GESTURE2VEC: CLUSTERING GESTURES USING REPRESENTATION LEARNING METHODS FOR CO-SPEECH GESTURE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Co-speech gestures are a principal component in conveying messages and enhancing interaction experiences between humans. Similarly, the co-speech gesture is a key ingredient in human-agent interaction including both virtual agents and robots. Existing machine learning approaches have yielded only marginal success in learning speech-to-motion at the frame level. Current methods generate repetitive gesture sequences that lack appropriateness with respect to the speech context. In this paper, we propose a Gesture2Vec model using representation learning methods to learn the relationship between semantic features and corresponding gestures. We propose a new conceptual framework that considers gestures as a non-verbal language itself. Our approach first converts gesture sequences into symbolic chunks, using frame and sequence level autoencoders and rigorous training techniques to learn the vocabulary. Using this higher-level representation, we then take advantage of a machine translation model to learn translations of text to discrete sequences of associated gesture chunks in the learned gesture space. Ultimately, we use these quantized gestures as input to the autoencoder's decoder to produce gesture sequences. The resulting gestures can be applied to both virtual agents and humanoid robots. Ablation studies support that our gesture chunking approach, fixed decoder weights, and vector quantization are the main drivers of diversity in objective gesture diversity measures. Further subjective and objective evaluations confirm the success of our approach in terms of appropriateness, human-likeness, and diversity.

1 INTRODUCTION

Non-verbal behaviour is an indispensable part of our daily communication. Prior research stated that $70-93\%$ of communication is non-verbal, including facial expression, hand gesture, body pose, and vocal tones (Mehrabian, 2017; Lapakko, 2007). People spontaneously gesticulate to complement verbal channels during the speech to convey messages (McNeill, 2011; De Ruiter et al., 2012; Cassell et al., 1999). Hence, integrating non-verbal communication skills into social robots and virtual agents is crucial for compelling interactions (Minato et al., 2004; Woods et al., 2004; Breazeal et al., 2005). Gestures originate from semantic features(Chu & Kita, 2016) and are characterized by speech context (Lücking et al., 2013). Therefore, co-speech gesture generation has been typically addressed using textual and acoustic features to generate relevant gestures.

Gesture synthesis can be categorized into deterministic and probabilistic models. Deterministic models predict a single output for a given input, while probabilistic models estimate a plausible output probability distribution conditioned on the given input. The limitation of deterministic data-driven methods is that they neglect to span the variation of many dimensions in data space (Fragkiadaki et al., 2015; Ferstl et al., 2019). Probabilistic models are potentially able to capture a broader gesture space, but probabilistic models also suffer from the posterior-collapse problem (Bowman et al., 2015) which results in repetitive gestures close to the average. The problem arises when the one-to-one mapping assumption disregards the one-to-many relationship between speech and gesture. This is a common problem among generative models for gestures (Yoon et al., 2019; Ginosar et al., 2019), as well as in other domains, e.g. image generation models that produce a limited set of blurry and similar images (Lucas et al., 2019a), and early attempts for natural language generation (Bowman et al., 2015; Kingma et al., 2016; He et al., 2019). Literature attempted to address this posterior collapse problem using different techniques such as adversarial training (Goodfellow et al., 2014; Ginosar et al., 2019; Arjovsky et al., 2017; Srivastava et al., 2017), variational autoencoders (Higgins et al., 2016; Mi et al., 2018; Ling et al., 2020), normalizing flow (Alexanderson et al., 2020), vector quantization (Oord et al., 2017) (Razavi et al., 2019), weakened decoders (Yang et al., 2017; Semeniuta et al., 2017). However, as reported, prior work (Yoon et al., 2019; Ferstl et al., 2019; Kucherenko et al., 2020b) were not successful in capturing long-term dependencies. Indeed, the model parameters focus on imperceptible and local features such as continuity among consecutive frames, most likely due to mode collapse.

Inspired by advances in other machine learning fields, especially natural language processing, we model gestures as a language that contains a vocabulary, and then perform text-to-gesture as a language translation task. Specifically, we integrate unsupervised representation learning methods and a machine translation algorithm. First, we reduce body pose dimensionality using a Denoising Autoencoder at the frame level. Then, we perform a discretized motion representation learning to cluster similar motion sequences to a symbol. This vector quantization method is a form of cluster pattern recognition where each motion sequence is assigned to a particular word from a codebook. Finally, we train a machine translation model to translate between utterances and their accompanying gestural motion symbols. This method is effective since it mitigates the complexity of gesture space and focuses on longer dependencies. In addition to traditional evaluation measurements, we also present new objective and subjective metrics to evaluate the diversity of gestures.

2 Related Work

We first review co-speech gesture generation methods. Next, we discuss recent gesture generation approaches and their limitations. Finally, we introduce the deep neural network that we used and its advantages over prior work.

Early attempts for gesture generation use blending to smooth feasible motion clips selected from a database. Rule-based algorithms (Cassell et al., 2004; Huang & Mutlu, 2012), hidden Markov models (Levine et al., 2009), conditional random fields (Levine et al., 2010), and hybrid systems (Kipp, 2005; Neff et al., 2008) (Chiu et al., 2015) were used to select proper gestures conditioned on a given input. It is now recognized that these kind of systems require extensive efforts to annotate data and cannot generate gestures for unseen inputs. Also, they provide a single prediction for a given input and lack variation of generated movement. Furthermore, scheduling the gestures with speech is challenging since they may not be precisely synchronized (Butterworth & Hadar, 1989; Kendon, 2004) despite originating from the same source (McNeill, 2011).

Recent deep generative models, e.g. VAEs (Kingma & Welling, 2013; Higgins et al., 2016), GANs (Goodfellow et al., 2014; Abdal et al., 2019), and transformers (Vaswani et al., 2017; Brown et al., 2020; Dehghani et al., 2018), achieved notable results on different tasks. A generative model comprises the joint probability of given data and output. Therefore, it can generate new plausible instances by taking samples from the learned distribution (Hasegawa et al., 2018). Generative models have been used for human motion generation (Yan et al., 2018; Hernandez et al., 2019; Ling et al., 2020) as well as co-speech gesture generation (Ginosar et al., 2019; Yoon et al., 2019; Ferstl et al., 2021; Li et al., 2021). However, co-speech gesture generation is more challenging since the relationship between speech and motion is complex (Butterworth & Hadar, 1989).

