Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity

Preprint · September 2020

CITATIONS
0

READS
85

7 authors, including:

Youngwoo Yoon
Electronics and Telecommunications Research Institute
29 PUBLICATIONS **188** CITATIONS
SEE PROFILE

Joo-Haeng Lee
Electronics and Telecommunications Research Institute
74 PUBLICATIONS **478** CITATIONS
SEE PROFILE

Minsu Jang
Electronics and Telecommunications Research Institute
82 PUBLICATIONS **322** CITATIONS
SEE PROFILE

Jaeyeon Lee
Electronics and Telecommunications Research Institute
69 PUBLICATIONS **392** CITATIONS
SEE PROFILE

Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity

YOUNGWOO YOON, ETRI, KAIST
BOK CHA, University of Science and Technology, ETRI
JOO-HAENG LEE, ETRI, University of Science and Technology
MINSU JANG, ETRI
JAEYEON LEE, ETRI
JAEHONG KIM, ETRI
GEEHYUK LEE, KAIST

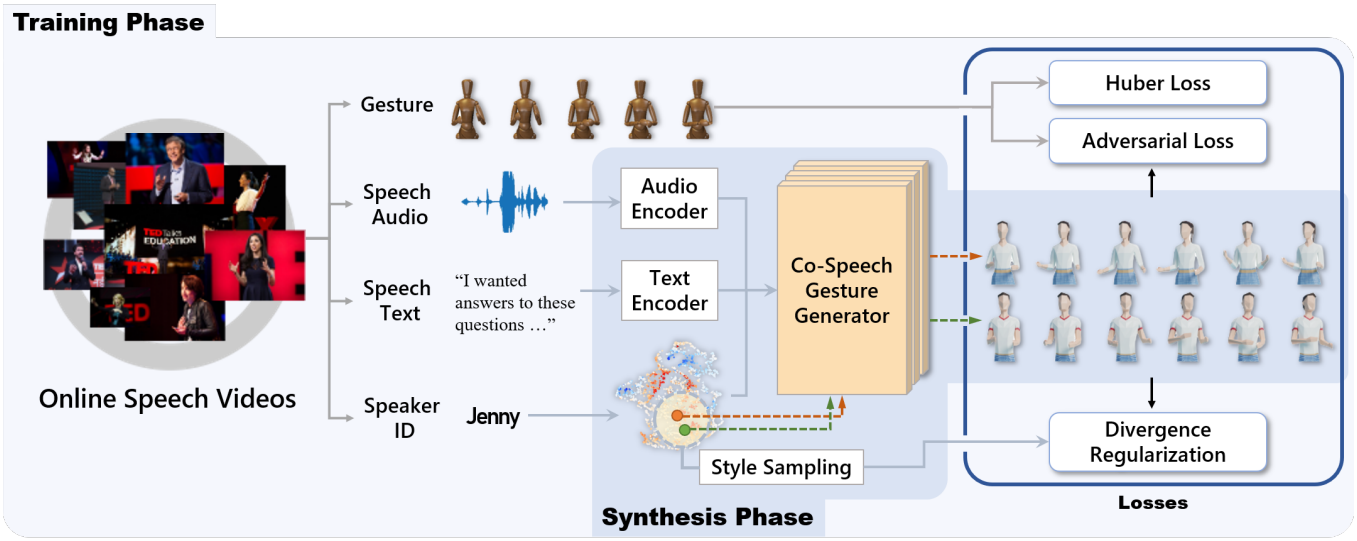

Fig. 1. Overview of the proposed gesture generation model that considers the trimodality of speech text, audio, and speaker identity. The model is trained on online speech videos demonstrating co-speech gestures. At the synthesis phase, we can manipulate gesture styles by sampling a style vector from the learned style embedding space.

For human-like agents, including virtual avatars and social robots, making proper gestures while speaking is crucial in human–agent interaction. Co-speech gestures enhance interaction experiences and make the agents look alive. However, it is difficult to generate human-like gestures due to the lack of understanding of how people gesture. Data-driven approaches attempt to learn gesticulation skills from human demonstrations, but the ambiguous and individual nature of gestures hinders learning. In this paper, we present an automatic gesture generation model that uses the multimodal context of speech text, audio, and speaker identity to reliably generate gestures. By incorporating a multimodal context and an adversarial training scheme, the proposed model outputs gestures that are human-like and that match with speech content and rhythm. We also introduce a new quantitative evaluation metric for gesture generation models. Experiments with the introduced metric and subjective human evaluation showed that the proposed gesture generation model is better than existing end-to-end generation models. We further confirm that our model is able to work with synthesized audio in a scenario where contexts are constrained, and show that different gesture styles can be generated for the same speech by specifying different speaker identities in the style embedding space that is learned from videos of various speakers. All the code and data is available at https://github.com/ai4r/Gesture-Generation-from-Trimodal-Context.

CCS Concepts: •**Computing methodologies** → **Animation;** *Supervised learning by regression;*

Additional Key Words and Phrases: nonverbal behavior, co-speech gesture, neural generative model, multimodality, evaluation of a generative model

ACM Reference format:
Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Trans. Graph.* 39, 6, Article 222 (December 2020), 16 pages.
DOI: 10.1145/3414685.3417838

DOI: 10.1145/3414685.3417838

1 INTRODUCTION

The continued development of graphics and robotics technology has prompted the development of artificial embodied agents, such as virtual avatars and social robots, as a popular interaction medium. One of the merits of the embodied agent is its nonverbal behavior, including facial expressions, hand gestures, and body gestures. In the present paper, we focus on upper-body gestures that occur with speech. Such co-speech gestures are a representative example of nonverbal communication between people. Appropriate use of gestures is helpful for understanding speech (McNeill 1992) and increases persuasion and credibility (Burgoon et al. 1990). Gestures are important not only in human–human interaction, but also in human–machine interaction. Gestures performed by artificial agents help a listener to concentrate and understand utterances (Bremner et al. 2011) and improve the intimacy between humans and agents (Wilson et al. 2017).

Interactive artificial agents, such as game characters, virtual avatars, and social robots, need to generate gestures in real time in accord with their speech. Automatically generating co-speech gestures is a difficult problem because machines must be able to understand speech, gestures, and the relationship between them. Two representative gesture generation methods are rule-based and data-driven approaches (Kipp 2005; Kopp et al. 2006). The rule-based approach, as the name suggests, defines various rules mapping speech to gestures; it requires considerable human effort to define the rules, but it is widely used in commercial robots because these models are relatively simple and intuitive. The data-driven approach learns gesticulation skills from human demonstrations. This approach requires more complex models and large amounts of data, but they do not require human effort in designing rules. As large gesture datasets are becoming more available, research on data-driven approaches is increasing, e.g., (Chiu et al. 2015; Ginosar et al. 2019; Huang and Mutlu 2014; Kipp 2005; Yoon et al. 2019).

One data-driven approach, called the end-to-end method (Ginosar et al. 2019; Yoon et al. 2019), is unlike others in that it uses raw gesture data without intermediate representation such as predefined unit gestures. Such less restrictive representation increases the method's expressive capacity, enabling it to generate more natural gestures. Previous studies have successfully demonstrated end-to-end gesture generation methods. However, they were limited by their consideration of only a single modality, either speech audio or text. Since human gestures are associated with various factors, such as speech content, speech audio, interlocutor interaction, individual personality, and surrounding environment, generating gestures from a single speech modality can produce a very limited model. In the study of human gestures (McNeill 1992), researchers have defined four categories, called iconic, metaphoric, deictic, and beat gestures, which are related to different contexts. Iconic gestures illustrate physical actions or properties (e.g., raising one's hands while saying "tall") and metaphoric gestures describes abstract concepts (e.g., moving one's hands up and down to depict a wall while saying "constraint"). Both iconic and metaphoric gestures are highly related to the speech lexicon. Deictic gestures are indicative motions that point to a specific target or space, and are related to both the speech lexicon and the spatial context in which the gesture is made. Beat gestures are rhythmic movements that are closely related to the speech audio. In addition, even with the same speech and in the same surrounding environment, each person makes different gestures every time due to inter-person and intra-person variability of human gestures, and the inter-person variability may be attributed to individual personality. Various modalities related to speech should be considered in order to generate more meaningful and human-like gestures.

In the present study, we propose an end-to-end gesture generation model that uses the multimodal context of text for speech content, audio for speech rhythm, and speaker identity (ID) for style variations. To integrate these multiple modalities, a temporally synchronized encoder–decoder architecture is devised based on the property of temporal synchrony found between speech and gestures in human gesture studies (Chu and Hagoort 2014; McNeill 2008). We experimentally confirm that each modality is effective. Especially, a style embedding space is learned from speaker IDs to reflect inter-person variability, so we can create different styles of gestures for the same speech by sampling different points in the style embedding space. Figure 1 provides an overview of the proposed gesture generation model and its training. The model is trained on a dataset derived from online videos exhibiting speech gestures with a training objective to generate human-like and diverse gestures. Our task is to develop a general gesture generator, a model that is supposed to generate convincing gestures for previously unseen speech.

A major hurdle in gesture generation studies is determining how to evaluate results. There is no single ground truth in gesture generation and well-defined evaluation methods are not yet available. Subjective human evaluation is the most reasonable method, but it is not cost effective and difficult to reproduce results. Some studies have used the mean absolute error (MAE) of the positions of body joints between human gesture examples and generated gestures for the same speech (Ginosar et al. 2019; Joo et al. 2019). The MAE evaluation method is objective and reproducible, though it is hard to ascertain to what extent the MAE between joints correlates with perceived gesture quality. In the present paper, we apply the Fréchet inception distance (FID) concept proposed in image generation research (Heusel et al. 2017) to our problem of gesture generation. FID compares fitted distributions on a latent image feature space between the sets of real and generated images. We introduce the Fréchet gesture distance (FGD), which compares samples on a latent gesture feature space. With synthetic noisy data and comparing to human judgements, we validate that the proposed metrics are more perceptually plausible than computing the MAE between gestures.

Our contributions can be summarized as follows:

- A new gesture generation model using a trimodal context of speech text, audio, and speaker identity. To the best of our knowledge, this is the first end-to-end approach using trimodality to generate co-speech gestures.
- The proposal and validation of a new objective evaluation metric for gesture generation models.
- Extensive experiments to verify the usability of the proposed model. We show style manipulations with the trained style embedding space, the model's response to altered speech text, and the gestures' incorporation with synthesized audio.

