ABSTRACT

During speech, people spontaneously gesticulate, which plays a key role in conveying information. Similarly, realistic co-speech gestures are crucial to enable natural and smooth interactions with social agents. Current end-to-end co-speech gesture generation systems use a single modality for representing speech: either audio or text. These systems are therefore confined to producing either acoustically-linked beat gestures or semantically-linked gesticulation (e.g., raising a hand when saying “high”): they cannot appropriately learn to generate both gesture types. We present a model designed to produce arbitrary beat and semantic gestures together. Our deep-learning based model takes both acoustic and semantic representations of speech as input, and generates gestures as a sequence of joint angle rotations as output. The resulting gestures can be applied to both virtual agents and humanoid robots. Subjective and objective evaluations confirm the success of our approach. The code and video are available at the project page svito-zar.github.io/gesticulator.

KEYWORDS

Gesture generation; virtual agents; socially intelligent systems; co-speech gestures; multi-modal interaction; deep learning

1 INTRODUCTION

When speaking, people often spontaneously produce hand gestures, also referred to as co-speech gestures. These co-speech gestures can accompany the content of the speech – what is being said – on all levels, from partial word meanings to situation descriptions [25].

Gestures and speech are both integral parts of human communication, and their combination can significantly enhance the conveyance of information. In this work, we present a framework for generating semantically-aware speech-driven gesture generation, which we call Gesticulator. The framework is designed to produce arbitrary beat and semantic gestures together, using both acoustic and semantic representations of speech as input. Subjective and objective evaluations confirm the success of our approach, and the code and video are available at the project page svito-zar.github.io/gesticulator.
While there are several theories on how gestures are produced by humans [5, 10, 33], there is a consensus that speech and gestures correlate strongly [18, 23, 31, 39]. In this section, we review some concepts relevant to our work, namely gesture classification, the temporal alignment between gestures and speech as well as the gesture-generation problem formulation.

2.1 Background

While there are several theories on how gestures are produced by humans [5, 10, 33], there is a consensus that speech and gestures correlate strongly [18, 23, 31, 39]. In this section, we review some concepts relevant to our work, namely gesture classification, the temporal alignment between gestures and speech as well as the gesture-generation problem formulation.

2.1.1 Co-Speech Gesture Types. Our work is informed by the gesture classification by McNeill [33], who distinguished the following gesture types:

1. **Iconic** gestures represent some aspect of the scene;
2. **Metaphoric** gestures represent an abstract concept;
3. **Deictic** gestures point to an object or orientation;
4. **Beat** gestures are used for emphasis and usually correlate with the speech prosody (e.g., intonation and loudness).

The first three gesture types, also called *representational gestures*, depend on the content of the speech – its semantics – while the last type instead depends on the audio signal – the acoustics. Hence, systems that ignore either aspect of speech can only learn to model a subset of human co-speech gesticulation.

2.1.2 Gesture-Speech Alignment. Gesture-speech alignment is an active research field covering several languages, including French [12], German [4], and English [18, 31, 39]. We focus on prior work on gesture-speech alignment for the English language.

In English, gestures typically lead the corresponding speech by, on average, 0.22 s (std 0.13 s) [31]; specifically, Pouw et al. [39] aligned different gesture types with the peak pitch of the speech audio and found that the onset of beat gestures usually precedes the corresponding speech by 0.35 s (std 0.3), the onset of iconic gestures precedes speech by 0.45 s (std 0.4), and the onset of pointing gestures precedes speech by 0.38 s (std 0.4).

Informed by these works, we take the widest range among the studies, plus some margin, for the time-span of the speech used to predict the corresponding gesture, and consider 1 s of future speech and 0.5 s of past speech as input to our model detailed in Sec. 4.

2.1.3 The Gesture-Generation Problem. We frame the problem of speech-driven gesture generation as follows: given a sequence of speech features \( s = \{s_t\}_{t=1:T} \) the task is to generate a corresponding pose sequence \( \tilde{g} = \{\tilde{g}_t\}_{t=1:T} \) of gestures that an agent might perform while uttering this speech. Here, \( t = 1 : T \) denotes a sequence of vectors for \( t \) in 1 to \( T \).