Input to the generative system can be supplied from either speech-text, audio, or both modalities with a broad range of semantic and acoustic features (Kucherenko et al., 2021c). Systems that use acoustic features led to generated beat gestures according to the speech rhythm (Hasegawa et al., 2018; Ferstl et al., 2020; Ginosar et al., 2019; Kucherenko, 2018; Pouw & Dixon, 2019), and text-based systems (Yoon et al., 2019; Ishi et al., 2018) produced more semantically aware gestures. Although text-based models capture communicative features, they lack the strong effect of speech acoustics, i.e. intonation, prosody, and loudness, on expressed gestures (Pouw et al., 2020). Recent work benefiting from both modalities generated more compelling co-speech gestures in terms of appropriateness and naturalness (Yoon et al., 2020).

Typically, co-speech gesture generation systems build upon an encoder and decoder architecture (Hasegawa et al., 2018; Kucherenko et al., 2019; Yoon et al., 2019; Kucherenko et al., 2020a; Ferstl

& McDonnell, 2018). Recurrent Neural Networks (RNN) structures have been used extensively for the encoder and decoder (Hasegawa et al., 2018). However, RNN based models suffer from the error accumulation problem and are not good at capturing long-term human motion dependencies (Hernandez et al., 2019). While Convolutional Neural Network (CNN) based gesture generation models are not vulnerable to accumulation problems, CNNs are prone to regress to the average motion and generate repetitive gestures (Li et al., 2021). Variational Autoencoders were also proposed for the co-speech gesture generation task to generate more realistic and diverse gestures, e.g. (Rezende et al., 2014; Li et al., 2021; Kucherenko, 2018). VAEs with strong decoders tend to ignore the latent variable and learn the mode of the output data. This problem, known as "posterior collapse" causes the model to generate slightly similar outputs close to the average (Lucas et al., 2019b).

Literature considerably explored adversarial training to generate more realistic and diverse gestures (Ginosar et al., 2019; Ahuja et al., 2020; Yoon et al., 2020; Li et al., 2021; Ferstl et al., 2020; 2021). Although they brought attractive results, GANs are hard to train (Lucic et al., 2017) and suffer from mode collapse (Tulyakov et al., 2018). Normalizing-flow based models have GANs advantages while replacing adversarial loss by classical likelihood maximization training (Kingma & Dhariwal, 2018) and efficient probabilistic inference of VAEs (Henter et al., 2020) (Kingma & Dhariwal, 2018). Furthermore, autoregressive models on discrete data achieved impressive results in many sequences-to-sequence tasks such as machine translation (Wang et al., 2019), image generation(Salimans et al., 2017) (Razavi et al., 2019), and speech synthesis (Oord et al., 2016) (Gârbacea et al., 2019). VQ-VAE models learn a discrete latents space, enable us to use autoregressive models on the posterior, and does not suffer from "posterior collapse" (Oord et al., 2017). Ordinarily, gesture generation systems use a generator network that produces gestures from a latent code. To better regress the data and cover a broad range of motion, (Li et al., 2021; Kucherenko et al., 2021b;a) uses a motion-specific code space that represents motion attributes.

All of the state-of-the-art models mentioned above generate gestures frame-by-frame. We suggest that this architecture drains model capabilities on local dependencies at the expense of global features and diversity. We therefore reformulate the problem as a machine translation task. Inspired by VQ-VAE, we propose a method that combines VQ-VAE, Denoising Autoencoders and weakened decoders to learn a discretized latent space. Finally, we use an autoregressive model on the quantized motions to produce co-speech gestures from the input.

3 METHOD

We considered Yoon et al. (2019) as our baseline since we also focus on word embedding textual features (Bojanowski et al., 2017) as the control signal for gesture generation. The authors in (Yoon et al., 2019) used an RNN based Encoder-Decoder architecture with a soft attention mechanism (Bahdanau et al., 2014; Cho et al., 2014). The encoder extracts textual features, and the decoder produces co-speech gestures frame-by-frame. We extend the baseline model by utilizing representation learning approaches. Our proposed system has the following steps.

- Pose representation learning at the frame-level
- Discrete motion representation learning at the sequence-level.
- Translation from text to the learned discrete motion representation space.

Learning powerful representations without supervision is of utmost importance to reduce problem complexity. Autoencoders are unsupervised models that learn significant data features and discard spurious patterns by minimizing the reconstruction error (Kingma & Welling, 2013). They consist of an encoder network followed by a decoder network. An autoencoder's bottleneck finds a shared data representation (Goodfellow et al., 2009; Bengio et al., 2007) by learning the correlations between input dimensions and reconstructs them from a low-dimensionality representation. We used two different autoencoders with different components, both at the frame-level and sequence-level, as explained in the following sections.

3.1 FRAME-LEVEL AUTOENCODER

Inspired from Kucherenko et al. (2019; 2021a) we used a Denoising Autoencoder (DAE) architecture (Vincent et al., 2010; Goodfellow et al., 2016) to reduce dimensionality at the frame level.

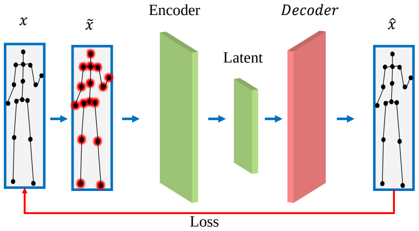


Figure 1: Pose DAE: Representation learning at the frame-level.

