Agents that gesticulate: A framework for semantically-aware speech-driven gesture generation

Paper #1641

ABSTRACT

When speaking, people spontaneously gesticulate (i.e., produce co-speech gestures) which plays a key role in conveying information. Similarly, realistic co-speech gestures are crucial to enable natural and smooth human interactions with social agents. Current data-driven 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 simultaneously. We present a model designed to produce arbitrary, as opposed to pre-animated, 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. We illustrate the model's efficacy with subjective and objective evaluations.

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 [23], with the purpose to reference or represent other movements, objects, or abstract ideas [32].

The debate on the origin of gestures in humans is ongoing. While a group of work supports the claim that co-speech gestures are solely generated from the speech production process (*speech production hypothesis*) [28], another group of work supports the claim that co-speech gestures stem from semantic features (*lexical retrieval hypothesis*) [8]. One consensus is however, that the generation of co-speech gestures is intimately linked with the speech production.

Realistic gestures are crucial to accurately simulate real human interactions, in which nonverbal behavior plays a key role in conveying information [13]. Virtual agents —that typically look like humans and interact with them through verbal and nonverbal cues [41]— have been developed for a diverse set of applications, such as serious gaming [27], interpersonal skills training [30], and therapy systems [36]. Interactions with these virtual agents have shown to be more engaging when the agent's verbal behavior is accompanied by appropriate nonverbal behavior [39]. Moreover, it has been shown that manipulating gesture properties of the virtual agent can influence people's perception of agent's emotions [4].

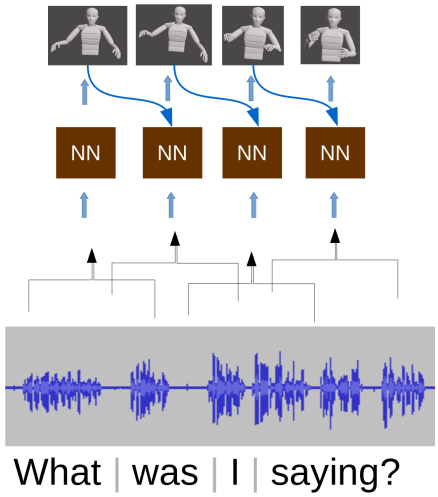

Figure 1: Overview of the autoregressive model.

Traditionally, gesture generation for virtual agents has been done by various rule-based systems [3, 17, 40]. Alongside recent advances in deep learning, data-driven approaches have increasingly gained interest for gesture generation [15, 24, 46]. While early work has considered gesture generation task as a classification which aims to deduce a specified gesture class [6, 31], more recent work has considered it as a regression task which aims to produce continuous motion. We focus on the latter task: *continuous gesture generation*. To date, prior work on continuous gesture generation has used a single input modality: either acoustic or semantic. In contrast, our work makes use of both these input modalities to allow for semantic-aware speech-driven continuous gesture generation. The contributions of this work are the following:

(1) A first model which maps speech acoustic and semantic features into the corresponding continuous 3D gestures and is applicable to both virtual agents and humanoid robots;
(2) A comparison and contrast of the effects of different architectural and modelling choices;
(3) Evaluation of the effect of the two modalities of the speech – audio and semantics — on the resulting gestures in terms of objective measures (such as motion statistics) and observer's subjective perception of the generated gestures.

We additionally extend the publicly available corpus of 3D co-speech gestures, Trinity College dataset [11], with manual text transcriptions.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 2020, Auckland, New Zealand

https://doi.org/doi

2 BACKGROUND

Crucial for interactions with others, verbal and nonverbal behaviors are governed by a complex set of implicit rules [2]. While there are several theories on how gestures are being produced by humans [2, 8, 28], there is a consensus that speech and gestures correlate strongly [20, 26, 34]. In this section, we review some concepts relevant to our work: gesture classification, gesture generation problem formulation and gesture-speech temporal alignment.

2.1 Co-speech gesture types

There are several definitions of gesture types and their properties. In this work, we follow the gesture classification by McNeil [28], who distinguished the following four gesture types:

(1) *Iconic* gestures represent some aspect of the scene being described in speech;
(2) *Metaphoric* gestures represent an abstract concept;
(3) *Deictic* gestures point to an object or orientation;
(4) *Beats* gestures are used to put emphasis and usually correlate with the speech prosody.