Each speech segment \( s_t \) is represented by several different features, such as acoustic features (e.g., spectrograms), semantic features (e.g., word embeddings) or a combination of the two. The ground-truth pose \( g_t \) and the predicted pose \( \tilde{g}_t \) at the same time instance \( t \) can be represented in 3D space as a sequence of joint rotations: \( g_t = [x_t, y_t, z_t]_{t=1:n} \) \( n \) being the number of keypoints of the body and \( a, \beta \) and \( \gamma \) representing rotations in three axes.

2.2 Related Work

As this work contributes toward data-driven gesture generation, we confine our review to these methods.

2.2.1 Audio-Driven Gesture Generation. Most prior work on data-driven gesture generation has used the audio-signal as the only speech-input modality in the model [14, 15, 19, 28, 42]. For example, Sadoughi and Busso [42] trained a probabilistic graphical model to generate a discrete set of gestures based on the speech audio-signal, using discourse functions as constraints. Hasegawa et al. [19] developed a more general model capable of generating arbitrary 3D motion using a deep recurrent neural network, applying smoothing as postprocessing step. Kuschenko et al. [28] extended this work by applying representation learning to the human pose and reducing the need for smoothing. Recently, Ginosar et al. [15] applied a convolutional neural network with adversarial training to generate 2D poses from spectrogram features. However, driving either virtual avatars or humanoid robots requires 3D joint angles. Ferslt et al. [14] followed the approach of adversarial training and applied it to a recurrent neural network together with a gesture phase classifier. Our model differs from these systems in that it leverages both the audio signal and the text transcription for gesture generation.

2.2.2 Text-Transcription-Driven Gesture Generation. Several recent works mapped from text transcripts to co-speech gestures. Ishi et al. [22] generated gestures from text input through a series of probabilistic functions: Words were mapped to word concepts using WordNet [34], which then were mapped to a gesture function (e.g., iconic or beat), which in turn were mapped to clusters of 3D hand gestures. Yoon et al. [48] learned a mapping from the utterance text to gestures using a recurrent neural network. The produced gestures were aligned with audio in a post-processing step. Although these works capture important information from text transcriptions, they may fail to reflect the strong link between gestures and speech acoustics such as intonation, prosody, and loudness [40].

2.2.3 Multimodal Gesture-Generation Models. Only a handful of works have used multiple modalities of the speech to predict matching gestures. The model in Neff et al. [37] predicted gestures based on text, theme, rhyme, and utterance focus. They also incorporated text-to-concept mapping. Concepts were then mapped to a set of 28 discrete gestures in a speaker-dependent manner. Chiu et al [9] used both audio signals and text transcripts as input, to predict a total of 12 gesture classes using deep learning. Our approach differs from these works, as we aim to generate a wider range of gestures: rather than predicting a discrete gesture class, our model produces arbitrary gestures as a sequence of 3D poses.

2.2.4 Regarding Motion Continuity. Separate from the input modalities of the system is the aspect of visual motion quality. Continuous
gesture generation can avoid the concatenation-point discontinuities exhibited by playback-based approaches such as motion graphs [3, 26]. That said, comparatively few approaches to continuous gesture generation explicitly try to enforce continuity in the generated pose sequence. Instead, they rely on postprocessing to increase smoothness as in [19]. Yoon et al. [48] include a velocity penalty in training that discourages jerkly motion. The recurrent connections used in several models [13, 19, 48] can also act as a pose memory that may help the model to produce smooth output motion. Autoregressive motion models have recently demonstrated promising results in probabilistic audio-driven gesture generation [2]. In this paper, we similarly investigate autoregressive connections for improving motion quality, which explicitly provide the most recent poses as input to the model when generating the next pose.

3 TRAINING AND TEST DATA

We develop our gesture generation model using machine learning: we learn a gesture estimator $\hat{g} = P(s)$ based on a dataset of human gesticulation, where we have both speech information $s$ (acoustic and semantic) and gesture data $g$. For this work, we specifically used the Trinity Gesture Dataset [13], comprising 244 minutes of audio and motion capture recordings of a male actor speaking freely on a variety of topics. We removed lower-body data, retaining 15 upper-body joints out of the original 69. Fingers were not modelled due to poor data quality.

To obtain semantic information for the speech, we first transcribed the audio recordings using Google Cloud automatic speech recognition (ASR), followed by thorough manual review to correct recognition errors and add punctuation for both the training and test parts of the dataset. The same data was used by the GENEVA 2020 gesture generation challenge1 and has been made publicly available in the original dataset repository2.