The remainder of this paper is organized as follows. We first introduce related research (Section 2), then describe the proposed model (Section 3) and its training in detail (Section 4). Section 5 introduces a metric for evaluating gesture generative models and Section 6 describes human evaluation to validate the proposed metric. Section 7 presents qualitative and quantitative results. Finally, Section 8 concludes the paper with a discussion of the limitations and future direction of the present research.

2 RELATED WORK

We first review automatic co-speech gesture generation methods for artificial agents. Next, we introduce previous data-driven gesture generation approaches. Related work discussing gesture styles, multimodality, and evaluation methods are also introduced.

Co-speech Gesture Generation for Artificial Agents. Motion capture and retargeting human motions to artificial agents is widely used to generate motions, especially in commercial systems, because of its high-quality motion from human actors (Menache 2000). Nonverbal behavior can also be generated by retargeting human motion (Kim and Lee 2020). However, the motion capture method has a critical limitation: the motion should be recorded beforehand. Therefore, the motion capture method can only be used in movies or games that have specified scripts. Interactive applications, in which the agents interact with humans with various speech utterances in real time, mostly use automatic gesture generation methods. The typical automatic generation method is rule-based generation (Cassell et al. 2004; Kopp et al. 2006; Marsella et al. 2013). For example, the robots NAO and Pepper (Softbank 2018) have a predefined set of unit gestures and have rules that connect speech words and unit gestures. This rule-based method requires human effort to design the unit gestures and hundreds of mapping rules. Research into data-driven methods has aimed to reduce the human effort required for rule generation; these methods find gesture generation rules in data using machine learning techniques. Probabilistic modeling for speech–gesture mapping has also been studied (Huang and Mutlu 2014; Kipp 2005; Levine et al. 2010) and a neural classification model selecting a proper gesture for given speech context (Chiu et al. 2015) was also proposed. The review paper (Wagner et al. 2014) provides a comprehensive summary of the gesture generation research and rule-based approaches.

End-to-end Gesture Generation Methods. Gesture generation is a complex problem that requires understanding speech, gestures, and their relationships. To reduce the complexity of this task, previous data-driven models have divided speech into discrete topics (Sadoughi and Busso 2019) or represented gestures as predefined unit gestures (Huang and Mutlu 2014; Kipp 2005; Levine et al. 2010). However, with recent advancements in deep learning, an end-to-end approach using raw gesture data is possible. There are studies using the end-to-end approach (Ferstl et al. 2019; Ginosar et al. 2019; Kucherenko et al. 2019, 2020; Yoon et al. 2019) that have formulated gesture generation as a regression problem rather than a classification problem. This continuous gesture generation does not require crafting unit gestures and their rules and also removes the restriction that gesture expressions must be selected from predetermined unit gestures.

One study used an attentional Seq2Seq network that generates a sequence of upper body poses from speech text (Yoon et al. 2019). The network consists of a text encoder that processes speech text and a gesture decoder that generates a pose sequence. Other studies generated gestures from speech audio (Ferstl et al. 2019; Ginosar et al. 2019; Kucherenko et al. 2019). These audio-based generators also based on the neural architectures generating a sequence of poses, and some studies used adversarial loss to guide generated gestures to become similar to actual human gestures. The main difference between the previous models is the use of different speech modalities. Both semantics and acoustics are important for generating co-speech gestures (McNeill 1992), so, in this paper, we propose a model that uses multimodal speech information, audio and text together. Note that there is a concurrent work considering both audio and text information, but it trained and validated the generative model on a limited dataset of a single actor (Kucherenko et al. 2020).

Learning Styles of Gestures. People make different gestures even when they say the same words (Hostetter and Potthoff 2012). Similarly, artificial agents must also learn different styles of gestures. The agents should be able to make extrovert- or introvert-style gestures according to their emotional states, interaction history, user preferences, and other factors. Stylized gestures also give the agents a unique identity similar to appearances and voices. Previous studies have attempted to generate such stylized gestures (Ginosar et al. 2019; Levine et al. 2010; Neff et al. 2008). In these studies, generative models were trained separately for each speaker or style. This approach is an obvious way of learning individual styles, but requires a substantial amount of training data for each individual style. Because of this limitation, only three and ten individual styles were trained in (Levine et al. 2010) and (Ginosar et al. 2019), respectively. In the present study, we aim to build a style embedding space, so that we can manipulate styles through sampling the space into which different styles are embedded, rather than replicating a particular style as the previous papers did. Another study proposed more detailed style manipulation by using control signals of hand position, motion speed, or moving space (Alexanderson et al. 2020).

Processing Multimodal Data. The present study considers four modalities: text, audio, gesture motion, and speaker identity. Generally, multimodal data processing includes the representation of each modality, alignment between modalities, and translation between modalities (Baltrušaitis et al. 2018). There are two approaches to representation: one is that all modalities share the same representation and the other is that modalities are represented separately, and later alignment or translation stages integrate them. We can find both representation approaches related to gesture generation. A study by (Ahuja and Morency 2019) represented both human motion and descriptive text as vectors in the same embedding space. In other studies, different representations are used for different modalities (Roddy et al. 2018; Sadoughi and Busso 2019). We use separate representations, owing to the difficulty of learning a cross-modal representation for co-speech gestures arising from the weak and ambiguous relationship between speech and gestures.

Alignment between modalities is also an important factor for time-series data. In (Ginosar et al. 2019), a feature vector encoding input speech was passed to a decoder to generate gestures, and the

alignment between the modalities is not explicitly handled. A neural encoder and decoder implicitly processed the alignment as well as the translation from speech to gesture. In (Yoon et al. 2019), a similar encoder–decoder architecture was used, but they guided the model to learn sequential alignment more explicitly by incorporating an attention mechanism (Bahdanau et al. 2015). In (Kucherenko et al. 2020), speech audio and text were aligned but not with gestures. Our model uses explicitly aligned speech and gesture because speech and gesture are synchronized temporally (Chu and Hagoort 2014), allowing the network to concentrate on the translation from input speech to gestures.

Evaluating Generative Models. Recently, as research into generative models has expanded, interest in evaluating generative models has increased. In a generation problem considering speech synthesis, image generation, and conversational text generation, human evaluation is the most plausible evaluation method because there is no clear ground truth to compare with. However, the results of human evaluation cannot easily be reproduced. A reliable computational evaluation metric is necessary for reproducible comparisons with state-of-the-art models and would accelerate research. Previous studies have measured gesture differences between generated and human gestures (Ginosar et al. 2019; Joo et al. 2019), though this method is limited because pose-level differences do not measure the perceptual quality of the generated gestures. Some studies have used other metrics to evaluate human motion, for example, the motion statistics of jerk and acceleration (Kucherenko et al. 2019) and Laban parameters from a study of choreography (Aristidou et al. 2015). However, the aforementioned metrics compute distances for each sample, so they cannot measure how the generated results are diversified, which is crucial in generation problems. In the image generation problem, the inception score (Salimans et al. 2016) and FID (Heusel et al. 2017) have recently become de facto evaluation metrics because they can measure the diversity of generated samples as well as their quality, and this concept was successfully applied to other generation problems (Kilgour et al. 2018; Unterthiner et al. 2019). In this study, we have applied the concept of FID to the gesture generation problem to measure both perceptual quality and diversity.

3 METHOD

3.1 Overall Architecture

Gesture generation in this paper is a translation problem that generates co-speech gestures from a given speech context. Our goal is to generate gestures that are human-like and match well with any given speech. We propose a neural network architecture consisting of three encoders for input speech modalities and a decoder for gesture generation. Figure 2 shows the overall architecture. Three modalities—text, audio, and speaker identity (ID)—are encoded with different encoder networks and transferred to the gesture generator.

A gesture is represented as a sequence of human poses, and the generator, which is a recurrent neural network, generates poses frame-by-frame from an input sequence of feature vectors containing encoded speech context. Speech and gestures are temporally synchronized (Chu and Hagoort 2014; McNeill 2008), so we configured the generator to use part of the speech text and audio near

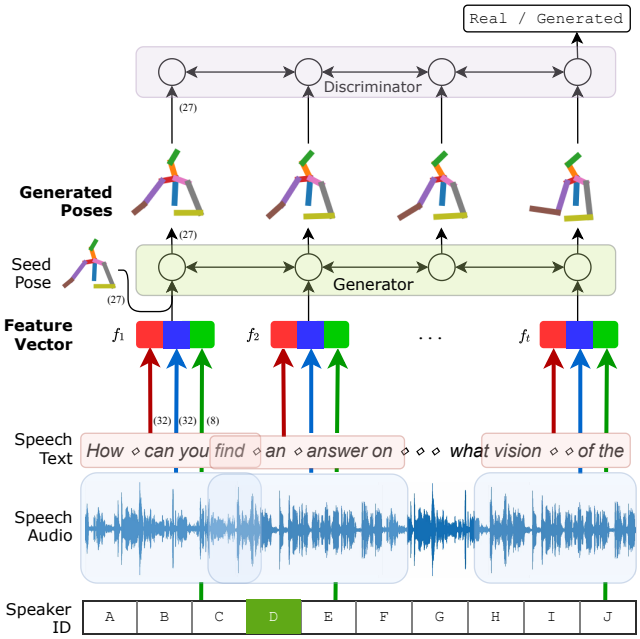

Fig. 2. The architecture of the proposed gesture generation model. The generator generates a sequence of human poses from a sequence of context feature vectors that contain the encoded features of speech text, speech audio, and speaker identity (ID). The features of text, audio, and speaker ID are depicted as red, blue, and green arrows, respectively. The seed poses are also used to ensure continuity between consecutive syntheses. The discriminator is a binary classifier that distinguishes between real human gestures and generated gestures. The number in parentheses indicates the data dimension. The poses are in 27 dimensions since there are nine directional vectors in 3D coordinates.

the current time step instead of the whole speech context. Gesture style does not change in the short term, so the same speaker ID is used throughout the synthesis. In addition, we used seed poses for the first few frames for better continuity between consecutive syntheses. See appendix A for the figures of detailed architecture.