The first three types of gestures depend on the content of the speech, its semantics, while the last type only depends on the audio signal, speech acoustics. Hence, systems that ignore semantics of the speech can represent only a small fraction of all the gesture types and to fully model gestures, semantics need to be included.

2.2 Gesture generation

We frame the problem of speech-driven gesture generation as follows: given a sequence of speech features $\boldsymbol{s} = [\boldsymbol{s}_t]_{t=1:T}$ the task is to generate a corresponding gesture sequence $\hat{\boldsymbol{g}} = [\hat{\boldsymbol{g}}_t]_{t=1:T}$ that an agent might perform while uttering this speech, where $t = 1 : T$ means that we consider a sequence of vectors for t in 1 to T.

A speech segment s_t can be represented by a sequence of different features, such as acoustic features (e.g., spectrograms), semantic features (e.g., word embeddings) or a combination of the two. The ground truth gestures g_t and predicted gestures $\hat{g}_t$ can be represented in 3D space as a sequence of joint rotations: $\boldsymbol{g}_t = [\alpha_{i,t}, \beta_{i,t}, \gamma_{i,t}]_{i=1:n}$, n being the number of keypoints of the human body , α, β and γ represent rotations in three axes.

2.3 Gesture-speech alignment

Alignment between gestures and speech is an active research field. It has been investigated for several languages, including French [10], German [1], and English [26, 34]. Below, we focus on prior work which analyzed gesture-speech alignment for the English language.

Loehr [26] found that gestures are typically 0.22s (std 0.13s) earlier than the corresponding speech. Pou et al. [34] did a more detailed analysis aligning different types of gestures with the speech audio peak pitch. The authors found that the onset of beat gestures usually precedes the corresponding speech by 0.35s (std 0.3), that of iconic gestures precedes speech by 0.45s (std 0.4), and the onset of pointing gestures precedes speech by 0.38s (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: we consider 1s of the future speech and 0.5s of the past speech as an input to our model, which is described in Section 5.

3 RELATED WORK

There is extensive previous research on rule-based systems (for review, see [44]), however, with recent advances in deep learning, there has been a shift from rule-based systems [3, 17, 40] to data-driven approaches [15, 24, 46]. Our work contributes to the line of data-driven approaches and we confine our review of related work to these methods. The rest of this section is organized by the input modality used in prior work.

3.1 Audio-signal driven gesture generation

A majority of the work on data-driven gesture generation has been using audio-signal as a single modality of speech used in the model [5, 12, 15, 38]. For example, Sadoughi et al. [38] trained a probabilistic graphical model to generate gestures based on the speech audio-signal and discourse function. The evaluation considered three different hand gestures and two head motions. Hasegawa et al. [15] developed a more general model capable of generating any 3D motion using a deep recurrent neural network, after smoothing was applied as postprocessing step. Kucherenko et al. [24] extended this work by applying representation learning to the human pose and reducing the need for smoothing. Recently, Ginosar et al. [12] applied a convolutional neural network to generate 2D poses from spectrogram features. However, virtual avatars and humanoid robots typically require 3D joint angles and cannot be driven by 2D poses. Our model generates 3D joint angles and uses both audio-signal and text-transcription.

3.2 Text-transcription driven gesture generation

Several recent works mapped from text-transcripts to the corresponding gestures [18, 19, 46]. Ishi et al. [18] generated gesture for input text through a series of probabilistic functions. First, they mapped text from words to word concepts using WordNet [29]. Then, these word concepts were mapped to a gesture function (e.g. iconic and beat) and finally, from a gesture function they mapped to the hand gesture clusters (from the 3D positions). Yoon et al. [46] collected a dataset of 2D gestures and corresponding text transcriptions from YouTube videos, and learned a mapping from the utterance's text to the corresponding gestures by sequence2sequence recurrent neural network (RNN). 2D gestures were mapped to 3D and executed on a humanoid robot NAO. Although these works capture important information from text transcriptions, they may fail to seize the strong dependency of gestures on the acoustic aspects of the speech, such as intonation, prosody, and loudness [35].

3.3 Multimodal gesture generation models

Only a handful of works have used multiple modalities of the speech to predict the corresponding gestures. Neff et al. [31] had a multimodal input to their gesture generation model containing the text, theme, rheme, and focus of each utterance. They also used text-to-concept mapping. From the data on each speaker, a concept-to-gesture mapping for a set of 28 discrete gestures was learned for a particular speaker. Chiu & Marsella [6] used both audio signal and

text transcription as input to their deep-learning based system, to predict a total of 12 gesture classes. Our approach differs from these works, as we aim to generate a wider range of gestures represented in the 3D pose space. Rather than predicting a discrete gesture, our model takes both audio and text as input to produce arbitrary gestures as a sequence of 3D poses.