3.1 Test-Segment Selection

Two 10-minute recordings from the dataset were held out from training. We selected 50 segments of 10 s for testing: 30 random segments and 20 semantic segments, in which speech and recorded gestures were semantically linked. Three human annotators marked time instants where the recorded gesture was semantically linked with the speech content. Instances where all three annotators agreed (within 5 s tolerance) were used as semantic segments in our experiments.

3.2 Audio-Text Alignment

Text transcriptions and audio typically have different sequence lengths. To overcome this, we encode words into frame-level features as illustrated in Figure 2. First, the sentence, excluding filler words, is encoded by BERT [11], which is the state-of-the-art model in natural language processing (NLP). We encode filler words and silence, which do not contain semantic information, as special, fixed vectors $V_f$ and $V_s$, respectively. Filler words typically indicate a thinking process and can occur with a variety of gestures. Therefore, we set the text feature vector $V_f$ during filler words equal to the average of the feature vectors for the most common filler words in the data. Silence typically has no gesticulation [17], so the silence feature vector $V_s$ was made distinct from all other encodings, by setting all elements equal to −15. Finally, we use timings from the ASR system to nonuniformly upsample the text features, such that both text and audio feature sequences have the same length and timings. This is a standard text-speech alignment method in the closely-related field of speech synthesis [47].

4 SPEECH-DRIVEN GESTURE GENERATION

This section describes our proposed method for generating upper-body motion from speech acoustics and semantics.

4.1 Feature Types

We base our features on the state of the art in speech audio and text processing. Throughout our experiments, we use frame-synchronized features with 20 fps.

Like previous research in gesture generation [13, 15], we represent speech audio by log-power mel-spectrogram features. For this, we extracted 64-dimensional acoustic feature vectors using a window length of 0.1 s and hop length 0.05 s (giving 20 fps).

For semantic features, we use BERT [11] pretrained on English Wikipedia: each sentence of the transcription is encoded by BERT, resulting in 768 features per word, aligned with the audio as described in Sec. 3.2. We supplement these by five frame-wise scalar features, listed in Table 1.

To extract motion features, the motion-capture data was downsampled to 20 fps and the joint angles were converted to an exponential map representation [16] relative to a T-pose; this is common in computer animation. We verified that the resulting features did

---

1 genea-workshop.github.io/2020/#gesture-generation-challenge
2 trinityspeechgesture.scs.tcd.ie
not contain any discontinuities. Thereafter, we reduced the dimensionality by applying PCA and keeping 92% of the variance of the training data, similar to [48]. This resulted in 12 components.

4.2 Model Architecture and Training

We believe that a simple model architecture is preferable to a more complex one, everything else being equal. Hence, the intent of this work was to develop a straightforward model that solves the studied task. Figure 3 illustrates our model architecture. First, the text and audio features of each frame are jointly encoded by a feed-forward neural network to reduce dimensionality. To provide more input context for predicting the current frame, we pass a sliding window spanning 0.5 s (10 frames) of past speech and 1 s (20 frames) of future speech features over the encoded feature vectors. These time spans are grounded in research on gesture-speech alignment, as reviewed in Sec. 2.1.2. The encodings inside the context window are concatenated into a long vector and passed through several fully-connected layers. The model is also autoregressive: we feed preceding model predictions back to the model as can be seen in the figure, to ensure motion continuity. To condition on the information from the previous poses, we use FiLM conditioning [38], which generalizes regular concatenation. FiLM applies element-wise affine transforms $FiLM(x, \alpha, \beta) = x \ast \alpha + \beta$ to network activations $x$, where scaling $\alpha$ and offset $\beta$ vectors are produced by a neural net taking other information (here previous pose) as input. The final layer of the model and of the conditioning network for FiLM are linear to not restrict the attainable output range.

4.3 Training Procedure

We train our model on sequences of aligned speech audio, text, and gestures from the dataset. Each training sequence contains 70 consecutive frames from a larger recording. The first 10 and the last 20 frames establish context for the sliding window, while the 40 central frames are used for training. The model is optimized end-to-end for 100 epochs using stochastic gradient descent (SGD) and Adam [24] to minimize the loss function $\text{loss}(g, \hat{g}) = \text{MSE}(g, \hat{g}) + \lambda \text{MSE}(\Delta g, \Delta \hat{g})$, here $g$ and $\Delta g$ are the ground-truth position and velocity, $\hat{g}$ and $\Delta \hat{g}$ are the same quantities for the model prediction and MSE stands for Mean Squared Error. The weight $\lambda$ was set empirically to 0.6. Our velocity penalty can be seen as an improvement on the penalty used by Yoon et al. [48]. Instead of penalizing the absolute value of the velocity, we enforce velocity to be close to that of the ground truth.