DAE learns a lower-dimensionality representation of data while corrupting the data intentionally through additive isotropic Gaussian noise. The bottleneck in the middle of DAE's encoder and decoder is generated from the noisy input, and the DAE learns to reconstruct the original input from it. DAEs can capture fundamental structures in the input distribution while preventing it from simply learning the identity function as each feature is encoded and decoded independently to the others (Vincent et al., 2010). The frame-level DAE includes an encoder and decoder as follows. The DAE $Encoder$ maps a noise injected input $\tilde{m}_f$ from pose space to the representation z_f and decodes the representation z_f back to a single frame $\hat{m_f}$ in motion space.

$$z_f = Encoder_{DAE}(\tilde{m}_f) \ ; \ \tilde{m}_f = m_f + N(0, I) \tag{1}$$

$$\hat{m} = Decoder_{DAE}(z_f) \tag{2}$$

As the DAE minimizes the reconstruction mean square error (MSE) loss, it learns a more informative representation to resemble the original input closely. Figure. 1 illustrates our representation learning for poses at the frame level.

3.2 Sequence-Level Autoencoder

In this step, we aim to create a vocabulary over the gesture space in which we will have a specific token for a set of similar motion sequences in the real world. We use a variational autoencoder (VAE) framework armed with a discrete latent representation (Oord et al., 2017). Given a set of observations, a Vector Quantized Variational Autoencoder (VQ-VAE) learns to create a motion vocabulary by parameterizing the posterior distribution of discrete latents.

In more detail, the VAE defines the posterior distribution as $p(z|x) \propto p(x|z)p(z)$. Typically, the prior $p(z)$ has been considered $z \sim N(o; I)$ on the latent variable $z \in \mathbb{R}^D$ where D is the bottleneck's dimensionality. Accordingly, the VAE encoder creates a posterior distribution $q(z|x)$ over the latent representation of input x. Meanwhile, for a chosen approximate posterior $q(z|x)$, the decoder is trained on the reparametrized sample $\tilde{z} \sim q(z|x)$ instead of deterministic encoded z (Kingma & Welling, 2013). The VAE arranges the latent space such that motions with similar movements are projected close to each other while reducing the reconstruction loss (Kingma & Welling, 2013; Stewart et al., 2021).

Additionally, we consolidated a denoising characteristic to the variational autoencoder framework (DVAE), presenting a more flexible and robust posterior distribution approximation than the standard VAE (Im Im et al., 2017). It can be considered as a standard VAE with the denoising criterion, which samples a noise injected input $\hat{x} \sim p(\hat{x}|x)$ rather than x itself. Afterward, it samples $\tilde{z} \sim q(z|\hat{x})$ using an encoder network, and samples the reconstructed input from the $p(x|z)$.

Furthermore, we discretize the latent space by decomposing it into a set of embedding vectors (Oord et al., 2017). Previous work has shown that VQ led to better representation learning, prevented mode collapse, and provided high reconstruction resolution (Oord et al., 2017; Razavi et al., 2019; Chorowski et al., 2019; Baevski et al., 2019). Additionally, discretization allows employing algorithms from the NLP community to model long-range temporal dependencies rather than imperceptible details at the frame level.

In this model, shown in Figure 2, the posterior $q(z|\hat{x})$ and prior $p(z)$ distributions are categorical. Indeed, samples from these distributions map to tokens from a codebook of K embedding vectors $e_i \in \mathbb{R}^D, i \in 1, 2, ..., K$. The discrete latent variables $z^\ddagger$ is determined from the continuous VAE's encoder output z. VQ-DVAE finds the z nearest neighbour embedding vector e_i in the K embedding vectors of $\mathbb{R}^D$. As shown in equation 3, the posterior categorical distribution $q^\ddagger(z|x)$

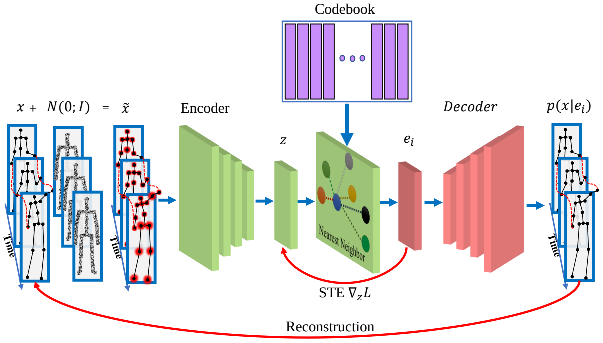


Figure 2: Motion VQ-DVAE: Discretized representation learning at the sequence-level.

probabilities represent a one-hot. Afterward, the corresponding embedding e_i is fed to the decoder for the reconstruction process.

$$q^\ddagger(z^\ddagger = k|x) = \begin{cases} 1 & \text{for } k \arg\min_j \|z \; - e_j\| \\ 0 & \text{Otherwise} \end{cases} \tag{3}$$

Despite the standard VAE, the prior is a uniform distribution over the K elements in the codebook. Therefore, the KL-divergence term usually incorporated into the ELBO is constant and safely removed. equation 4 defines the loss term for the VQ-DVAE from Oord et al. (2017). Since equation 3 is not differentiable, similar to the straight-through estimator (Bengio et al., 2013), we consider gradients from decoder inputs e_i for encoder outputs z.

$$\mathcal{L} = \log p(x|z^\ddagger) + \|sg[z] - e\|_2^2 2 + \beta\|z - sg[e]\|_2^2; \tag{4}$$

The first term stands for the reconstruction loss and trains the encoder and decoder. Although the velocity does not appear on the loss function explicitly, we feed the input concatenated with its derivative as input to the VQ-DVAE. In fact, we encourage the model to reconstruct not only the input sequence frame by frame, but also its derivative. This can be counted as a proxy for motion dynamics (Yoon et al., 2019). The second term encourages the embeddings to move closer to the encoder output. On the other hand, the third loss term, named commitment loss, encourages the encoder to generate latents close to the assigned embeddings from equation 3. Meanwhile, it prevents the volume of the embedding space from growing arbitrarily.

Note that sg means “stop-gradient”, so we do not propagate the gradient w.r.t. that term. Therefore, the first term updates both the encoder and decoder, the second term updates the codebook embeddings, and the third term updates the encoder to commit to its embedding.