3.4 Multimodal non-verbal behavior generation

Several works used multimodal input to generate other aspects of non-verbal behavior, such as facial expression, gaze, and head nods [7, 37, 43]. Chu et al [7] used both text and facial expressions of the conversational partner to generate appropriate response, based on encoder-decoder RNN. Sadoughi et al [37] also developed an RNN, which can generate expressive lip movements based on the speech signal and emotional cues. Vougioukas et al. [43] used audio and face images to generate photo-realistic videos using a Generative Adversarial Network (GAN). However, gesticulation generation is a principally different task, since there is much higher variability in hand gestures than in facial movements or head nods and has different governing principles.

4 TRAINING AND TESTING DATA

We develop the gesture generation model using machine learning: we learn a gesture estimator $\hat{\boldsymbol{g}} = F(\boldsymbol{s})$ based on a dataset of human gesticulating, where we have both speech information $\boldsymbol{s}$ (both acoustic and semantic) and gesture $\boldsymbol{g}$ data. In this section, we describe the dataset we used to train our model as well as how we aligned text with audio.

4.1 Speech-Gesture dataset

A prerequisite for deep-learning based gesture generation is the availability of large data-sets. Collecting such datasets directly from internet videos, as in [12, 21, 46], may be an attractive idea, but such data only yields 2D landmarks of joint locations in pixel space, and lacks absolute 3D joint angle data required for most virtual avatars and humanoid robots. Consequently, we resorted to a motion capture dataset, the Trinity Gesture Dataset [11], which is publicly available for research purposes [1]. This dataset consists of audio and motion capture recordings of one male actor speaking freely on various topics, and comprises 244 minutes of data in total. We synchronized files for which the motion and speech were not aligned (due to different starting points and frame drops in the motion capture), and removed the data for the lower-body and fingers. This resulted in 15 upper-body joints out of the original 69. The lower body motion was removed to separate the gesture from locomotion, as the actor was moving extensively in the recordings. The fingers were removed due to poor data quality.

To obtain semantic data, we transcribed the audio recordings in the data. This was accomplished by an initial processing using Google Automatic Speech Recognition (ASR), followed by a manual correction of speech recognition errors, the adding of punctuation, and finally the expansions of contractions (e.g., "I'm" to "I am"). We plan to share this data to facilitate further research.

[1] trinityspeechgesture.scss.tcd.ie

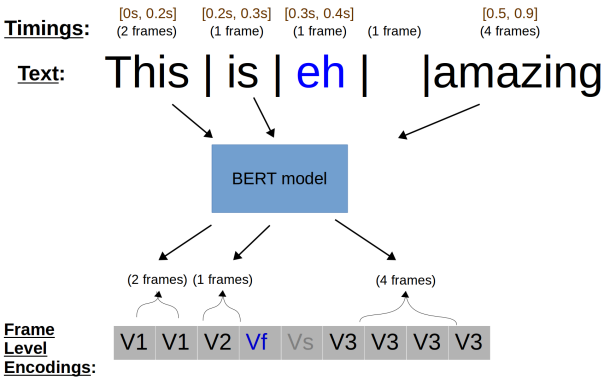

Figure 2: How we encode text into frame-level features? First, the sentence (excluding filler words) is encoded by the BERT model [9]. Then we repeat each vector according to how long the corresponding word was. Filler words and silence are encoded by special fixed vectors, denoted here by Vf and Vs respectively.

4.2 Test segments selection

Two recordings of the dataset, 10 minutes each, were held out for testing. We selected 50 segments of 10s for testing: 30 random segments and 20 semantic segments, in which the ground truth gestures were semantically linked with the speech. Three human annotators marked the times where the ground truth gesture was semantically linked with the speech content. Segments for which all three annotators agreed were used as semantic gestures in our experiments. We used 5s tolerance to find agreement between the annotators.