During development, we observed that information from previous poses (the autoregression) tended to overpower the information from the speech: our initial model moved independently of speech input and quickly converged to a static pose. This is a common failure mode in generative sequence models, cf. [8, 20]. To counteract this, we pretrain our model without autoregression for the first seven epochs (a number chosen empirically), before letting the model receive autoregressive input. This pretraining helps the network learn to extract useful features from the speech input, an ability which is not lost during further training. Additionally, while full training begins without any teacher forcing (meaning that the model receives its own previous predictions as autoregressive input instead of the ground-truth poses), this is annealed over time: after one epoch, the model receives the ground-truth poses instead of its own prediction (for two consecutive frames) every 16 frames, which increased to every eight frames after another epoch, to every four frames after the next epoch, and then to every single frame after that. Hence, after five epochs of training with autoregression, our model has full teacher forcing: it always receives the ground-truth poses for autoregressive. This procedure greatly helps with learning a model that properly integrates non-autoregressive input.

4.4 Hyper-Parameter Settings

For the experiments in this paper, we used the hyper-parameter search tool Tune [30]. We performed random search over 600 configurations with velocity loss as the only criterion, obtaining the following hyper-parameters: Speech-encoding dimensionality 124 at each of 30 frames, producing 3720 elements after concatenation. The three subsequent layers had 612, 256, and 12 or 45 nodes (the output dimensionality with or without PCA). Three previous poses were encoded into a 512-dimensional conditioning vector. The activation function was tanh, the batch size was 64 and the learning rate $10^{-4}$. For regularization, we applied dropout with probability 0.2 to each layer, except for the pose encoding, which had dropout 0.8 to prevent the model from attending too much to past poses.

5 EVALUATION MEASURES

In this section we describe the objective and subjective measures we used in our experiments (Secs. 6 and 7).

5.1 Objective Measures

There is no consensus in the field about which objective measures should be used to evaluate the quality of generated gestures. As a step towards common evaluation measures for the gesture generation field, we primarily use metrics proposed by previous researchers. Specifically, we evaluated the average values of root-mean-square error (RMSE), acceleration and jerk (rate of change of acceleration), and acceleration histograms of the produced motion, in line with Kucherenko et al. [28]. To obtain these statistics, the gestures were converted from joint angles to 3D joint positions.

The acceleration and jerk were averaged over all frames for all 14 3D joints (except for the hips, which were fixed). To investigate the motion statistics in more detail, we also computed velocity histograms of the generated motion and compared those against
histograms derived from the ground-truth test data. We calculated the relative frequency of different velocity values over time-frames in all 50 test sequences, split into bins of width 1 cm/s.

5.2 Subjective Measures
To investigate human perception of the gestures we conducted several user studies that all followed the same protocol and procedure.

5.2.1 Experiment Design. We assessed the perceived human-likeness of the virtual character’s motion and how the motion related to the character’s speech using measures adapted from recent co-speech gesture generation papers [15, 48]. Specifically, we asked the questions “In which video...”: (Q1) “...are the character’s movements most human-like?” (Q2) “...do the character’s movements most reflect what the character says?” (Q3) “...do the character’s movements most help to understand what the character says?” (Q4) “...are the character’s voice and movement more in sync?”

We used attention checks to filter out inattentive participants. For four of the six attention checks, we picked a random video in the pair and heavily distorted either the audio (in the 2nd and 17th video pairs) or the video quality (in the 7th and 21st video pairs). Raters were asked to report any video pairs where they experienced audio or video issues, and were automatically excluded from the study upon failing any two of these four attention checks. In addition, the 13th and 24th video pairs presented the same video (from the random pool) twice. Here an attentive rater should answer “no difference”.

5.2.2 Experimental Procedure. Participants were recruited on Amazon Mechanical Turk (AMT) and assigned to one specific comparison of two systems; they could complete the study only once, and were thus only exposed to one system pair. Each participant was asked to evaluate 26 same-speech video pairs on the four subjective measures: 10 pairs randomly sampled from a pool of 28 random segments, 10 from a pool of 20 semantic segments, and 6 attention checks (see above). These video pairs were then randomly shuffled.