3.3 Translation from Text to Gesture Vocabulary

Finally, having a gesture vocabulary, the co-speech gesture generation resembles a machine translation task of English to the Gesture domain. Consequently, we apply a purely autoregressive model of sequence-to-sequence known as a machine translation model. For instance, n-gram models (Jurafsky & Martin, 2009), convolutional neural language models (Dauphin et al., 2017; Bai et al., 2018) attention-based models (Vaswani et al., 2017; Wang et al., 2019), etc. yielded attractive results. We use a two-layered bidirectional gated recurrent neural network (GRU) (Cho et al., 2014), with a soft attention mechanism (Bahdanau et al., 2014). After the translation task of input text to gesture tokens is complete, we use the VQ-DVAE’s decoder to reproduce the entire gesture sequence followed by further post-processes such as the Savitzky-Golay smoothing filter (Savitzky & Golay, 1964).

4 Experiment

This section describes our implementation and experiment details, such as the specifics of the training data, model, and the training techniques we used. We obtained our training set from the Trinity dataset (Ferstl & McDonnell, 2018), which contains 23 clips, each approximately 10 minutes in length, with 244 minutes of aligned speech text, audio, and gesture training data in total. Using the GENEA challenge train and test set (Kucherenko et al., 2020b) also enables us to benchmark our results over provided test cases, including baselines and submitted models. Since this research focuses on gaining the most semantic information from text inputs, we do not use the provided audio data in our approach.

The dataset is captured with a 53 marker setup and 20 Vicon cameras at 59.95 frames per second (FPS). The motion data was stored as a time series of Euler rotations for each joint in the BioVision Hierarchy format (BVH) with 59.95 frames per second. Since we focus on the upper body, we obtain the corresponding 15 upper-body joints out of the available 69 body joints. Moreover, Euler angles were converted to joint positions in 3D space and normalized regarding shoulder length. We also down-sampled gestures to the frame rate of 20 FPS and removed finger motions due to the low accuracy of recordings. Although the dataset provides aligned word transcription in JSON format, it is inaccurate and might mislead the model. We hence applied the Gentle forced aligner algorithm (Ochshorn & Hawkins, 2017) to obtain exact timing information for each word.

4.1 DAE

We train our system on 15 joints from the upper body, including: Spine, Spine1, Spine2, Spine3, Neck, Neck1, Head, RightShoulder, RightArm, RightForeArm, RightHand, LeftShoulder, LeftArm, LeftForeArm, and LeftHand. Furthermore, we used 3x3 rotational matrices instead of 3D coordinates of each 15 joints to train the DAE network. Thus, the size of input and output vectors for pose representation at frame-level was 3x3x15=135. We also standardized each dimension to a mean of zero and a maximum absolute value of one for fast convergence in training.
The DAE consists of one linear layer of input size to the bottleneck dimensionality followed by a Tanh activation as an encoder and a linear layer of bottleneck to input size as a decoder and learning rate of 0.001. We injected Gaussian noise of standard deviation 0.1 to input data and the dropout rate was set to 0.2. To decide proper dimensionality for the bottleneck, we trained the model with different hidden dimensions fo 20 epochs. Similar to Thangthai et al. (2021); Kucherenko et al. (2019), we picked 40 dimensions to represent every single motion frame.

4.2 VQ-DVAE

In this section, we explain further details of the VQ-DVAE model as well as the training process. At this stage, we aimed to map a sequence of s_i to a latent representation. We selected sequences with length of 30 frames and a stride size of 10 frames. Also, considering the joint velocity as a proxy of motion dynamics (Yoon et al., 2019; Kucherenko et al., 2020a; 2021a), we concatenated the derivative of poses representations to the input . Therefore, we feed 40+40 = 80 features per frame to the autoencoder to encode and reproduce it. VQ-DVAE consists of an encoder, a quantized latent space embeddings as bottleneck, and a decoder. In order to learn sequential data, by stacking Bi-directional GRU networks on top of each other, we defined a multi-layer bi-directional recurrent neural network architecture for both encoder and decoder networks with the hidden size of 200. Consequently, we compress $80 * 30 = 2400$ dimensions to the continuous latent space of $200 * 2 = 400$ dimensions. Afterward, we quantize this variable z into the nearest neighbour embedding $e_i \in \mathbb{R}^D, D = 400, i \in 1, 2, ..., K$ and feed it to the decoder as discussed before. We empirically chose the number of embedding vectors k equal to 300.

4.3 Weakened decoders

As mentioned earlier, VAEs suffer from posterior collapse when a strong decoder network is used. One simple yet effective solution to this problem could be weakened decoders (Yang et al., 2017; Semeniuta et al., 2017). A conditional RNN decoder receives the last generated output frame as input to the current step. In our problem, continuity, short-range correlation, is a strong assumption (Alexanderson et al., 2020; Srivastava et al., 2015). For instance, the joint positions of a frame are the same as the previous one with subtle changes. A conditioned decoder simply determines these correlations and neglects extremely subtle movements requiring long-term dependencies (Srivastava et al., 2015). An unconditioned decoder that does not receive that input enforces the encoder to find this information and put it into the encoded vector. To learn a more robust and informative representation, we weaken the decoder by freezing its weights while minimizing the ELBO loss. Training the model with no gradient affecting the decoder enforces the encoder to learn informative and robust representation. In this setup, the decoder only propagates the encoder output through time, and the encoder carries reconstruction loss. However, after 20 epochs, we also trained the decoder in a conditioned fashion for another 20 epochs concerning smooth reproducibility at inference time.

5 Evaluation

In this section, we elucidate our comprehensive experiments to validate the effectiveness of the proposed method. We present an ablation study in the Appendix to substantiate the effectiveness of all the components of VQ-DVAE. We describe the baselines systems as well as the objective and subjective measures we used to assess our method.