4.3 Audio-Text alignment

Text transcriptions and audio can have different sequence length and the same length audio sequence can correspond to a different number of words. In other words, text and audio features are not aligned, which complicates modeling. Therefore, to facilitate the training, we align the text with the audio, such that they have the same sequence length. Figure 2 illustrates how we encode words into frame-level features. First, the sentence, excluding filler words, is encoded by BERT model [9], which is the state-of-the-art in many tasks in natural language processing (NLP). Then we repeat each vector according to how long the corresponding word was. Filler words and silence are encoded by special fixed vectors, denoted here by V_f and V_s respectively. They were encoded differently because they represent no semantic information. Filler words typically indicate a thinking process and would occur with different gestures. Therefore, the feature vector for the filler words V_f is calculated as an average of the feature vectors for most common filler words in the dataset. Silence typically have no gestures [14], so silence feature vector V_s is set to be distinct from all the other encoding: it has each dimension equal to -15, while other embeddings have feature values between -10 and 10.

Table 1: Text features used.

BERT encoding generated by a pretrained BERT model [9]
How far we are in the current word
How much is left until the end of the word
Word length
What is the progress within the word (in %)
What is the speaking speed of this word (in syllable per second).

5 AUTOREGRESSIVE SPEECH-DRIVEN GESTURE GENERATION

This section describes our method for generating upper body motion based on speech acoustics and semantics. First, we describe how the speech features were computed, followed by a description of our model architecture. Finally we discuss the training scheme.

5.1 Feature selection

We choose our features following the state-of-the-art in speech audio and text processing. The motion representation was also following common practice. Throughout our experiments, we consequently use frame-synchronised features with 20 fps.

Common to previous research in gesture generation [11, 12], we represent speech audio with mel-spectrogram features in a log-scale. For this, we extracted 64-dimensional feature-vectors using a window length of 0.1 s and hop length 0.05 s (resulting in 20 fps).

For semantic features, we use the BERT model [9], pretrained on an English Wikipedia: each sentence of the text transcription was encoded by BERT resulting in 768 features per word. We added five more features related to the current frame in the word or the word generally, which are provided in Table 1. To match audio frequency, word-level features were first transformed into frame-level features, as described in Section 4.3 and then upsampled from 10fps to 20fps.

To extract motion features, we down-sampled the motion capture data to 20 fps, and converted the joint angles to an exponential map representation expressed relative to a reference T-pose. This representation is commonly used in computer animation. Thereafter, we reduced the dimensionality by applying PCA and kept the minimum number of components that contained at least 95% of the variance of the training data, following [46], resulting in 12 components.

5.2 Model architecture

In order to generate co-speech gestures as a sequence of 3D poses we process the input audio and text vectors sequentially using overlapping sliding windows as illustrated in Figure 1. Each sliding window contains 0.5s (10 frames) of past speech and 1s (20 frames) of future speech features, which is grounded in the research on gesture-speech alignment, as described in Section 2.3.

Our model architecture is illustrated in Figure 3. First, each frame's text and audio features are encoded by the same feed-forward neural network to reduce dimensionality. The resulting encodings inside the current window are then concatenated into a long vector. Afterwards, several fully connected layers are applied. Our model is autoregressive: we feed the model predictions back to the model as can be seen in the model illustration. This is done

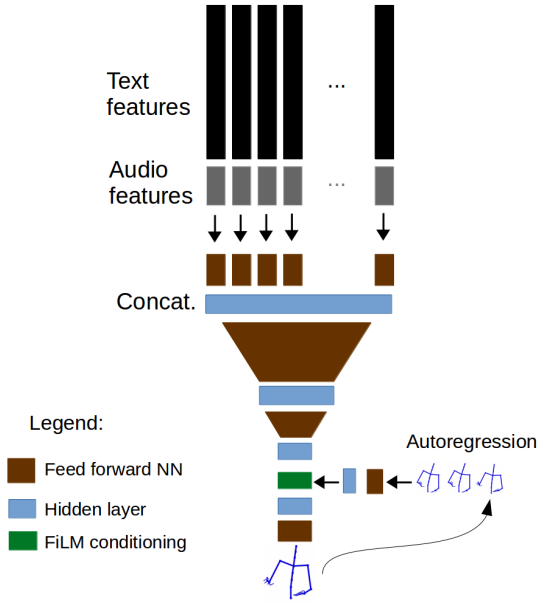

Figure 3: The autoregressive model architecture. The text and audio features of each frame are encoded independently and resulting encodings are then concatenated. Afterwards, several fully connected layers are applied. The output pose is fed back into the model in an autoregressive fashion via FiLM conditioning.

to ensure motion continuity. As a way to condition on the information from the previous poses, we use FiLM conditioning [33], which generalizes regular conditioning. FiLM applies an affine transform to network activations $\boldsymbol{x}$, where scaling $\boldsymbol{\alpha}$ and offset $\boldsymbol{\beta}$ vectors are produced by a neural net taking previous poses as input,

$$FiLM(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{x} * \boldsymbol{\alpha} + \boldsymbol{\beta} \quad (1)$$

where both multiplication and summation are element-wise. The final layer of the model and final layer of conditioning parameters generation contain one linear layer at the end to not restrict the resulting values by the activation function.