Every participant first completed a training phase to familiarize themselves with the task and interface. This training consisted of five items not included in the analysis, with video segments not present in the study, showing gestures of different quality. Then, during the experiment, the videos in each pair were presented side by side in random order and could be replayed as many times as desired. For each pair, participants indicated which video they thought best corresponded to a given question (one of Q1 through Q4 above), or that they perceived both videos to be equal in regard to the question.

6 ABLATION STUDY
In this section, we evaluate the importance of various model components by individually ablating them, training seven different system variants including the full model (see Table 2). Comparisons against other gesture-generation approaches are reported in Sec. 7.

6.1 Objective Evaluation
In this section we report objective metrics, as described in Sec. 5.1.

6.1.1 Average Motion Statistics. Table 3 illustrates acceleration and jerk, as well as RMSE, averaged over 50 test samples for the ground

<table>
<thead>
<tr>
<th>System</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model</td>
<td>The proposed method</td>
</tr>
<tr>
<td>No PCA</td>
<td>No PCA is applied to output poses</td>
</tr>
<tr>
<td>No Audio</td>
<td>Only text is used as input</td>
</tr>
<tr>
<td>No Text</td>
<td>Only audio is used as input</td>
</tr>
<tr>
<td>No FiLM</td>
<td>Concatenation instead of FiLM</td>
</tr>
<tr>
<td>No Velocity loss</td>
<td>The velocity loss is removed</td>
</tr>
<tr>
<td>No Autoregression</td>
<td>The previous poses are not used</td>
</tr>
</tbody>
</table>

Table 2: The seven system variants in the ablation study

<table>
<thead>
<tr>
<th>System</th>
<th>Accl. (cm/s²)</th>
<th>Jerk (cm/s³)</th>
<th>RMSE (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model</td>
<td>37.6 ± 4.3</td>
<td>830 ± 89</td>
<td>11.4 ± 11.8</td>
</tr>
<tr>
<td>No PCA</td>
<td>63.8 ± 8.3</td>
<td>1332 ± 192</td>
<td>13.0 ± 14.7</td>
</tr>
<tr>
<td>No Audio</td>
<td>26.9 ± 3.9</td>
<td>480 ± 67</td>
<td>11.3 ± 11.7</td>
</tr>
<tr>
<td>No Text</td>
<td>27.0 ± 1.9</td>
<td>715 ± 63</td>
<td>10.9 ± 11.3</td>
</tr>
<tr>
<td>No FiLM</td>
<td>44.2 ± 6.6</td>
<td>931 ± 181</td>
<td>11.0 ± 11.5</td>
</tr>
<tr>
<td>No Velocity loss</td>
<td>36.4 ± 4.1</td>
<td>779 ± 93</td>
<td>11.4 ± 12.3</td>
</tr>
<tr>
<td>No Autoregression</td>
<td>120.3 ± 19.2</td>
<td>3890 ± 637</td>
<td>11.2 ± 12.0</td>
</tr>
<tr>
<td>Ground truth</td>
<td>144.7 ± 36.6</td>
<td>2322 ± 538</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3: Objective evaluation of our systems: mean and standard deviation over 50 samples.
which we evaluated participants’ perception of a virtual character’s gestures as produced by the seven variants of our model described in Table 2. The experimental procedure and evaluation measures (see Sec. 5.2) were identical across all perceptual studies, including this one. Video samples from all systems in this study can be found at vimeo.com/showcase/6737868.

Figure 4: Velocity histograms of the wrist joints for the ablation study.

Figure 4a illustrates the velocity histogram of the wrist joints for the different model architectures and loss functions we considered. We observe two things: 1) the distributions are not influenced strongly by either FiLM conditioning or by velocity loss; and 2) autoregression reduces the amount of fast moments, making the velocity histogram more similar to the ground truth.

Velocity histograms for different input/output data are shown in Figure 4b. Removing PCA increases velocity, making the distribution more similar to the ground truth. In other words, training our model in the PCA space leads to reduced variability, which makes sense. We observe that excluding the text input makes the velocity smaller. This agrees with Table 3 and probably means that without semantic information the model produces mainly beat gestures, whose characteristics differ from other gesture types. While these numerical evaluations are valuable, they say very little about people’s perceptions of the generated gestures.