3.2 Encoding Speech Context

This section describes how the speech modalities of text, audio, and speaker ID are represented and the details of the encoder networks. We have four modalities, including the output gesture, in different time resolution. We first ensure that all input data have the same time resolution as the output gestures, so all modalities share the same time steps and the proposed sequential model (Figure 2) can process speech input and generate poses frame by frame.

The speech text is a word sequence, with the number of words varying according to speech speed. We insert padding tokens (◇) into the word sequence to make a padded word sequence ($word_1$, $word_2$, ..., $word_t$) that is the same length as the gestures. Here, t is the number of poses in a synthesis (fixed as 34 throughout the paper, see Section 4). We assume the exact utterance time of words is known, so the padding token is inserted to make the words temporally match the gestures. For instance, for the speech text *"I*

love you", if there were a short pause between *"I"* and *"love"*, then the padded word sequence would be *"I ⋄ ⋄ love you"* when t is 5. All words in the padded word sequence are then transformed into word vectors in 300 dimensions via a word embedding layer. Next, these word vectors are encoded by a temporal convolutional network (TCN) (Bai et al. 2018) to make 32-D feature vectors for speech text modality (f_1^{text}, f_2^{text}, ..., f_t^{text}). TCN processes sequential data through convolutional operations, and showed competitive results over the recurrent neural networks in diverse problems (Bai et al. 2018). In this paper, we used a four-layered TCN, where each f^{text} has a receptive field of 16. Thus, f_i^{text} encodes 16 padded words around at time step i. For our training dataset the average and the largest number of non-padding words in this receptive field were 3.9 and 16, respectively.

We used FastText (Bojanowski et al. 2017), a pretrained word embedding, and update these embeddings during training. There was the concern that word embeddings pretrained by filling a missing word in a sentence (Mikolov et al. 2013) may not suitable to gesture generation. For instance, if we query words that are close to *large*, then *small* appears in the top-3 list in both GloVe (Pennington et al. 2014) and FastText (Bojanowski et al. 2017) even though they have opposite meanings. This problem with pretrained word embedding has also been raised in text-based sentiment analysis, where the sentiment of words is important (Fu et al. 2018). We tested three different settings: 1) pretrained embeddings without weight updating, 2) pretrained embeddings with fine-tuning weights, and 3) learning word embeddings from scratch. In our problem, using pretrained embeddings with fine-tuning was the most successful. FastText (Bojanowski et al. 2017) was favored over GloVe (Pennington et al. 2014) since FastText is using subword information so that it gives accurate representation for unseen words.

For the speech audio modality, a raw audio waveform goes through cascaded one-dimensional (1D) convolutional layers to generate a sequence of 32-D feature vectors (f_1^{audio}, f_2^{audio}, ..., f_t^{audio}). Audio frequency is usually fixed, so we adjusted the sizes, strides, and padding in the convolutional layers to obtain equally many audio feature vectors as there were output motion frames. In our experiments, each feature vector had a receptive field of about a quarter of a second. The quarter-second receptive field may not be large enough to cover occasional asynchrony between speech and gesture (the standard deviation of the temporal differences is about a half second according to (Bergmann et al. 2011)), but our use of a bidirectional GRU in the gesture generator that sends information forwards and backwards can compensate for the asynchrony.

The model also uses speaker IDs to learn a style embedding space. Human gestures are not the same even for the same speech. We utilize the speaker IDs to reflect characteristics of each speaker in the dataset, and we call this individuality as 'style' in the present paper. Note that our purpose is to build an embedding space capturing different styles not to replicate gestures of each speaker. The speaker IDs are represented as one-hot vectors where only one element of a selected speaker is nonzero. A set of fully connected layers maps a speaker ID to a style embedding space of much smaller dimension (8 in the present study). To make the style embedding space more interpretable, variational inference (Kingma and Welling 2014; Rezende et al. 2014) that uses a probabilistic sampling process is used. The same feature vector f^{style} on the style embedding space is used for all time steps in a synthesis.

3.3 Gesture Generator

The generator $G(\cdot)$ takes encoded features as input and generates gestures. The gesture is a sequence of human poses p_i consisting of 10 upper body joints (spine, head, nose, neck, L/R shoulders, L/R elbows, and L/R wrists). All poses were spine-centered. When we train the model, we represent each pose as directional vectors which represent the relative positions of the child joints from the parent joints. There are nine directional vectors for spine–neck, neck–nose, nose–head, neck–R/L shoulders, R/L shoulders–R/L elbows, and R/L elbows–R/L wrists. The directional vectors are favored for training the proposed model because this representation is less affected by bone lengths and root motion. In the representation of joint coordinates, a small translation of neck, which is the parent joint of both arms, can have an excessive effect on all coordinates of the arms. We denote human poses represented as directional vectors by d_i, and all directional vectors were normalized to the unit length. We note that forearm twists were not considered in this paper.

For gesture generation, we use a multilayered bidirectional gated recurrent unit (GRU) network (Cho et al. 2014). Encoded features of speech text, audio, and speaker ID are concatenated to form a concatenated feature vector $f_i = (f_i^{\text{text}}, f_i^{\text{audio}}, f^{\text{style}})$ for each time instant i. The generator takes the feature vector f_i as input and generates the next pose $\hat{d}_{i+1}$ iteratively.

For a long speech, the speech is divided into 2-second chunks and the generator synthesizes gestures for each chunk. The use of seed poses helps to make transitions between consecutive syntheses smooth. Seed poses $d_{i=1,\ldots,4}$, the last four frames of the previous synthesis, are concatenated with the feature vector for the early four frames of the next synthesis as (f_i, d_i), and an additional bit is used to indicate the presence of a seed pose.

3.4 Adversarial Scheme

An adversarial scheme (Goodfellow et al. 2014) is applied in training the model to generate more realistic gestures. The adversarial scheme uses a discriminator, which is a binary classifier distinguishing between real and generated gestures. By alternate optimization of generator and discriminator, the generator improves its performance to fool the discriminator. For the discriminator, we use a multilayered bidirectional GRU that outputs binary output for each time step. A fully connected layer aggregate the t binary outputs and gives a final binary (real or generated gesture) decision.

4 TRAINING WITH "IN-THE-WILD" VIDEOS

4.1 TED Gesture Dataset

The gesture generation model is trained on the TED gesture dataset (Yoon et al. 2019), which is a large-scale, English-language dataset for data-driven gesture generation research. The dataset includes speech from various speakers, so it is suitable for learning individual gesture styles. We added 471 additional TED videos to the data of (Yoon et al. 2019), for a total of 1,766 videos. Extracted human poses from TED videos, speech audio, and transcribed English speech text are available. We further converted all human poses to 3D by using the

3D pose estimator (Pavllo et al. 2019) which convert a sequence of 2D poses into 3D poses. The pose estimator uses temporal convolutions that lead to temporally coherent results despite of a few of inaccurate 2D poses. We used the manual speech transcriptions available on each TED talk, with onset timestamps of each word extracted using the Gentle forced aligner (Ochshorn and Hawkins 2016) to insert padding tokens. The forced aligner reported successful alignment of 97% of the total words.

From the videos, only the sections of videos in which upper body gestures were clearly visible were extracted; the total duration of the valid data was 97 h. The gesture poses were resampled at 15 frames per second, and each training sample having 34 frames was sampled with a stride of 10 from the valid video sections. The initial four frames were used as seed poses and the model was trained to generate the remaining 30 poses (2 seconds). We excluded non-informative samples having little motion (i.e., low variance of a sequence of poses) and erratic samples having lying poses (i.e., low angle of the spine–neck vector).

The dataset was divided into training, validation, and test sets. The division was done at the video level. Because all presentations in the TED dataset were given by different speakers, the number of unique speaker IDs is the same as the number of videos and there is no overlap of speaker IDs between split sets. We used the training set for training the model, the validation set for tuning the systems, and the test set for qualitative results and human evaluation. The final number of 34-frame sequences in each data partition were 199,384; 26,795; and 25,930.

4.2 Training Loss Function

The model is trained using the losses below. We use L_G to train the encoders and gesture generator and L_D to train the discriminator.

$$L_G = \alpha \cdot L_G^{\text{Huber}} + \beta \cdot L_G^{\text{NSGAN}} + \gamma \cdot L_G^{\text{style}} + \lambda \cdot L_G^{\text{KLD}} \tag{1}$$

$$L_G^{\text{Huber}} = \mathbb{E}[\frac{1}{t}\sum_{i=1}^{t} \text{HuberLoss}(d_i, \hat{d}_i)] \tag{2}$$

$$L_G^{\text{NSGAN}} = -\mathbb{E}[\log(D(\hat{d}))] \tag{3}$$

$$L_G^{\text{style}} = -\mathbb{E}\left[\min\left(\frac{\begin{array}{r}\text{HuberLoss}(G(f^{\text{text}}, f^{\text{audio}}, f^{\text{style}_1}) \\ - G(f^{\text{text}}, f^{\text{audio}}, f^{\text{style}_2}))\end{array}}{\|f^{\text{style}_1} - f^{\text{style}_2}\|_1}, \tau\right)\right] \tag{4}$$

$$L_D = -\mathbb{E}[\log(D(d))] - \mathbb{E}[\log(1 - D(\hat{d}))] \tag{5}$$

where t is the length of the gesture sequence, d_i represents the ith pose, represented as directional vectors, in a training sample. When training the encoder and gesture generator, we minimized the difference between human poses d in the training examples and the corresponding generated poses $\hat{d}$ using the Huber loss (Huber 1964). This loss L_G^{Huber} can be interpreted as a once-differentiable combination of the L1 and L2 losses, and is therefore sometimes called the smooth L1 loss. The adversarial losses L_G^{NSGAN} and L_D are from the non-saturating generative adversarial network (NS-GAN) (Goodfellow et al. 2014). We use sample mean to approximate the expectation terms.

A generative model conditioned on multiple input contexts often suffers from posterior collapse where weak context is ignored. In the proposed model, various gestures can be generated only from text and audio, so the style features from speaker IDs might be ignored during training. Thus, we use diversity regularization (Yang et al. 2019) to avoid ignoring style features. L_G^{style} is the Huber loss between the gestures generated from different style features normalized by the differences of the two style features, so it guides style features in the embedding space to generate different gestures. τ is for value clamping for numerical stability. In Equation 4, f^{style_1} is the style feature corresponding to the speaker ID of a training sample, and f^{style_2} is the style feature for a speaker ID selected randomly. L_G^{KLD}, the Kullback–Leibler (KL) divergence between $\mathcal{N}(0, I)$ and the style embedding space assumed Gaussian, prevents the style embedding space from being too sparse (Kingma and Welling 2014).