5.3 Training procedure

We train our model on sequences of aligned speech audio, text, and corresponding gestures from the dataset. Each sequence for the training contains 70 frames from a larger recording. The first 10 frames are used to set the past context and last 20 to set the future context, hence, only 40 frames of those sequences were used to optimize the model. The model was optimized end-to-end using stochastic gradient descent (SGD) with Adam optimizer [22] with the following loss function:

$$\boldsymbol{loss} = MSE(\boldsymbol{g}, \hat{\boldsymbol{g}}) + \alpha\, MSE(\Delta\boldsymbol{g}, \Delta\hat{\boldsymbol{g}}) \quad (2)$$

where $\boldsymbol{g}$ and $\Delta\boldsymbol{g}$ are the ground truth position and velocity, $\hat{\boldsymbol{g}}$ and $\Delta\hat{\boldsymbol{g}}$ are the positions and velocities of the model prediction and MSE stands for Mean Squared Error. The coefficient α was set empirically to 0.6, based on subjective evaluations by the authors. The velocity penalty can be seen as an improvement on the penalty as used

by Yoon et al. [46], which simply penalized the absolute value of velocity. We enhance it by enforcing velocity to be close to the velocity of the ground truth, rather than just to be small.

We have observed that information from the previous poses which are fed back autoregressively tend to dominate information from the speech: our initial model moved independently of speech and quickly converged to a static pose. This is a common hurdle in generative models of speech feature sequences [16, 42, 45]. To contract this, we pretrain our model without autoregression for the first seven epochs (the number of epochs was chosen empirically). After seven epochs the model starts receiving autoregressive inputs. This approach helps the network to learn useful features from the speech features during those pretraining epochs and this ability to attend to speech is not lost during further training.

Additionally, when we start providing previous predictions to the model, we initially train it using teacher forcing: instead of its prediction the model receives the ground truth poses from the previous frames. Teacher forcing is annealed during training over time: initially, the model receives its prediction instead of the ground truth poses every 16 frames (for two consecutive frames), which increased to every eight frames after the first epoch, to every four frames after the next epoch, and to every single frame after that. Therefore, after five epochs of training with autoregression, our model stops receiving teacher forcing, and receives its own predictions instead. This approach vastly helps the model to recover from its own mistakes.

5.4 Hyper-parameters

For the experiments in this paper, we chose hyper-parameter searching tool called Tune [25]. We performed random search over 600 configurations with velocity loss as criterion. We used the following hyper-parameters: Speech encoding dimensionality at each frame is 124, resulting in a 124*30 = 3720-dimensional speech segment encoding in the first layer. The subsequent three layers had a dimensionality of 612, 256, and 45 (which is equal to the output dimensionality), respectively. Three previous poses were encoded by a neural network into a 512-dimensional vector containing both α and β for the FiLM conditioning. We used tangent hyperbolic (TanH) as an activation function and trained our model with the batch size 64 and the learning rate 0.0001. For regularization, we applied dropout probability 0.2 on each layer, except for the previous poses encoding, where the dropout was increased 0.8 to prevent the model from attending too much to the previous poses.

6 OBJECTIVE EVALUATION

Our work aims to allow semantic-aware gesture generation by combining both speech acoustic and semantic information to produce continuous gestures, as opposed to a discrete gesture class.

As previous work on continuous gesture generation has exclusively used a single modality, it is difficult to find baselines that allow for a fair comparison with our model. Therefore, we evaluate the importance of various components of our system by individually manipulating them, resulting in seven different variants of the system including the full model (see Table 2). In this section, we describe the objective measures, we report and discuss the experimental results.

Table 2: The seven different variants of the system as used in the evaluation.

