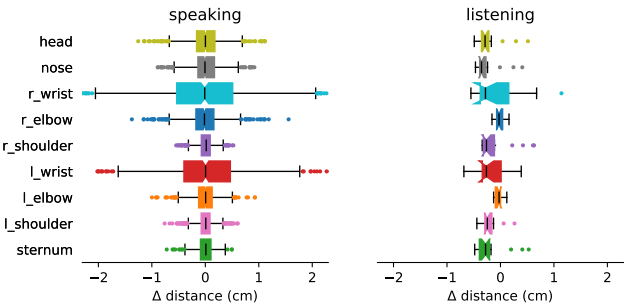

Figure 7: Comparing statistics for speaking and listening in the ground truth. Using the speech activity label, we divide the data to speaking and listening segments. Here we show box plots for the first derivative of the distance to the mean pose. This provides a sense of the level of activity for during each speaking mode.

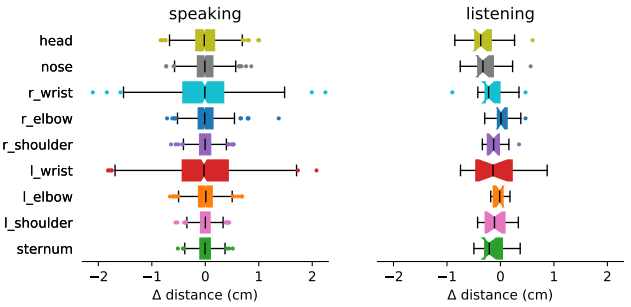

Figure 8: Comparing statistics for speaking and listening for Flow-VAE. Again we show box plots for the first derivative of the distance to the mean pose for both speaking and listening modes. Here we can see, in a similar manner to the ground truth, activity differs between the two modes.

from our method matching the characteristics of the ground truth very closely. The baseline CVAE performs least well here, with rather limited head pose.

6.3 Speaking and Listening

When listening, the avatar must realistically transmit back-channel signals to give the appearance of engagement with the off camera interlocutor. This is an important part of the behaviour of a conversational agent [Greenwood et al. 2017], so we further explore the behaviour of our model during listening.

Our held out data is labelled for speech activity. We separate the Speaker A activity to speaking and listening modes. For each of these modes we then take the first derivative of the distance from the mean pose. The derivative is chosen here to properly show the difference in motion activity rather than the space occupied by the speaker. For example, while Speaker B is talking the subject may hold a pose for a number of seconds. We show this view for GT in Figure 7, and for our method in Figure 8. Our model shows very similar statistical changes when switching from speaking and listening modes, when we compare to the same ground truth sequence. In particular, when listening to Speaker B, the actor moves towards the mean pose, so the gradients (in both cases) become negative.

7 DISCUSSION

An example of the retargeted motion can be seen in Figure 1, which shows the animated pose at every 38 frames of a 15s duration clip along with the audio waveform and a transcription of the speech. Video examples of the animated results can be found in our supplementary material. It is clear that the Flow-VAE successfully replicates gesturing style from the ground truth and produces new motions with recognisable characteristics of the speaker. Our model generates animation that is expressive, and gestures that are diverse but plausible for the speech. Our model generates smaller movements that are closer to the mean pose on Speaker A when Speaker B is speaking, which is both intuitively correct and consistent with ground truth behaviour.

In total there are 34M, 180M and 3M trainable parameters in the Flow-VAE, MoGlow and CVAE respectively. Our Flow-VAE contains 80% fewer trainable parameters than MoGlow, and this introduces a substantial reduction on training time. CVAE has fewer parameters, but generates less realistic animation.

7.1 Limitations

Since the model lacks any notion of semantics, it typically produces beat gestures that coincide with the timing and intensity of speech. On occasion it also generates head nodding at plausible locations, but typically misses head nods during the short statements of agreement when Speaker B is talking. Our dataset is biased towards Speaker A, and does not contain many long sequences of Speaker B's speech. We expect that a more balanced dataset might further improve the diversity of gestures that our model generates during listening.

8 CONCLUSION

We have presented a technique for automatically animating speech gestures from audio for *conversational* agents using a Flow-VAE architecture. Flow-VAEs overcome the limitations of conventional VAEs by using normalising flows for variational inference. By embedding normalising flows in an autoencoder framework, we are able to speed up training time and reduce model complexity compared to using normalising flows in isolation, and generate more expressive animation than conventional CVAEs. Flow-VAEs are non-deterministic and generate natural looking speech gestures for both speaking and non-speaking segments of a dyadic conversation. A qualitative evaluation indicated that the estimated behaviours are consistent with ground truth motion.

REFERENCES

Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. 2020. Style Transfer for Co-speech Gesture Animation: A Multi-speaker Conditional-Mixture Approach. In *European Conference on Computer Vision (ECCV)*. Cham, 248–265.

Chaitanya Ahuja, Shugao Ma, Louis-Philippe Morency, and Yaser Sheikh. 2019. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. In *International Conference on Multimodal Interaction*. 74–84.

Simon Alexanderson. 2020. The StyleGestures entry to the GENEA Challenge 2020. https://doi.org/10.5281/zenodo.4088600

Simon Alexanderson, Gustav Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Computer Graphics Forum* 39, 2 (05 2020), 487–496.