L_D is to train the discriminator D, and the generator and discriminator are alternately updated with L_G and L_D as in conventional GAN training (Goodfellow et al. 2014). $D(\cdot)$ is trained to output 1 for human gestures and 0 for generated gestures.

The model was trained for 100 epochs. An Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ was used, and the learning rate was 0.0005. Weights for the loss terms were determined experimentally ($\alpha = 500$, $\beta = 5$, $\gamma = 0.05$, and $\lambda = 0.1$). In addition, there was a warm-up period of 10 epochs in which the adversarial loss was not used ($\beta = 0$). τ was 1000.

The trained encoders and generator are used at the synthesis stage. As the model is lightweight enough, the synthesis can be done in real time. A single synthesis generating 30 poses takes 10 ms on a GPU (NVIDIA RTX 2080 Ti) and 80 ms on a CPU (Intel i7-5930K).

5 OBJECTIVE EVALUATION METRIC

It is difficult to evaluate gesture generation models objectively because no perceptual quality metric is available for human gestures. Although a human evaluation method in which participants rate generated gestures subjectively is possible, objective evaluation metrics are still required for fair and reproducible comparisons between state-of-the-art models. No proper and widely used evaluation metric is yet available for the gesture generation problem.

Image generation studies have proposed the FID metric (Heusel et al. 2017). Latent image features are extracted from the generated images using a pretrained feature extractor and FID calculates the Fréchet distance between the distributions of the features of real and generated images. Because FID uses feature vectors that describe visual characteristics well, FID is more perceptually appropriate than measurements over raw pixel spaces. FID can also measure the diversity of the generated samples by using the samples' distribution rather than simply averaging the differences between the real and generated samples. The diversity of generation has been thought to be one of major factors in evaluating generative models (Borji 2019). Diversity is also crucial for the gesture generation problem because the use of repetitive gestures makes artificial agents look dull.

5.1 Fréchet Gesture Distance

In applying the concept of FID to the gesture generation problem, there is a hurdle that no general feature extractor is available for gesture data. The paper proposing FID used an inception network trained on the ImageNet database for image classification, but there is no analog of the pretrained inception network for gesture motion data to the best of our knowledge. Accordingly, we trained a feature extractor based on autoencoding (Rumelhart et al. 1985), which can be trained in unsupervised manner. The feature extractor consists of a convolutional encoder and decoder; the encoder encodes a sequence of direction vectors d to a latent feature $z^{gesture}$ and the decoder then attempts to restore the original pose sequence from the latent $z^{gesture}$ (see appendix A for the detailed architecture). This unsupervised learning is unlike the supervised learning of the inception network used in FID. However, both supervised and unsupervised learning have proven to be effective for learning perceptual quality metrics (Zhang et al. 2018).

The encoder part of the trained autoencoder was used as a feature extractor. We defined $\mathrm{FGD}(X, \hat{X})$ as the Fréchet distance between the Gaussian mean and covariance of the latent features of human gestures X and the Gaussian mean and covariance of the latent features of the generated gestures $\hat{X}$ as follows:

$$\mathrm{FGD}(X, \hat{X}) = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}) \quad (6)$$

where μ_r and Σ_r are the first and second moments of the latent feature distribution Z_r of real human gestures X, and μ_g and Σ_g are the first and second moment of the latent feature distribution Z_g of generated gestures $\hat{X}$.

For training the feature extractor, we used the Human3.6M dataset (Ionescu et al. 2013) containing motion capture data of 7 different actors and 17 different scenarios including discussion and making purchases showing co-speech gestures. The total duration of the training data was about 175 m. All poses were frontalized based on two hip joints.

5.2 Experiment with Synthetic Noisy Data

We explored the properties of the proposed FGD metric using synthetic noisy data. Five types of noisy data were considered. Gaussian noise and Salt&Pepper (S&P) noise were added to the joint coordinates of poses; the same noise data were added to all poses in a sequence, so that there is no artificial temporal discontinuity. Temporal noise was simulated by adding Gaussian noise to only a few time frames. Multiplicative transformation in "eigenposes" p_i^{eigen} (Yoon et al. 2019) converted from p_i using principal component analysis (PCA) was used to generate monotonous or exaggerated gestures. Mismatched gestures were also generated to examine how the metric responds to discrepancies between speech and gestures. The following shows how the noisy data were synthesized. The parameter ζ controls the overall disturbance levels. The dimension of a pose, K, is 30 (10 joints in 3D coordinates).

- Gaussian noise: $\tilde{p}_i = p_i + x; x \sim \mathcal{N}_K(0, \zeta I)$

Fig. 3. Samples of noisy gesture data to validate evaluation metrics. (a) None (original data), (b) Gaussian noise ($\zeta = 0.001$), (c) Salt&Pepper noise ($\zeta = 0.1$), (d) Temporal noise ($\zeta = 5$), (e–f) Multiplicative transformation in eigenposes ($\zeta = 0.0, 2.0$), (g) Mismatched sample

- Salt&Pepper noise: $\tilde{p}_i = p_i + x$

$$x_{k=1,\ldots,K} = \begin{cases} 0.2 & \text{if } u \le \zeta/2; u \sim U(0,1) \\ -0.2 & \text{if } \zeta/2 < u \le \zeta \\ 0 & \text{otherwise} \end{cases}$$

- Temporal noise:

$$\tilde{p}_i = \begin{cases} p_i + x; x \sim \mathcal{N}_K(0,\ 0.003I) & \text{if } r \le i < r + \zeta; \\ & r \text{ is a random time step} \\ p_i & \text{otherwise} \end{cases}$$

- Multiplicative transformation: $\tilde{p}_i^{eigen} = \zeta \cdot p_i^{eigen}$
- Mismatched samples: Select a fraction ζ of all samples and associate the input speech to random gestures in the TED test dataset to make mismatched samples.

Figure 3 shows samples of the synthetically noisy data. The Gaussian noise introduced changes across all joints, whereas the S&P noise produces impulsive noise in a few joints. The temporal noise introduced discontinuities in motion. Multiplicative transformation was applied to eigenposes, so it controls the overall motion range. The mismatch noise shows a sample of nonmatching content and speech rhythms.

We measured FGD and mean absolute error of joint coordinates (MAEJ) which is calculated as $\mathrm{MAE}(\tilde{p}, p)$. Figure 4 shows the experimental results. For the Gaussian and S&P noise, both FGD and MAEJ showed increasing distances as the disturbance level increases, but FGD showed larger distances for S&P noise than Gaussian noise on average, unlike MAEJ. As shown in Figure 3 (b) and (c), the sample with Gaussian noise still look like human poses, though with some distortions, whereas the sample with S&P noise show unrealistic poses where the neck is out of the upper body. The samples with Gaussian noise are more perceptually plausible gestures than those

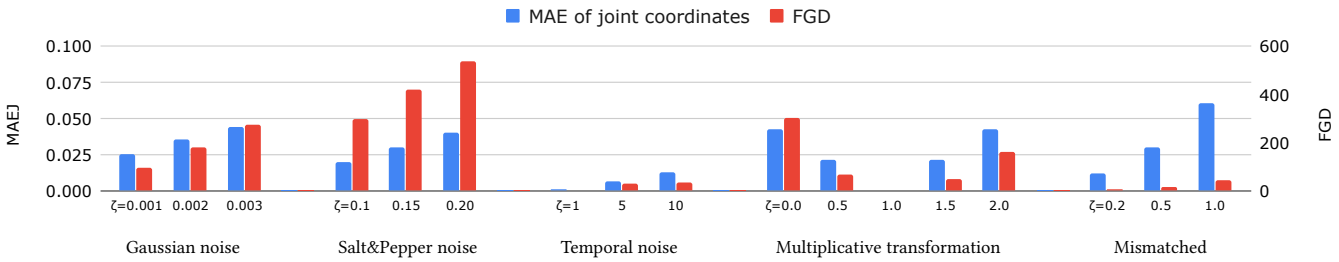

Fig. 4. Results of the metric validation experiment on the synthetic noisy dataset showing four types of noise. The disturbance level increases as ζ increases except for the multiplicative transformation. The disturbance level is lowest when $\zeta = 1.0$ for the multiplicative transformation.

with S&P noise, so, in our view, having larger distances for S&P noise is acceptable.

Both FGD and MAEJ showed relatively low values for the temporal noise even though discontinuous motion is perceptually unnatural. MAEJ calculates errors in each time frame independently, so it is obvious that MAEJ is not able to capture motion discontinuity. However, FGD, which encodes a whole sequence, also showed low distances unexpectedly. The primary reason is that the feature extractor used in FGD were not able to discriminate enough between the sequences with and without temporal noise. When we examined the reconstructed motion from the autoencoder, we found that the autoencoder tended to remove temporal noise.

For the multiplicative transformation, both metrics showed increasing distances as the disturbance level increased (larger or smaller than $\zeta = 1.0$). MAEJ showed similar distances for $\zeta = 0.0$ and 2.0, but FGD showed a much larger distance when $\zeta = 0.0$. As shown in Figure 3 (e) and (f), $\zeta = 0.0$ and 2.0 make mean and exaggerated poses. If we consider several results and their diversity, exaggerated poses are perceptually favored over having the same mean poses regardless of input speech. Thus, it is reasonable to have larger distances for $\zeta = 0.0$ for than 2.0, as FGD does.

Lastly, for the mismatched samples, both MAEJ and FGD showed increasing distances for more mismatched samples, but the increase in FGD was smaller than MAEJ. This result is not surprising since FGD considers a distribution formed by a set of gestures and is not aware of the input speech.

In this experiment, we found that FGD gives perceptually plausible results for the different gesture data of Gaussian noise, S&P noise, and multiplicative transformation and has the limitation that it is not able to measure well enough temporal noise and match of speech and gestures. We found the characteristics of FGD; however, it is difficult to argue that the metric is suitable for use based on an experiment with synthetic data wherein only one human gesture example is used for each speech even though many-to-many mappings exist between speech and gestures. To further investigate the effectiveness of FGD and MAEJ, we compare these metrics to human judgements in the following section.