6.2 First Perceptual Study

To investigate human perception of the gestures we conducted several user studies. This section reports on Perceptual Study 1, in which we evaluated participants’ perception of a virtual character’s gestures as produced by the seven variants of our model described in Table 2. The experimental procedure and evaluation measures (see Sec. 5.2) were identical across all perceptual studies, including this one. Video samples from all systems in this study can be found at vimeo.com/showcase/6737868.

In the comparison of system ablations (Perceptual Study 1), 123 participants (μ age = 41.8 ± 12.3; 52 male, 70 female, 1 other) remained after exclusion of 477 participants who failed the attention checks, experienced technical issues, or stopped the study prematurely. The majority were from the USA (N = 120). Each sub-study had between 19 and 21 participants. We conducted a binomial test excluding ties with Holm-Bonferroni correction of p-values to analyze the responses. (24 responses that participants flagged for technical issues were excluded.) Our analysis was done in a double-blind fashion such that the conditions were obfuscated during analysis and only revealed to the authors after the statistical tests had been performed. The results are shown in Figure 5.

We can see from the evaluation of the “No Text” system that removing the semantic input drastically decreases both the perceived human-likeness of the produced gestures and how much they are linked to speech: participants preferred the full model over the one without text across all four questions asked with p < .0001. This confirms that semantics are important for appropriate automatic gesture generation.

The “No Audio” model is unlikely to generate beats, and might not follow an appropriate speech rhythm. Results in Figure 5 confirm this: participants preferred the full model over the one without audio across all four questions asked (p < .0001).

Removing autoregression from the model only affected perceived naturalness, where it performed significantly worse (p < .0001), as shown in Figure 5. This aligns with the findings from the objective evaluation: without autoregression the model produces jerky, unnatural-looking gestures, but the jerkiness does not influence whether gestures are semantically linked to the speech content.

There was no statistical difference between the Full model and the model without FiLM conditioning in terms of Q1 and Q2, but the model without FiLM was preferred with p < .02 for Q3 and p < .04 for Q4. This suggests that FiLM conditioning was not helpful for the model and regular concatenation worked better.

Removing the velocity penalty did not have a statistical difference on user responses, except for reducing user preference on Q4 with p < .04, suggesting that this component is not critical for the model.

The model without PCA gave unexpected results. In videos, we see that removing PCA improved gesture variability. While for human-likeness, there was no statistical difference, “No PCA” was significantly better (p < .0001) on Q2, Q3 and Q4 (see Figure 5). In summary, participants preferred the system without PCA, so it was chosen as our final model for the remaining comparisons (in Sec. 7).

6.3 Relation between objective and subjective evaluations

Objective and subjective evaluations each have their pros and cons. In this subsection we analyze the empirical correlation between the two for the experiments reported here.

From, the “No FiLM” and “No Velocity Loss” conditions, we see that user ratings did not change much for ablations that only produced minor changes in motion statistics. This is not surprising. For models with low jerk compared to the ground truth, we can see that participants preferred the models where the jerk was closer to that of the ground truth motion (“No PCA”). However, too high jerk
Figure 5: Results of Perceptual Study 1: comparing different ablations of our model in pairwise preference tests. Four questions, listed above each bar chart, were asked about each pair of videos. The bars show the preference towards the full model (higher values mean stronger preference) with 95% confidence intervals.

was associated with unnatural motion (“No Autoregression”). These results seem to indicate that jerk analysis provides information about the human-likeness of the motion.

7 ADDITIONAL EVALUATIONS AND COMPARISONS
The primary goal of this work is to develop the first model for continuous gesture generation that takes into account both the semantics and the acoustics of the speech. That said, we also benchmark our model against the state of the art in gesture generation. We compare the proposed approach to the model by Ginosar et al. [15], which is based on CNNs (convolutional neural networks) and GANs (generative adversarial networks), and therefore denoted CNN-GAN.

The hyper-parameters for the baseline method were fine-tuned by changing one parameter at a time and manually inspecting the visual quality of the resulting gestures on the validation dataset. The final hyper-parameters for the CNN-GAN [15] model were: batch size = 256, number of neurons in the hidden layer = 256, learning rate = 0.001, training duration = 300 epochs and $\lambda$ coefficient for the discriminator loss = 5. This tuned system was compared against the best system (“No PCA”) identified in Sec. 6.