5.1 Baseline Systems

We assessed our system against the ground-truth, and three recently published co-speech gesture generation systems. Our ground-truth was taken from the natural motions of the actor for a given speech segment. Baseline Text (BT) from Yoon et al. (2019) is the most similar work to ours since we both used Fasttext word vectors (Bojanowski et al., 2017) as our features from the input text transcript. Also, we both used the Bahaduna RNN Encoder-Decoder architecture with a soft attention mechanism (Bahdanau et al., 2014; Cho et al., 2014). The encoder extracts text features, and the decoder generates poses frame-by-frame. However, our sequence-to-sequence model maps extracted features from the text into a series of motion symbols, with each symbol representing 30 frames of motion. Baseline BA proposed by Kucherenko et al. (2019) used audio as input to generate co-speech gestures composed of an encoder-decoder structure. In this model, the encoder maps audio input to a sequence of learned pose representations while the decoder projects them back to poses. The third baseline model (BTA) (Korzun et al., 2020) achieved impressive results (Kucherenko et al., 2020b) using both audio and text modalities. In order to extract acoustic and textual features, BTA combined two separate encoders, one for each modality. This work was interesting since it used both modalities and scored the highest in human-likeness and second-highest in terms of appropriateness among (Yoon et al., 2019; Kucherenko et al., 2019; Alexanderson et al., 2020; Lu et al., 2021; Thangthai et al., 2021).

5.2 Objective Measures

Although improving subjective measures is our ultimate goal, they are costly, time-consuming, and require human labour. Moreover, evaluating generative models is tricky since there is not a one-and-only true motion sequence for a given speech utterance. Accordingly, we support our experimental results with numerical evaluations recently proposed to unveil gesture qualities.

Average jerk and velocity metrics have been used in prior work to quantitatively assess gesture systems (Kucherenko et al., 2020b). We used average jerk and velocity metrics proposed by Kucherenko et al. (2019). Jerk, the third derivative of joint positions, characterizes smoothness by calculating the rate of acceleration in a movement. We used the average over third derivative of joints as an objective metric to compare systems. We also posit that plausible generated gestures should follow similar velocity characteristics to the ground truth. Note that the gestures were converted from joint rotational angles to 3D joint positions. Hellinger distance (Hellinger, 1909) has extensively been used to compare two distributions. Accordingly, we compare each system to the ground truth to see how close the gestures are to the natural motion w.r.t that structural aspect.

Additionally, we introduce an objective metric to evaluate gesture diversity over the long term, thanks to our proposed discretized latent space at the sequence level. Indeed, we used our trained VQ-DVAE model to cluster predicted gestures for each condition at the sequence level. We adopted the Hellinger metric to evaluate how a system's output distribution follows the real-world gesture vocabulary distribution. The Hellinger distance metric was applied between each system and the ground truth to evaluate its closeness in terms of diversity.

5.3 Subjective Measures

Our ultimate purpose is to generate realistic gestures that are as valid and natural as the ground truth. Therefore, we conducted a human study to evaluate our performance subjectively. Comparable to

Table 1: Objective Evaluation Results

System	**Average Jerk**	**Hellinger Distance**		
		Velocity	***Acceleration***	***Diversity***
GT	1588.63 ± 2899.61	0	0	0
BT	795.09 ± 73.27	0.0988	0.1651	0.7495
BA	1275.43 ± 78.21	0.0793	0.1592	0.6298
BTA	904.19 ± 81.61	0.0710	0.1268	**0.5715**
Proposed	$\mathbf{1601.22 \pm 104.03}$	**0.0511**	**0.06465**	0.59184

Table 2: Subjective Evaluation Results

System	**Human-Likeness**		**Appropriateness**		**Diversity**	
	Mean	***STD***	***Mean***	***STD***	***Mean***	***STD***
GT	0.69	0.22	0.71	0.23	0.72	0.21
BT	0.41	0.23	0.31	0.21	0.43	0.20
BA	0.39	0.20	0.42	0.21	0.45	0.19
BTA	**0.50**	0.22	**0.53**	0.22	0.45	0.20
Proposed	**0.50**	0.20	0.45	0.22	**0.53**	0.18

the prior work in this area (Ginosar et al., 2019; Salvi et al., 2009; Kucherenko et al., 2020b; Alexanderson et al., 2020; Thangthai et al., 2021; Lu et al., 2021; Yoon et al., 2019; Korzun et al., 2020; Kucherenko et al., 2019), we evaluated generated gestures regarding perceived *human-likeness* and *appropriateness* of the motion given the speech context. These two measures indeed disentangle motion quality from the relevancy to the speech context. In addition, we also assessed systems in terms of diversity as opposed to repetitive movements over longer intervals. Specifically, we asked the following questions in our human-study: "How human-like does the motion appear?", "How appropriate are the gestures for the speech?" and "How diverse does the motion appear?".

For the appropriateness evaluation, we selected 40 speech segments each contains a complete sentence or phrase with an average of 10 seconds. Similarly, we selected 40 muted segments randomly for the human-likeness study and diversity study with a fixed length of 10 seconds and 20 seconds, respectively.

5.4 Experimental Procedure

We recruited participants from the Amazon Mechanical Turk (MTurk) to evaluate the results subjectively. MTurk is a crowdsourcing website proven effective in recruiting more diverse participants than in college experiments (Keith et al., 2017). MTurk provides the option to set requirements for the intended population based on different factors, i.e. age, gender, nationality, HIT approval rate. The HIT approval rate is defined as the percentage of completed work that other requestors have approved for that specific person. We posted a Human Intelligence Task (HIT) on MTurk containing an external link to our web interface. More details about our human-study interface are provided in the Appendix section. Participants were asked to complete the HIT, obtain a survey completion code, and enter it in MTurk to be compensated based on the minimum wage per hour in their country. We restricted our intended population to collect high-quality results by setting the minimum HIT approval rate to 95% and approved HITs greater than 5000. We also required participants to be located in Canada due to ethical review board policies.

6 Results

We recruited 24 participants and filtered out 6 judgments based on attention checks and ratings to the ground-truth gestures as a superior system over all conditions. The average age was 33.1 (STD=8.9) years with 12 men and 6 women, all English native speakers living in Canada. Among accepted judgments, the average experiment duration was approximately 41 minutes. The mean and standard deviation (STD) of ratings for the three studies are presented in Table 2.

Objective results on the 20 minute test set for all conditions is summarized in Table 3. We report the mean and standard deviation of the average jerk. Table 1 also presents the Hellinger distance of velocity and acceleration histograms to the ground truth. The diversity column shows the Hellinger distance on vocabulary frequency based upon assigned labels to constituent sequences (30 frames) obtained from VQ-DVAE.