6 USER STUDY TO VALIDATE EVALUATION METRICS

In this section, we validated FGD by comparing with subjective ratings from humans. We followed the overall experimental setting in the paper introducing Fréchet video distance (Unterthiner et al. 2019), but we had two separate user study sessions with 14 noise models and 10 trained gesture generation models. In the first session, 14 noise models (excluding Mismatched with $\zeta = 0.2$ and 0.5) were used. In the second session, 10 gesture generation models showing different FGD were selected among models trained in the course of this study. We tried to select models equidistant in terms of FGD. The selected models are in different architectures, configurations, and training stages; see appendix B for the complete list of the selected models and their configurations. We also included the human gesture in both sessions.

We made videos showing a dummy character making gestures for each model. In the evaluation, pairwise preference comparisons were used instead of a Likert-scale rating. Co-speech gestures are subtle, so participants would have struggled to rate them on a five- or seven-point scale. Using pairwise preference comparisons reduces participants' cognitive load and yields reliable results, as discussed in (Clark et al. 2018). The participants watched two videos of two different models and responded to one of three questions asking about their preference, human-likeness of motion, and speech–gesture matching: 1) "Which gesture motion do you prefer?," 2) "Which gesture motion is more natural and human-like?," and 3) "Which gesture motion is more appropriate with the speech audio and words?" The answer options were "Video A," "Video B," and "Undecidable." For the question on human-likeness of motion, the videos were played without speech audio to make the participants assess only the motion. Each participant was asked to answer one randomly selected question in all of his/her trials, since the three questions are substantially correlated and the participants are prone to give the same answer if we ask three questions at the same time.

For the evaluation, speech samples with lengths of 5–10 s were drawn randomly from the TED test dataset. We only reviewed the quality of the extracted 3D human poses of the samples to exclude faulty samples that may mislead the performance of human gestures (the top line). Thirty speech samples were used in the evaluation after excluding four faulty samples where the speaker is manipulating an object, sitting on a chair, and occluded by a podium. Two models were randomly selected for each pairwise comparison to eliminate ordering effects.

Native or bilingual English speakers were recruited from Amazon MTurk. Each participant responded to 30 pairwise comparisons which were chosen randomly among all possible pairwise combinations (30 sentences $\times$ $\{{}_{15}C_2$ or ${}_{11}C_2\} \times 3$ questions). They took

Table 1. Agreement of the evaluation metric to human judgements in the user study on (a) noise models and (b) trained gesture generation models. We also report the agreements between human subjects as a top line. Higher numbers are better.

Metric	Agreement (%)		
	Preference	Human-likeness of motion	Speech–gesture match
(a) Noise models			
MAE of joint coordinates (MAEJ)	50.9	55.9	60.5
MAE of acceleration	46.5	47.7	46.3
FGD	64.8	63.6	66.3
Between human subjects	83.3	72.2	85.7
(b) Trained gesture generation models			
MAE of joint coordinates (MAEJ)	37.7	48.2	32.8
MAE of acceleration	34.9	40.0	38.9
FGD	70.5	59.6	70.2
Between human subjects	73.1	78.8	94.4

15–30 min to complete the task, and 2.5 USD was given as a reward. We also included an attention check presenting two copies of the same video side by side. The participants who did not answer "undecidable" in this case were excluded. In the first session with the noise models, a total of 28 subjects participated, but we analyzed the results from 22 subjects after excluding six subjects failed the attention check. There were 13 male and 9 female subjects, and they were 36.9 ± 11.5 years old. In the second session with the trained generation models, a total of 51 subjects participated and 21 subjects were excluded. There were 15 male and 15 female subjects, and they were 42.8 ± 13.2 years old. The total numbers of answers were 660 and 900 at the first and second session, respectively.

We evaluated the objective evaluation metrics by comparing those with human judgements, and the results are shown in Table 1. MAE of acceleration was used to assess dance motion (Aristidou et al. 2015) and gestures (Kucherenko et al. 2019), and it focuses on motion rather poses. The agreement values were calculated as the number of comparisons in which each metric agreed human judgement divided by the total number of comparisons. "Undecidable" responses were not included in the analysis. In both sessions, FGD showed greater agreement with human judgements than did MAE of joint coordinates and MAE of acceleration on all questions. However, FGD was performed less than the agreements between humans; in particular, FGD showed the lowest agreement of 53.5% for temporal noise as discussed in Section 5.2.

By considering both experimental results on synthetic noisy data and human judgements, FGD is a plausible objective metric. In addition, when we examine learning curves, which are shown in Figure 5, FGD showed a decreasing trend when the distribution of generated gestures becomes more similar to the reference distribution as training continues. In contrast, MAEJ showed a flat learning curve. The lowest MAEJ is at Epoch 6, in which only static mean poses appear for all speech contexts. In the following experiments, we use FGD to compare models.

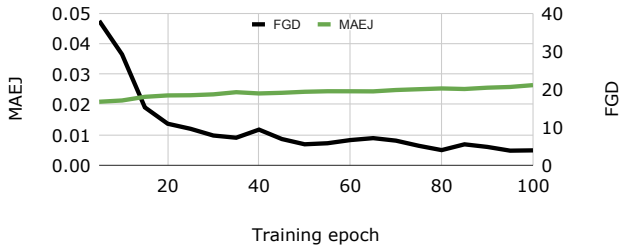

Fig. 5. Validation learning curves measured by mean absolute error of joint coordinates (MAEJ) and Fréchet gesture distance (FGD).

All subjects were asked to write the reasons for their selection. Most of them said they preferred gestures that were fit to speech words and audio, as we had assumed in the present paper. Opinions on gesture dynamics were mixed. Some participants liked dynamic or even exaggerated gestures, whereas some other participants preferred moderate gestures with a few large movements for emphasis. This implies that the gesture styles must be adapted as per the users' preference.

7 EXPERIMENTS AND HUMAN EVALUATION

7.1 Qualitative Results

Figure 6 shows the gesture generation results for the speech in the test set of the TED gesture dataset. The gestures are depicted using a 3D dummy character. The poses represented as directional vectors were retargetted to the character with fixed bone lengths, and the gesture sequences were upsampled using cubic spline interpolation to 30 FPS. We used the same retargeting procedure for all animations. The character makes metaphoric gestures when saying "civil rights," "30 million," or "great leadership." An iconic gesture also found for the words "to the point." Gesture generation depends on speech rhythm and presence or absence of speech as shown in the sample (a) and (e). A deictic gesture also appears in (c) when the character says "I." Please see the supplementary video for the animated results.

7.2 Comparisons with state-of-the-art models

We compared the proposed model with three models from previous studies. The first model compared is attentional Seq2Seq, which generates gestures from speech text (Yoon et al. 2019). We followed the original implementation provided by the authors but the gesture representation was modified to be identical to the proposed model. The second comparison model is Speech2Gesture (Ginosar et al. 2019), which generates gestures from speech audio using an encoder–decoder neural architecture and learns to generate human-like gestures by using an adversarial loss during training. Spectrograms were used to represent audio in this model. The third one is the joint embedding model (Ahuja and Morency 2019), which

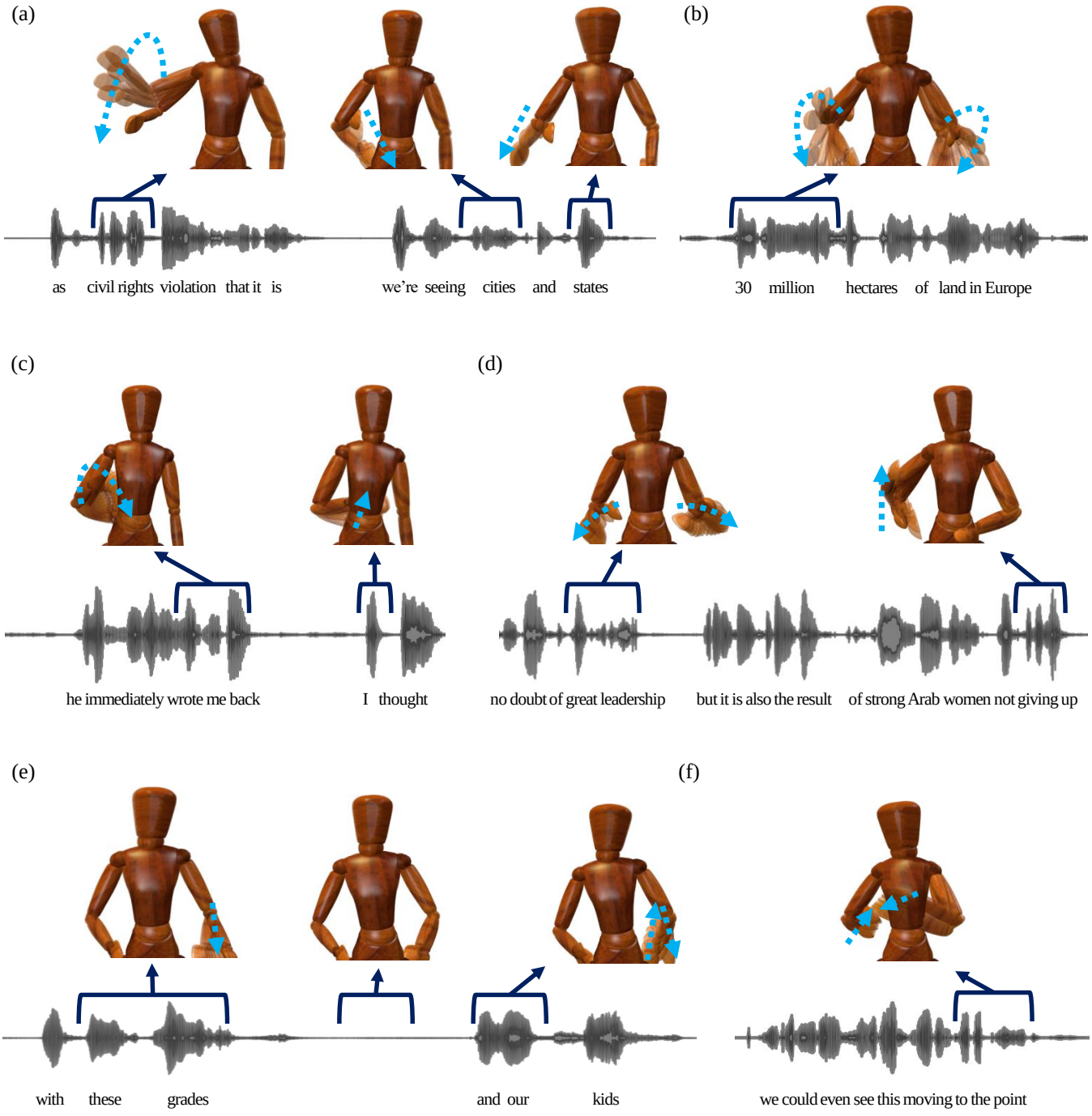

Fig. 6. Sample results of co-speech gesture generation from the trimodal speech context of text, audio, and speaker identity. Motion history images for some parts are depicted along with the speech text and audio signals. In (a), the character makes metaphoric gestures when saying "civil rights" and beat gestures for "cities and states." In (b) and (d), there are metaphoric gestures for the words of "30 million," "great leadership," and "giving up." In (c), a deictic gesture appears when the character says "I." In (e), we can find the character does not gesture in the middle of the silence. An iconic gesture is also found in (f).