Apratim Bhattacharyya, Michael Hanselmann, Mario Fritz, Bernt Schiele, and Christoph-Nikolas Straehle. 2019. Conditional Flow Variational Autoencoders for Structured Sequence Prediction. In *Bayesian Deep Learning NeurIPS 2019 Workshop*.

Brian Butterworth and Uri Hadar. 1989. Gesture, Speech, and Computational Stages: A Reply to McNeill. *Psychological review* 96 (02 1989), 168–74.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 43, 1 (2019), 172–186.

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 *Computer Graphics and Interactive Techniques*. 413–420.

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

Chung-Cheng Chiu and Stacy Marsella. 2014. Gesture generation with low-dimensional embeddings. In *International Conference on Autonomous Agents and Multi-agent Systems*. 781–788.

Chris Cremer, Xuechen Li, and David Duvenaud. 2018. Inference Suboptimality in Variational Autoencoders. In *International Conference on Machine Learning (ICML)*. 1078–1086.

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.

Cathy Ennis, Rachel McDonnell, and Carol O'Sullivan. 2010. Seeing is believing: Body motion dominates in multisensory conversations. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 1–9.

Ylva Ferstl and Rachel McDonnell. 2018. IVA: Investigating the use of recurrent motion modelling for speech gesture generation. In *Intelligent Virtual Agents (IVA)*. https://trinityspeechgesture.scss.tcd.ie

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

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3497–3506.

David Greenwood, Stephen Laycock, and Iain Matthews. 2017. Predicting Head Pose in Dyadic Conversation. In *Intelligent Virtual Agents (IVA)*. Springer International Publishing, 160–169.

Alexey A Gritsenko, Jasper Snoek, and Tim Salimans. 2019. On the relationship between Normalising Flows and Variational-and Denoising Autoencoders. *ICLR Workshop on Deep Generative Models for Highly Structured Data* (2019).

Björn Hartmann, Maurizio Mancini, and Catherine Pelachaud. 2005. Implementing expressive gesture synthesis for embodied conversational agents. In *International Gesture Workshop*. Springer, 188–199.

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

Matthew D Hoffman. 2017. Learning deep latent Gaussian models with Markov chain Monte Carlo. In *International Conference on Machine Learning (ICML)*. 1510–1519.

Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. 2020. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Intelligent Virtual Agents (IVA)*. 1–8.

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

Adam Kendon. 1994. Do gestures communicate? A review. *Research on Language and Social Interaction* 27, 3 (1994), 175–200.

Diederik P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1× 1 convolutions. (2018), 10236–10245.

Vladislav Korzun, Ilya Dimov, and Andrey Zharkov. 2020. The FineMotion entry to the GENEA Challenge 2020. https://doi.org/10.5281/zenodo.4088609

Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. In *Intelligent Virtual Agents (IVA)*. 97–104.

Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2020. The GENEA Challenge 2020: Benchmarking gesture-generation systems on common data. https://doi.org/10.5281/zenodo.4094697

Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture Controllers. *ACM Transactions on Graphics (TOG)* 29 (07 2010).

Sergey Levine, Christian Theobalt, and Vladlen Koltun. 2009. Real-Time Prosody-Driven Synthesis of Body Language. In *ACM SIGGRAPH* (Yokohama, Japan) *(SIGGRAPH Asia '09)*. Association for Computing Machinery, New York, NY, USA, Article 172, 10 pages. https://doi.org/10.1145/1661412.1618518

Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. 2013. Speech enhancement based on deep denoising autoencoder.. In *Interspeech*. 436–440.

R M Maatman, Jonathan Gratch, and Stacy Marsella. 2005. Natural Behavior of a Listening Agent. In *Intelligent Virtual Agents (IVA)*. Springer Berlin Heidelberg, Berlin, Heidelberg, 25–36.

Shweta Mahajan, Iryna Gurevych, and Stefan Roth. 2020. Latent Normalizing Flows for Many-to-Many Cross-Domain Mappings. In *International Conference on Learning Representations (ICLR)*.

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*. 25–35.

David Mcneill. 1985. So You Think Gestures are Nonverbal? *Psychological Review* 92 (07 1985), 350–371.

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

Lars Mescheder, S Nowozin, and Andreas Geiger. 2017. Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*. PMLR, 2391–2400.

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, Article 5 (March 2008), 24 pages. https://doi.org/10.1145/1330511.1330516

Kunkun Pang, Taku Komura, Hanbyul Joo, and Takaaki Shiratori. 2020. CGVU: Semantics-guided 3D Body Gesture Synthesis. https://doi.org/10.5281/zenodo.4090879

Danilo Jimenez Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *International Conference on Machine Learning (ICML)*, Vol. 37. 1530–1538.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31, 1 (1966), 1–10.

Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. 2018. Audio to body dynamics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7574–7583.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. 2020. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems* 33 (2020).

Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Guest Editorial: Gesture and speech in interaction: An overview. *Speech Communication* 57 (2014), 209–232.

Yanzhe Yang, Jimei Yang, and Jessica Hodgins. 2020. Statistics-based Motion Synthesis for Social Conversations. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 201–212.

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Transactions on Graphics (TOG)* 39, 6, Article 222 (Nov. 2020), 16 pages.

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 *International Conference on Robotics and Automation (ICRA)*. IEEE, 4303–4309.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5745–5753.

Zachary Ziegler and Alexander Rush. 2019. Latent normalizing flows for discrete sequences. In *International Conference on Machine Learning (ICML)*. PMLR, 7673–7682.