7 Discussion

Gesture generation is challenging, especially in terms of appropriateness for the speech. Consequently, the difference between systems is subtle and a long way from the natural gestures. Objective metrics are consistent with the subjective results; however, we can see that different systems performed differently on evaluation metrics.

The subjective study shows that our model is preferred over all the systems in terms of human likeness. This is also aligned with similarity metrics in objective evaluation, where the proposed method ranked first in fundamental factors, i.e. jerk, acceleration, and velocity. Indeed, we can conclude that our representation learning method efficiently captured essential features and reproduced more natural gestures. Prior research has also shown that vector quantization led to more diverse and high-quality outputs. The BTA is the best in terms of appropriateness and our method is ranked second. We only used text features similar to the BT, while BTA used more advanced semantic features alongside acoustic features. We interpret it as the effectiveness of quantization against continuous regression resulted in better performance. Furthermore, raters perceived our model outputs as more diverse, showing our model's capability to overcome the mode collapse problem. As shown in Table 3, the proposed gesture generation method is significantly closer to the ground truth than the BT and approximately similar to BA and BTA in terms of the appearance of gesture sequences distribution.

Subjective and objective results suggest that we avoided the average gesture problem and generated more diverse and natural outputs than repetitive gestures. Vector quantization of motion representation narrowed the gesture space complication down to a set of vocabulary in a codebook that clusters similar sequences into a specific motion symbol. Although we significantly reduced the dimensionality with discrete encoding, both objective and subjective measurements show that generated gestures maintain a high quality in terms of naturalness. Quantization of motion representation space promotes the gesture generation model to focus on longer dependencies rather than local correlation. Consequently, we can apply machine translation algorithms from the NLP community to perform co-speech gesture generation tasks using entropy family loss instead of regression type.

We found that weakening the decoder strongly impacts the learned representation during the training process, resulting in better clusterings. However, its reproducibility was not smooth, and we later trained the decoder for 20 epochs while other parameters were fixed.

7.1 Limitations and Future Work

The first limitation of our work was that the dataset used in this research was limited to one person speaking in a monologue situation. Different persons have their own gesticulating style and different environments lead to complex speech and gestures. In future work we should consider datasets with mixed speakers in different environments. Another limitation of our research is that we trained our model on uppe,r body while fingers were excluded. Lower body is also important e.g. stepping forward and backward motions, standing still, approaching, facial expressions and fingers motions.

Although our system showed a significant improvement over baselines, it is still far behind human-generated motions. Kucherenko et al. (2020b) reported that raters were inclined to rate mismatched speech gestures from the ground truth, higher than synthesized gestures for a given input. Therefore, it is possible that we achieved a high appropriateness rate, especially compared to the Baseline Text (BT), due to the higher human-likeness quality of our system instead of appropriateness.

In this study, we did not involve audio and only used speech text. Punctuations such as question marks were also not provided in the source corpus. Therefore, generated gestures may lack movements relevant to acoustic features such intonation at the end of a sentence to indicate question (Pouw et al., 2020). Current work can be improved by adding punctuation using an automatic punctuation restoration system (Courtland et al., 2020). The proposed method can be extended by looking for correspondences between motions and audio features.

8 Conclusion

We proposed a fully unsupervised co-speech gesture generation system that combines representation learning methods and a machine translation algorithm. To the best of our knowledge, this is the first method that uses a representation learning algorithm at the sequence-level for the text-to-gesture generation task and provides a new state-of-the-art baseline in in this area. We connected a pose (frame-level) representation learning method and a quantized motion (sequence-level) representation learning, each trained separately. We also introduced a new objective evaluation metric that calculates the Hellinger distance of motions occurrence distributions as a measure for diversity. Finally, we trained a machine translation model to translate English sentences to gesture vocabulary. We found that the discretized motion space causes the model to focus on longer dependencies and generate more diverse and human-like gestures. The main limitation of our work was that we did not include audio features. In future work, we will consider more advance semantic features, audio features as well as emotional features, i.e. valence and arousal (Lim et al., 2011). The dataset was also not large enough to generalize a high-performance model, especially in terms of appropriateness. This can be addressed by creating a proper dataset through an automated process from currently available datasets. Furthermore, we will investigate hierarchical vector quantization (Razavi et al., 2019), to generate more natural and diverse gestures.

9 Reproducibility

Towards reproducibility of our results, we include here a link to our code implementing the proposed model as well as statistical analysis used in this paper:

`https://osf.io/xznb3/?view_only=8ba7c43f839242678e89f811e4763b6b`

We also provide the following link to visualizations and the human-study interface described in our Experiments section:

`https://osf.io/65q4p/?view_only=aa54d9fea4f1452a844a6acf6f6f2ac4`

Finally, the Appendix includes more details regarding the human study interface, in order for others to reproduce the subjective results.

References

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4432–4441, 2019.

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

Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. In *Computer Graphics Forum*, volume 39, pp. 487–496. Wiley Online Library, 2020.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pp. 153–160, 2007.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ international conference on intelligent robots and systems*, pp. 708–713. IEEE, 2005.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Brian Butterworth and Uri Hadar. Gesture, speech, and computational stages: A reply to mcneill. 1989.

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

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

Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. Predicting co-verbal gestures: a deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents*, pp. 152–166. Springer, 2015.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Jan Chorowski, Ron J Weiss, Samy Bengio, and Aäron van den Oord. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing*, 27(12):2041–2053, 2019.

Mingyuan Chu and Sotaro Kita. Co-thought and co-speech gestures are generated by the same action generation process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42 (2):257, 2016.

Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. URL `http://www.blender.org`.

Maury Courtland, Adam Faulkner, and Gayle McElvain. Efficient automatic punctuation restoration using bidirectional transformers with robust inference. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pp. 272–279, 2020.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR, 2017.

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

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

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

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

Ylva Ferstl, Michael Neff, and Rachel McDonnell. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics*, 89:117–130, 2020.

Ylva Ferstl, Michael Neff, and Rachel McDonnell. Expressgesture: Expressive gesture generation from speech through database matching. *Computer Animation and Virtual Worlds*, pp. e2016, 2021.

Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4346–4354, 2015.