Table 2. Results of comparisons with state-of-the-art models. Lower numbers indicate better performance (Bold: **best**, Underline: second).

Method	FGD
Attentional Seq2Seq (Yoon et al. 2019)	18.154
Speech2Gesture (Ginosar et al. 2019)	19.254
Joint embedding model (Ahuja and Morency 2019)	22.083
Proposed	**3.729**

creates human motion from motion description text. This model maps text and motion to the same embedding space. We embedded the input speech text and audio together to the same space as the motion. The same encoders in our model were used to process the audio and text, and 4-layered GRUs were used for gesture generation. All models were trained on the same TED dataset for the same number of epochs. We modified the original architectures of the baselines to generate the same number of poses (i.e., 30) and to use four seed poses for consecutive syntheses. The learning rate and weights of loss terms in the baselines were optimized via grid search for best FGD.

Figure 7 shows sample results from each model for the same speech. The joint embedding model generated very static poses, failing to learn gesticulation skills. The relationship between speech and gestures are weak and subtle, making it difficult to map speech and gestures to a joint embedding space. All other models generated plausible motions, but there were differences depending on the modality and training loss considered. Attentional Seq2Seq generated different gestures for different input speech sentences, but the motion tended to be slow and we found a few discontinuities between the seed poses and generates poses. The Speech2Gesture model used an RNN decoder similar to attentional Seq2Seq, but it showed better motion with the help of its adversarial loss component. However, because it uses only a single speech modality, audio, Speech2Gesture generated monotonous beat gestures. The proposed model successfully generated large and dynamic gestures as shown in the supplementary video.

The proposed model performed the best in terms of FGD (Table 2). We also analysed the human evaluation results by computing ranks from pairwise comparisons using the Bradley—Terry model (Chu and Ghahramani 2005). Pairwise comparisons were collected from another 14 MTurk subjects that passed the same attention check as before. The same settings described in Section 6 were used, but only the four models in Table 2 and human gestures were compared. Figure 8 shows the results. For all the questions, the proposed method achieved better results than Attentional Seq2Seq, Speech2Gesture, and joint embedding methods, but the differences between the proposed method and Speech2Gesture were not distinct in the the human-likeness of motion and speech–gesture match questions. We also tested statistical significance between the proposed method and the others by using the Chi-Square Goodness of Fit test over the null hypothesis that the probabilities of the pairwise choices are equal to 50% (the choice of "undecidable" was not counted). In the preference, the difference between the proposed and joint embedding method was significant (p ¡ 0.01). In the speech–gesture match, Seq2Seq

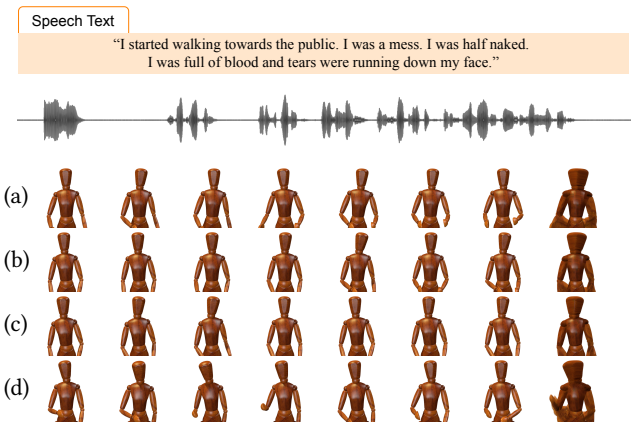

Fig. 7. Sample results of (a) attentional Seq2Seq, (b) Speech2Gesture, (c) joint embedding, and (d) the proposed model for the same input speech. Seven evenly sampled frames are shown for the resulting pose sequences. The last column shows motion history images in which all frames are superimposed. Please see the supplementary video for animated results.

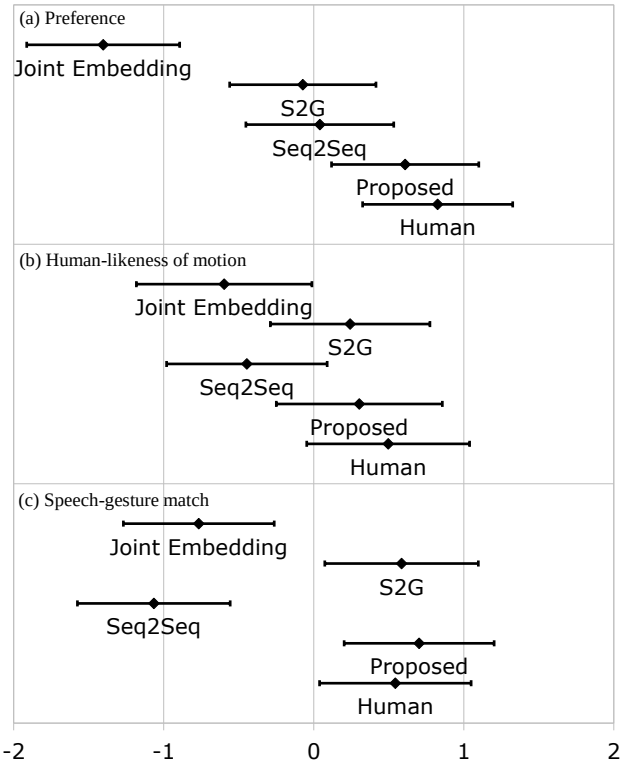

Fig. 8. The results of human evaluation for the three questions about (a) preference, (b) human-likeness of motion, and (c) speech–gesture match. The ranking is calculated using the Bradley-Terry model and the horizontal axis represents the winning probability against the other methods. Mean and standard deviation are depicted through Bayesian inference for the Bradley-Terry model (Chu and Ghahramani 2005). S2G denotes the Speech2Gesture method.

Table 3. Results of the ablation study for the proposed model. Lower numbers are better. Ablations are not accumulated.

Configuration	FGD
Proposed (no ablation)	3.729
Without speech text modality	4.701
Without speech audio modality	4.874
Without speaker ID	6.275
Without adversarial scheme	9.712
Without regularization terms L_G^{style} and L_G^{KLD}	5.756

and joint embedding methods were significantly different from the proposed method (p ¡ 0.01 and ¡ 0.05, respectively).

The proposed method showed better results than the previous methods objectively and subjectively. Also, the proposed method is mostly tied with human gestures in the user study. This indicates the superiority of the proposed method, but we cannot conclude that the proposed method performed equally well as humans since the human gestures used in the experiments were based on automatically extracted poses from TED videos and all motion was retargetted to a restricted character without face or hand expressions.

7.3 Ablation Study

An ablation study was conducted to understand the proposed model in detail. We eliminated components from the proposed model that was used in the comparison with the state-of-the-art models. Table 3 summarizes the results of the ablation study. Removing each modality of text, audio, and speaker ID reduced the model's performance; this shows that all three modalities used in the proposed model had positive effects on gesture generation. Among the loss terms, removing the adversarial term and regularization terms also worsened FGD. In particular, when we trained the model without the adversarial scheme, the model tended to generate static poses close to the mean pose.

Although, when ablating different modalities, excluding the speaker ID degraded the FGD the most, we could not find a noticeable degradation in our subjective impression of motion quality than ablating text or audio modalities. In our view, this is attributed to that overall diversity was reduced without the divergence regularization L_G^{style} and that the property of FGD that measures not only motion quality but also diversity. There is no concrete way to disentangle the factors of quality and diversity in FGD as well as FID. However, we hypothesise that the covariance matrix of the fitted Gaussian is more related to the diversity than to quality. The trace of the covariance matrix was 244, which is less than that of the human gestures and of the models without the text or audio modalities (299, 258, and 250, respectively). This indirectly suggests that generated gestures were less diverse without speaker IDs and L_G^{style}.

The text modality had the least effect on FGD. In the proposed model, speech text and audio are treated as independent modalities; however, strictly speaking, audio contains text information because we can transcribe text from audio. Although the above ablation study showed that the FGD worsened without the text modality, it was less significant than excluding audio or speaker IDs. We further verified the effect of the text with an additional experiment. Figure 9 shows how the generated gestures differed when a word was altered in the input speech with the model considering both text and audio and the model considering only audio. Although the model considering both text and audio generated different gestures (widening arms) when the word "hundreds" replaced the words "few," there was only a slight change in motion when we used the audio-only model. We synthesized speech audio using Google Cloud TTS (Google 2018) for both original and altered text.

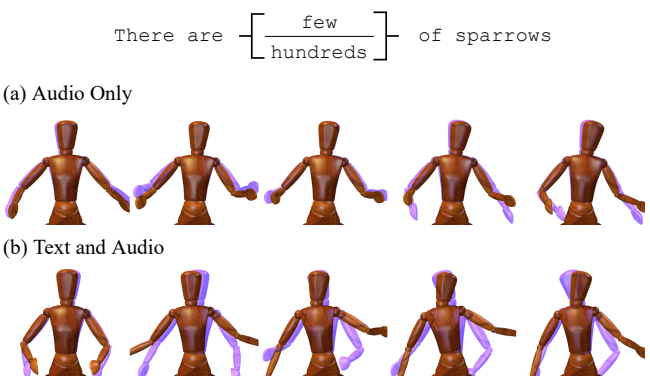

Fig. 9. Visualization of how a gesture changes when a word is changed in a sentence. We compare the results of (a) the ablated model without text modality and (b) proposed model considering both text and audio. The generated gesture for the original and altered sentences are overlaid for five evenly sampled frames. When we consider both text and audio, the model generates more different gestures for the changes in speech content.