Table 4: Objective comparison of our systems with the state-of-the-art: mean and standard deviation over 50 samples

<table>
<thead>
<tr>
<th>System</th>
<th>Accel. (cm/s²)</th>
<th>Jerk (cm/s³)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final model (no PCA)</td>
<td>63.8 ± 8.3</td>
<td>1330 ± 192</td>
</tr>
<tr>
<td>CNN-GAN [15]</td>
<td>254.7 ± 31.8</td>
<td>5280 ± 631</td>
</tr>
<tr>
<td>Ground truth</td>
<td>144.2 ± 35.9</td>
<td>2315 ± 530</td>
</tr>
</tbody>
</table>

7.1 Comparing with the state-of-the-art
Like the previous experiments, we follow the objective evaluation setup described in Sec. 5.1. Table 4 displays the average acceleration and jerk over 50 test sequences. We observe that the proposed method has acceleration and jerk values roughly half of those exhibited by the ground truth, while the CNN-GAN [15] baseline instead has twice the acceleration and jerk of the ground truth.

To investigate which model is preferred by human observers, we conducted another user study. We evaluated participants’ preference between the gestures as produced by the proposed models (No PCA) and CNN-GAN [15] (Perceptual Study 2). Video samples from this study can be found at vimeo.com/showcase/7127462.
7.2 Comparison with the Ground Truth

We also compared our model to the ground-truth gestures using the same procedure as before. In this study (Perceptual Study 3), 20 participants (μ age = 39.1 ± 8.4; 9 male, 11 female) remained after excluding 31 participants through the same criteria as in Perceptual Study 1. N = 18 were from the USA. There was a very substantial preference for the ground-truth motion (between 84 and 93%) across all questions. All differences were statistically significant according to Holm-Bonferroni-corrected binomial tests ignoring ties.

7.3 What Do “Semantic” Gestures Even Mean?

Finally, we evaluated if using text input helps our model to produce more semantically-linked gestures, such as iconic, metaphorical, and diegetic. To this end, we compared our best model (No PCA) with and without text information in the input: the first variant of this model received both audio and text, while the second one received only audio as input.

We asked three annotators to select which segments out of 50 test segments for both conditions that were semantically linked with the speech content. The annotators were all male and had an average age of 25.3 years. They were not aware of our research questions.

The results of this annotation were interesting and surprising: while all of them marked more gestures to be semantically linked with the speech content for the model that used text than the model without text (2 vs 0, 21 vs 9 and 9 vs 4), they had very low agreement: Cronbach’s alpha was below 0.5. The low agreement on which segment were semantic indicates that it is very subjective which gestures should be classified as semantically linked, which makes this and any similar evaluation challenging.

8 CONCLUSIONS AND FUTURE WORK

We have presented a new machine learning-based model for co-speech gesture generation. To the best of our knowledge, this is the first data-driven model capable of generating continuous gestures linked to both the audio and the semantics of the speech.

We evaluated different architecture choices and compared our model to an audio-based state-of-the-art baseline using both objective and subjective measures. All the study materials are publicly available at figshare.com/projects/Gesticulator/87128. Our findings indicate that:

1. Using both modalities of the speech – audio and text – can improve continuous gesture-generation models.
2. Autoregressive connections, while not commonplace in contemporary generation models, can enforce continuity of the gestures, without vanishing-gradient issues and with few parameters to learn. We also described a training scheme that prevents autoregressive information from overpowering other inputs.
3. PCA applied to the motion space (as used in [48]) can restrict the model by removing perceptually-important variation from the data, which may reduce the range of gestures.
4. The gestures from our model were preferred over the CNN-GAN [15] baseline by the study participants.

The main limitation of our work is that it requires an annotated dataset (with text transcriptions), which is labor-intensive. To overcome this, one could consider training the model directly on transcriptions from Automatic Speech Recognition.

Additionally, the vocabulary used in this dataset is sub-optimal.

As we can see in the frequency table provided at tinyurl.com/y22hört, out of the 50k total words there are 4230 unique words and the first 8 words account for 30% of all words spoken. This makes it challenging to learn semantic relations between gestures and text.

Future work also involves making the model stochastic (as in [2]), using larger datasets (such as [29]) and further improving the semantic coherence of the gestures, for instance by treating different gesture types separately.

ACKNOWLEDGEMENT

The authors would like to thank Andre Pereira, Federico Baldassarre and Marcus Klasson for helpful discussions. This work was partially supported by the Swedish Foundation for Strategic Research Grant No.: RIT15-0107 (EACare), by the Swedish Research Council projects 2017-05189 (CrowdVR) and 2018-05409 (StyleBot) and by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.