Cristina Gârbacea, Aäron van den Oord, Yazhe Li, Felicia SC Lim, Alejandro Luebs, Oriol Vinyals, and Thomas C Walters. Low bit-rate speech coding with vq-vae and a wavenet decoder. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 735–739. IEEE, 2019.

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

Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. *Advances in neural information processing systems*, 22:646–654, 2009.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

John K Haas. A history of the unity game engine. 2014.

Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. Evaluation of speech-to-gesture generation using bi-directional lstm network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp. 79–86, 2018.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*, 2019.

Ernst Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271, 1909.

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

Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7134–7143, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Chien-Ming Huang and Bilge Mutlu. Robot behavior toolkit: generating effective social behaviors for robots. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 25–32. IEEE, 2012.

Daniel Im Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. Denoising criterion for variational auto-encoding framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Carlos T Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters*, 3(4):3757–3764, 2018.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009. ISBN 0131873210.

Melissa G Keith, Louis Tay, and Peter D Harms. Systems perspective of amazon mechanical turk for organizational research: Review and recommendations. *Frontiers in psychology*, 8:1359, 2017.

Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29:4743–4751, 2016.

Michael Kipp. *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers, 2005.

Vladislav Korzun, Ilya Dimov, and Andrey Zharkov. The finemotion entry to the genea challenge 2020. In *Proc. GENEA Workshop. https://doi. org/10*, volume 5281, 2020.

Taras Kucherenko. Data driven non-verbal behavior generation for humanoid robots. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 520–523, 2018.

Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pp. 97–104, 2019.

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

Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. The genea challenge 2020: Benchmarking gesture-generation systems on common data. 2020b.

Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *International Journal of Human–Computer Interaction*, pp. 1–17, 2021a.

Taras Kucherenko, Rajmund Nagy, Patrik Jonell, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. Speech2properties2gestures: Gesture-property prediction as a tool for generating representational gestures from speech. In *Proceedings of the 21th ACM International Conference on Intelligent Virtual Agents*, pp. 145–147, 2021b.

Taras Kucherenko, Rajmund Nagy, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. Multimodal analysis of the predictability of hand-gesture properties. *arXiv preprint arXiv:2108.05762*, 2021c.

David Lapakko. Communication is 93% nonverbal: An urban legend proliferates. *Communication and Theater Association of Minnesota Journal*, 34(1):2, 2007.

Sergey Levine, Christian Theobalt, and Vladlen Koltun. Real-time prosody-driven synthesis of body language. In *ACM SIGGRAPH Asia 2009 papers*, pp. 1–10. 2009.

Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. Gesture controllers. In *ACM SIGGRAPH 2010 papers*, pp. 1–11. 2010.

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. *arXiv preprint arXiv:2108.06720*, 2021.

Angelica Lim, Tetsuya Ogata, and Hiroshi G Okuno. Converting emotional voice to motion for robot telepresence. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*, pp. 472–479. IEEE, 2011.

Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)*, 39(4):40–1, 2020.

JinHong Lu, TianHang Liu, ShuZhuang Xu, and Hiroshi Shimodaira. Double-dcccae: Estimation of body gestures from speech waveform. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 900–904. IEEE, 2021.

James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. Understanding posterior collapse in generative latent variable models. 2019a.

James Lucas, George Tucker, Roger B Grosse, and Mohammad Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. *Advances in Neural Information Processing Systems*, 32:9408–9418, 2019b.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*, 2017.

Andy Lücking, Kirsten Bergman, Florian Hahn, Stefan Kopp, and Hannes Rieser. Data-based analysis of speech and gesture: The bielefeld speech and gesture alignment corpus (saga) and its applications. *Journal on Multimodal User Interfaces*, 7(1):5–18, 2013.

David McNeill. *Hand and mind*. De Gruyter Mouton, 2011.

Albert Mehrabian. *Nonverbal communication*. Routledge, 2017.

Lu Mi, Macheng Shen, and Jingzhao Zhang. A probe towards understanding gan and vae models. *arXiv preprint arXiv:1812.05676*, 2018.

Takashi Minato, Michihiro Shimada, Hiroshi Ishiguro, and Shoji Itakura. Development of an android robot for studying human-robot interaction. In *International conference on Industrial, engineering and other applications of applied intelligent systems*, pp. 424–434. Springer, 2004.

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

RM Ochshorn and Max Hawkins. Gentle forced aligner. *github. com/lowerquality/gentle*, 2017.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

Wim Pouw and James A Dixon. Quantifying gesture-speech synchrony. In *the 6th gesture and speech in interaction conference*, pp. 75–80. Universitaetsbibliothek Paderborn, 2019.

Wim Pouw, Steven J Harrison, and James A Dixon. Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony. *Journal of Experimental Psychology: General*, 149(2):391, 2020.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pp. 14866–14876, 2019.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

Giampiero Salvi, Jonas Beskow, Samer Al Moubayed, and Björn Granström. Synface—speech-driven facial animation for virtual speech-reading support. *EURASIP journal on audio, speech, and music processing*, 2009:1–10, 2009.

Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. A hybrid convolutional variational autoencoder for text generation. *arXiv preprint arXiv:1702.02390*, 2017.

B Series. Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radiocommunication Assembly*, 2014.

Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3310–3320, 2017.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pp. 843–852. PMLR, 2015.

Kenneth Stewart, Andreea Danielescu, Lazar Supic, Timothy Shea, and Emre Neftci. Gesture similarity analysis on event data using a hybrid guided variational auto encoder. *arXiv preprint arXiv:2104.00165*, 2021.

Ausdang Thangthai, Kwanchiva Thangthai, Arnon Namsanit, Sumonmas Thatphithakkul, and Sittipong Saychum. Speech gesture generation from acoustic and textual information using lstms. In *2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 718–723. IEEE, 2021.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1526–1535, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*, 2019.

Sarah Woods, Kerstin Dautenhahn, and Joerg Schulz. The design space of robots: Investigating children's views. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*, pp. 47–52. IEEE, 2004.

Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 265–281, 2018.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *International conference on machine learning*, pp. 3881–3890. PMLR, 2017.

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 4303–4309. IEEE, 2019.