We also conducted the above text-altering experiment quantitatively. For 1,000 samples randomly selected from the validation set, a word in a speech sentence was changed to a synonym or antonym taken from WordNet (Miller 1995). If there were several synonyms or antonyms, the one closest in duration to the original word was selected to minimize the change in the length of the speech audio. Synthesized audio was used and the experiment was repeated 10 times due to the randomness in selecting samples and words. We report the FGD between the generated samples before and after text alteration; this measure is unlike all other FGD measures, which compare human motion and generated motion, in the paper. The model considering text and audio (2.433 ± 0.483) showed a significantly higher FGD than the model considering only audio (1.604 ± 0.275) (paired t-test, p ¡ 0.001), indicating that using text and audio modalities together helps to generate diverse gestures according to the changes in the speech text. This argument is also backed by the result that the FGD when a word was replaced by an antonym (2.567 ± 0.484) was significantly higher than when replaced by a synonym (2.299 ± 0.467) (paired t-test, p ¡ 0.05) in the model using both text and audio.

7.4 Incorporating Synthesized Audio

Many artificial agents use synthesized audio since recording a human speaking for every word is infeasible. We tested that the proposed model, trained with human speech audio, also can work with synthesized audio. Figure 10 and the supplementary video shows

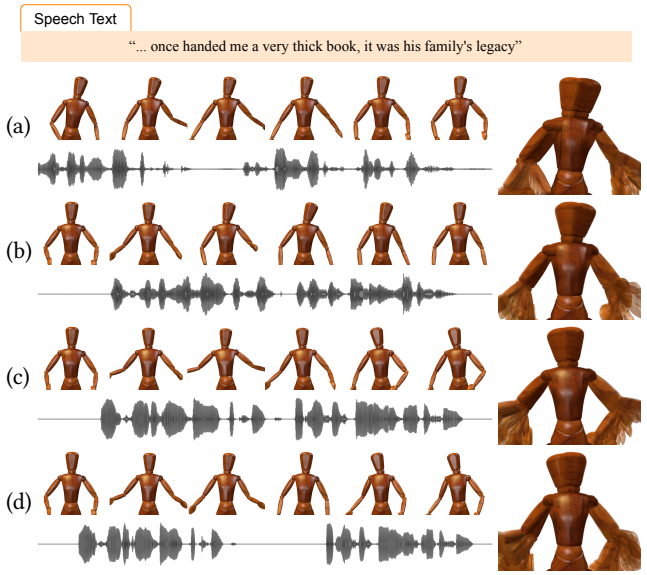

Fig. 10. Co-speech gesture generation results with (a) original human speech audio, (b) synthesized audio of a male voice, (c) synthesized audio of female voice, and (d) synthesized audio of a female voice with pauses. The proposed model can generate gestures from synthesized audio of different voices and rhythm.

some results using synthesized audio with different voices. Google Cloud TTS (Google 2018) was used in this experiment. The proposed model worked well with synthesized audio of different voices, prosody, speed, and pauses. When the speech is fast, the model generates rapid motion. The model also reacts to inserted speech pauses by generating static poses for the silence period.

7.5 Analysis of the Learned Style Embedding Space

The proposed model can generate different gesture styles for the same speech. Figure 11 visualizes the trained style embedding space and the gestures generated with different style vectors for the same input speech. To understand the style embedding space closely, we depict the motion statistics of the generated gestures for each style vector corresponding to speaker ID with the marker color and shape in the figure. Colors from red to blue correspond to higher and lower temporal motion variances. A larger motion variance can be called an extrovert style and the opposite is an introvert style. We also calculated the temporal motion variance for the right and left arms separately and used different marker shapes to indicate styles of handedness. Styles using the right and left arms more are depicted as ► and ◄ respectively, and the rest are depicted as ●. As shown in Figure 11, similar styles are clustered, and users can easily choose the desired style from the embedding space after traversing it.

8 CONCLUSIONS AND LIMITATIONS

In this paper, we presented a co-speech gesture generation model that generates upper-body gestures from input speech. We proposed a temporally synchronized architecture using the three input modalities of speech text, audio, and speaker ID. The trained model successfully generated various gestures matching the speech text and audio; different styles of gestures could be generated by sampling style vectors from a style embedding space. A new metric, FGD, was introduced to evaluate the generation results. The proposed metric was validated using synthetic noisy data and measuring the agreement with human judgements. The proposed generation method showed better results than previous methods both objectively and subjectively as determined by the FGD metric and human evaluation. We also highlighted different properties of the proposed model through various experiments. The model can generate gestures with synthesized audio of various prosody settings. Additionally, the style embedding space was trained to be a continuous space where similar styles are distributed closely.

There is room for improvement in the present research. First, it is difficult to control the gesture generation process. Although style manipulation is possible, users are not able to set constraints on gestures. For example, we might want an avatar to make a deictic gesture when the avatar says a specific word. Most end-to-end neural models have this controllability issue (Jahanian et al. 2020). It would be interesting to extend the current model to have further controllability, for example, by adding constraining poses in the middle of generation. Second, FGD need to be improved. In nonverbal behavior, subtle motion is as important as large motion, but the feature extractor trained by motion reconstruction might fail to capture subtle motion. It is also necessary to separately evaluate motion quality and diversity for in-depth comparisons between generation models. Third, we only considered the motion of upper body, whereas whole-body motion, including facial expressions and finger movements should be integrated. Taking a long-term view of creating an artificial conversational agent, we would pursue integrating our model with other nonverbal behaviors and with a conversational model. Gestures are deeply related to verbalization according to the information packaging hypothesis (Kita 2000), so an integrated model generating speech and gestures together could deliver information more efficiently.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their thorough and valuable comments. This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00162, Development of Human-care Robot Technology for Aging Society). Resource supporting this work were provided by the 'Ministry of Science and ICT' and NIPA ("HPC Support" Project).

REFERENCES

Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2Pose: Natural Language Grounded Pose Forecasting. In *International Conference on 3D Vision*. IEEE, 719–728.

Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 487–496.

Andreas Aristidou, Efstathios Stavrakis, Panayiotis Charalambous, Yiorgos Chrysanthou, and Stephania Loizidou Himona. 2015. Folk Dance Evaluation Using Laban Movement Analysis. *Journal on Computing and Cultural Heritage* 8, 4 (2015), 20.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representations*.

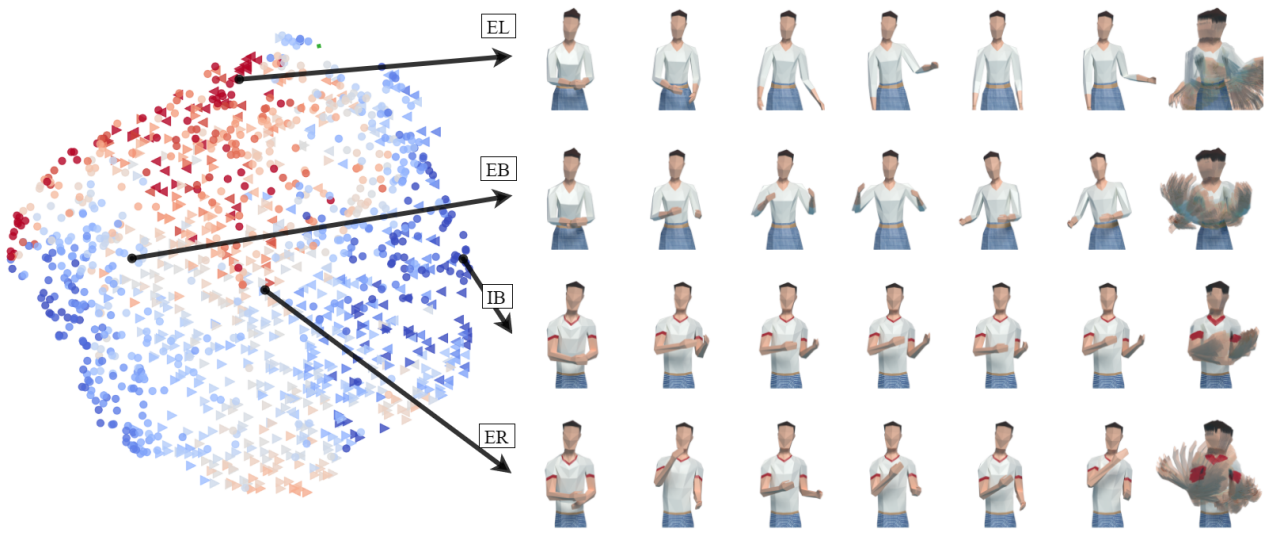

Fig. 11. Visualization of the style embedding space and sample generation results for the different style vectors. All speaker identities are mapped to style feature vectors f^{style}, and we visualize the feature vectors in two dimensions by using UMAP (McInnes et al. 2018). The points represent degrees of motion variance via color and degree of handedness by its marker types. We labeled the sampled style vectors according to the overall motion variance and handedness as 'IB' for the introvert style of moving both hands similarly, 'ER' for the extrovert style of moving the right hand more, and so on. All gesture results are generated from the same speech.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv:1803.01271* (2018).

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

Kirsten Bergmann, Volkan Aksu, and Stefan Kopp. 2011. The Relation of Speech and Gestures: Temporal Synchrony Follows Semantic Synchrony. In *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

Ali Borji. 2019. Pros and Cons of GAN Evaluation Measures. *Computer Vision and Image Understanding* 179 (2019), 41–65.

Paul Bremner, Anthony G Pipe, Chris Melhuish, Mike Fraser, and Sriram Subramanian. 2011. The Effects of Robot-Performed Co-Verbal Gesture on Listener Behaviour. In *IEEE-RAS International Conference on Humanoid Robots*. IEEE, 458–465.

Judee K Burgoon, Thomas Birk, and Michael Pfau. 1990. Nonverbal Behaviors, Persuasion, and Credibility. *Human communication research* 17, 1 (1990), 140–169.

Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. BEAT: the Behavior Expression Animation Toolkit. In *Life-Like Characters*. Springer, 163–185.

Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting Co-verbal Gestures: A Deep and Temporal Modeling Approach. In *ACM International Conference on Intelligent Virtual Agents*. Springer, 152–166.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Empirical Methods in Natural Language Processing*. 1724–1734.

Mingyuan Chu and Peter Hagoort. 2014. Synchronization of Speech and Gesture: Evidence for Interaction in Action. *Journal of Experimental Psychology: General* 143, 4 (2014), 1726.

Wei Chu and Zoubin Ghahramani. 2005. Extensions of Gaussian Processes for Ranking: Semi-supervised and Active Learning. *Learning to Rank* (2005), 29.

Andrew P Clark, Kate L Howard, Andy T Woods, Ian S Penton-Voak, and Christof Neumann. 2018. Why Rate When You Could Compare? Using the "EloChoice" Package to Assess Pairwise Comparisons of Perceived Physical Strength. *PloS one* 13, 1 (2018).

Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2019. Multi-Objective Adversarial Gesture Generation. In *Motion, Interaction and Games*. 1–10.

Peng Fu, Zheng Lin, Fengcheng Yuan, Weiping Wang, and Dan Meng. 2018. Learning Sentiment-Specific Word Embedding via Global Sentiment Representation. In *AAAI Conference on Artificial Intelligence*.

Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning Individual Styles of Conversational Gesture. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3497–3506.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

Google. 2018. Google Cloud Text-to-Speech. https://cloud.google.com/text-to-speech Accessed: 2020-03-01.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*. 6626–6637.

Autumn B Hostetter and Andrea L Potthoff. 2012. Effects of Personality and Social Situation on Representational Gesture Production. *Gesture* 12, 1 (2012), 62–83.

Chien-Ming Huang and Bilge Mutlu. 2014. Learning-Based Modeling of Multimodal Behaviors for Humanlike Robots. In *ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 57–64.

Peter J Huber. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* 35, 1 (1964), 73–101.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2013), 1325–1339.

Ali Jahanian, Lucy Chai, and Phillip Isola. 2020. On the "Steerability" of Generative Adversarial Networks. In *International Conference on Learning Representations*.

Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. 2019. Towards Social Artificial Intelligence: Nonverbal Social Signal Prediction in A Triadic Interaction. In *IEEE Conference on Computer Vision and Pattern Recognition*. 10873–10883.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2018. Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. *arXiv preprint arXiv:1812.08466* (2018).

Taewoo Kim and Joo-Haeng Lee. 2020. C-3PO: Cyclic-Three-Phase Optimization for Human-Robot Motion Retargeting based on Reinforcement Learning. In *International Conference on Robotics and Automation*.

Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*.

Michael Kipp. 2005. *Gesture Generation by Imitation: From Human Behavior to Computer Character Animation*. Universal-Publishers.

Sotaro Kita. 2000. How Representational Gestures Help Speaking. *Language and gesture* 1 (2000), 162–185.

Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a

Common Framework for Multimodal Generation: The Behavior Markup Language. In *ACM International Conference on Intelligent Virtual Agents*. Springer, 205–217.
Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing Input and Output Representations for Speech-Driven Gesture Generation. In *ACM International Conference on Intelligent Virtual Agents*. 97–104.
Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A Framework for Semantically-Aware Speech-Driven Gesture Generation. In *ACM International Conference on Multimodal Interaction*.
Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture Controllers. *ACM Transactions on Graphics* 29, 4 (2010), 1–11.
Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual Character Performance From Speech. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 25–35.
Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3 (2018).
David McNeill. 1992. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago press.
David McNeill. 2008. *Gesture and Thought*. University of Chicago press.
Alberto Menache. 2000. *Understanding Motion Capture for Computer Animation and Video Games*. Morgan Kaufmann.
Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*. 3111–3119.
George A Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41.
Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture Modeling and Animation Based on a Probabilistic Recreation of Speaker Style. *ACM Transactions on Graphics* 27, 1 (2008), 5.
Robert Ochshorn and Max Hawkins. 2016. Gentle: A Forced Aligner. https://lowerquality.com/gentle/ Accessed: 2020-01-06.
Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7753–7762.
Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing*. 1532–1543.
Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the International Conference on Machine Learning*, Vol. 32. 1278–1286.
Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Multimodal Continuous Turn-Taking Prediction Using Multiscale RNNs. In *ACM International Conference on Multimodal Interaction*. ACM, 186–190.
David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1985. *Learning Internal Representations by Error Propagation*. Technical Report. California Univ San Diego La Jolla Inst for Cognitive Science.
Najmeh Sadoughi and Carlos Busso. 2019. Speech-Driven Animation with Meaningful Behaviors. *Speech Communication* 110 (2019), 90–100.
Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*. 2234–2242.
Robotics Softbank. 2018. NAOqi API Documentation. http://doc.aldebaran.com/2-5/index_dev_guide.html Accessed: 2020-01-06.
Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. FVD: A new Metric for Video Generation. In *International Conference on Learning Representations Workshop*.
Petra Wagner, Zofia Malisz, and Stefan Kopp. 2014. Gesture and Speech in Interaction: An Overview. *Speech Communication* 57, Special Iss. (2014).
Jason R Wilson, Nah Young Lee, Annie Saechao, Sharon Hershenson, Matthias Scheutz, and Linda Tickle-Degnen. 2017. Hand Gestures and Verbal Acknowledgments Improve Human-Robot Rapport. In *International Conference on Social Robotics*. Springer, 334–344.
Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. 2019. Diversity-Sensitive Conditional Generative Adversarial Networks. In *International Conference on Learning Representations*.
Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots. In *International Conference on Robotics and Automation*. IEEE, 4303–4309.
Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition*. 586–595.

Table 4. The list of the gesture generation models used in the human evaluation. $*$ denotes the epoch having the best FGD.

Model	Training stage (epochs)	FGD
Proposed model	89*	3.729
Proposed model without regularization terms	83*	5.756
Proposed model without adversarial scheme	87*	9.712
Proposed model without text modality	20	12.144
Proposed model without audio modality	16	16.558
Attentional Seq2Seq	66*	18.054
Speech2Gesture	86*	19.254
Joint embedding model	98*	22.083
Proposed model without adversarial scheme and audio modality	17	26.328
Attentional Seq2Seq	20	28.273

A DETAILED ARCHITECTURES

Figure 12 shows the detailed architectures of the encoders, gesture generator, and discriminator. Figure 13 shows the architecture of the feature extractor used in the Fréchet gesture distance.

B MODELS IN HUMAN EVALUATION

Table 4 lists the models used in the human evaluation.

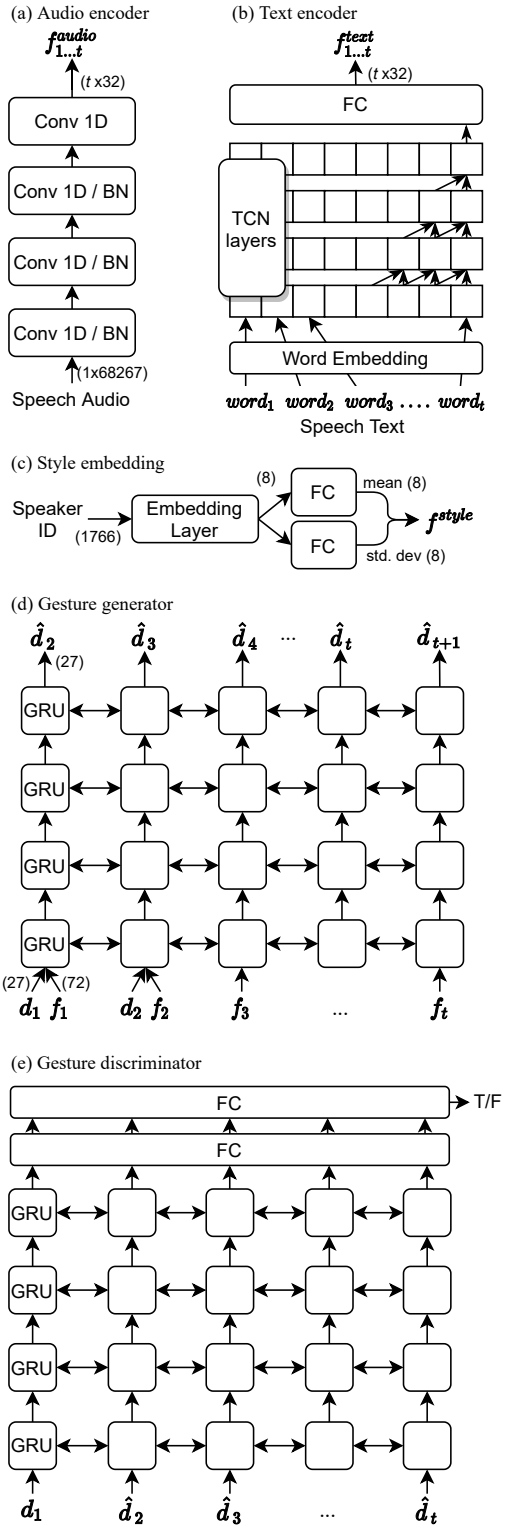

Fig. 12. Detailed architectures of the (a) audio encoder, (b) text encoder, (c) speaker embedding, (d) gesture generator (assumed two seed poses), and (e) discriminator. BN stands for batch normalization, FC for fully connected layer, and TCN for temporal convolutional network.

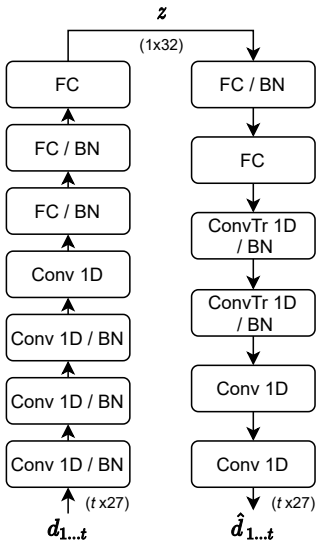

Fig. 13. Detailed architecture of the autoencoder of the Fréchet gesture distance. BN stands for batch normalization, FC for fully connected layer, and ConvTr for transposed convolution.