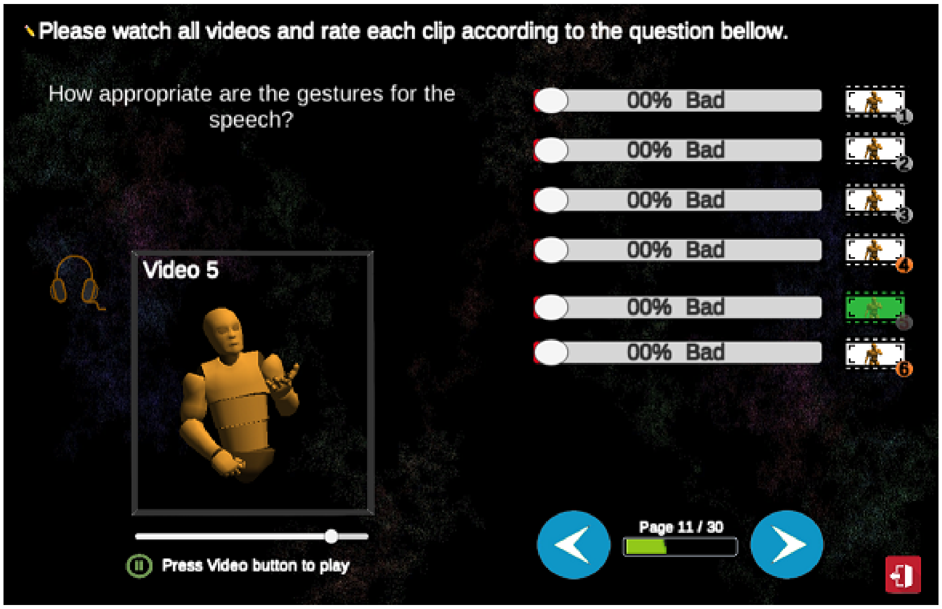
Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020.

A Appendix

A.1 Interface

The human-study interface, shown in Fig. 3, was implemented in Unity3D (Haas, 2014) and published in a WebGL format. We used Blender software (Community, 2018) to convert BioVision Hierarchy (BVH) files to a Filmbox (fbx) format that Unity3D can import as a humanoid animation. The avatar has 15 joints, excluding fingers in order to be consistent with our benchmarks.

The landing page of the experiment interface presented a consent form and instructions on how to use the interface with a quick guide tour. We randomly picked 10 clips from each of the three aforementioned segment pools to keep the experiment within 30 minutes and avoid exhausting raters. The first 10 clips shown to participants were from the human-likeness study. Next, raters were asked to use their headphone and test the audio settings to rate appropriateness stimuli. In the end, they were asked to rate 10 muted long clips from the diversity pool. To evaluate several systems simultaneously, we used a methodology similar to the (Kucherenko et al., 2020b) inspired by the Multiple Stimuli with Hidden Reference and Anchor test (Series, 2014). We used a page-wise strategy such that on each page, we included motions from all conditions corresponding to a specific segment assigned to that page. Therefore, we could employ pairwise statistical tests since stimuli were rated in parallel. We selected the 100-point rating scale and labelled them: "Bad", "Poor", "Fair", and "Excellent" within intervals of 20 points (Kucherenko et al., 2020b). Meantime, we randomized the order of clips on each page. After rating 30 stimuli, participants were asked to complete a demographic questionnaire. To see if a participant was paying enough attention and eliminate inattentive raters, we included an attention check on each page where we asked the participant to rate a specific value for a stimulus. On each page, the attention video and corresponding answer were selected randomly. Afterward, we excluded collected data with more than four failures in attention checks.



Please watch all videos and rate each clip according to the question bellow.

How appropriate are the gestures for the speech?

00% Bad

00% Bad

00% Bad

00% Bad

00% Bad

00% Bad

Figure 3: A screenshot of a page with stimuli from the human-study interface.

A.2 Ablation Study

To evaluate the essence of the contributions of each component in our method, we assess the latent representation space as well as the final output within an ablation study. To evaluate the latent representation, we analyze how two shifted sequences related to each other in latent space. We defined a Neighbour Sample Distance (NSD) metric, which measures the average distance between a sequence and its shifted version. We assume that in a good representation space, shifted samples' distances is smaller concerning the average distance in that space. We also apply the Fréchet gesture distance (FGD) Yoon et al. (2020) to compare distributions on the latent gesture space between

Table 3: Ablation Study Results

System	W	D	F	C	Der	VAE	VQ-VAE	Latent distance			FGD	Wasserstein
								10 Frames	*20 Frames*	*All*		
Proposed	30	✓	✓	✗	✓	✗	✓	0.50 ± 0.16	0.61 ± 0.16	24.64	4.83	18.69
Proposed - W	20	✓	✓	✗	✓	✗	✓	NA	NA	NA	7.95	22.36
Proposed - W	15	✓	✓	✗	✓	✗	✓	NA	NA	NA	6.62	37.02
Proposed - W	10	✓	✓	✗	✓	✗	✓	NA	NA	NA	7.84	18.31
Proposed - D	30	✗	✓	✗	✓	✗	✓	0.53 ± 0.16	0.63 ± 0.15	24.49	4.57	15.58
Proposed - F	30	✓	✗	✗	✓	✗	✓	0.35 ± 0.11	0.50 ± 0.13	24.39	7.71	22.12
Proposed - C	30	✓	✓	✓	✓	✗	✓	0.53 ± 0.14	0.68 ± 0.16	23.41	4.68	16.61
Proposed - Der	30	✓	✓	✗	✗	✗	✓	0.51 ± 0.15	0.62 ± 0.15	25.52	4.75	22.10
Proposed - VQ	30	✓	✓	✗	✓	✗	✓	0.75 ± 0.26	0.89 ± 0.24	43.65	5.23	22.13
Proposed - Vanilla	30	✗	✗	✓	✗	✗	✗	0.73 ± 0.20	0.78 ± 0.17	5.56	19.79	109.95
Proposed - D&Der	30	✗	✗	✓	✗	✗	✗	0.39 ± 0.14	0.53 ± 0.15	25.4	4.55	20.10

real and generated gestures. The more generated motions similar to the ground truth on the latent distribution, the smaller FGD value is.