Model	Description
Full	The proposed method
No FiLM	Concatanation instead of FiLM
No Velocity penalty	The velocity loss is removed
No Autoregression	The previous poses are not used
No PCA	No PCA is applied to the gesture poses
No Audio	Only text is used as input
No Text	Only audio is used as input

6.1 Objective Measures

There is no consensus in the field about which measure should be used to evaluate the quality of generated gestures. We strive to have a common evaluation measure for the gesture generation field, so we mainly use the metrics proposed by previous researchers.

Since the purpose of gesture generation is not to reproduce one specific *true* position, but rather to produce a plausible motion, as objective measures, we considered only motion statistics. In particular, we evaluated the average values of acceleration and jerk, and speed histograms of the produced motion, in line with Kucherenko et al. [24]. To obtain these statistics, the gestures were first converted from the joint angle space into the 3D coordinates.

6.2 Average motion statistics

Table 3 illustrates average acceleration and jerk over 50 test samples for the ground truth, proposed model and the different ablations to the proposed method. The ground truth statistics are given as the best achievable values because the perfect system would have the same motion statistics as the human motion.

We can observe that the proposed model is much slower than the original motion. Our model moves less probably because it is deterministic and hence produces gestures closer to the mean pose.

Not using PCA results in higher acceleration and jerk and made the model statistics closer to the ground truth. Our intuition for that is that PCA reduced variability in the data, which resulted in over-smoothed motion. These results might indicate that we should not use PCA. Having no audio or no text available made the produced gestures even slower. That is probably because there was a weaker input signal to the model and hence it gesticulated closer to the mean pose. Both FiLM conditioning and the velocity penalty seem to have very little effect on the motion statistics and are therefore not the most central aspects of the model. Autoregression is the key aspect of our system and we can clearly see it in this evaluation: without the autoregression model loses continuity and has a too high jerk of the motion, while having higher acceleration as well.

6.3 Motion acceleration histograms

The values above were averaged over all the time-frames and all the 3D points of the body. To investigate the motion statistics in more details, we computed acceleration histograms of the generated motions and compared those against histograms derived from the ground truth. We calculated the relative frequency of different acceleration values over time-frames in all 50 test sequences, split

Table 3: Objective evaluation of our systems: mean and standard deviation over 50 samples

Model	Acceleration	Jerk
Full model	0.33 ± 0.05	0.29 ± 0.03
No PCA	0.55 ± 0.09	0.46 ± 0.06
No Audio	0.26 ± 0.05	0.18 ± 0.03
No Text	0.21 ± 0.02	0.25 ± 0.03
No FiLM	0.34 ± 0.05	0.27 ± 0.04
No Velocity loss	0.34 ± 0.04	0.29 ± 0.03
No Autoregression	1.02 ± 0.19	1.64 ± 0.31
Ground truth	**1.42** ± 0.42	**1.12** ± 0.33

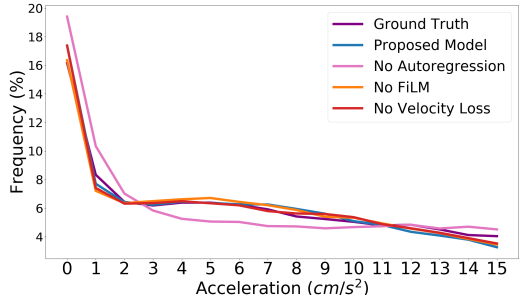

Figure 4: Acceleration distributions for different models and the ground truth.

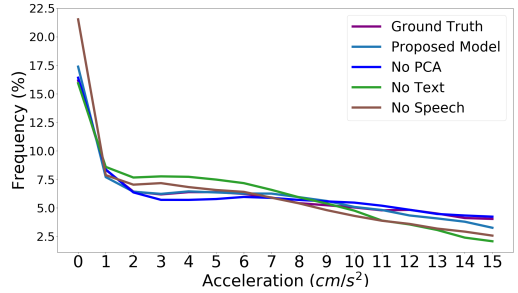

Figure 5: Acceleration distributions for different input or output data.

into bins of equal width. For easy comparison, our histograms are visualized as line plots rather than bar plots.

Figure 4 illustrates the acceleration histogram for the different changes in model architecture we considered. We observe that most of the model versions follow the acceleration distribution of the ground truth quite closely, while the model without autoregression deviates from it. That indicates two things: 1) model is not influenced strongly neither by FiLM conditioning nor by velocity loss; 2) autoregression is important to produce motion with similar motion statistics as in the human motion.

Acceleration histograms for different input/output data are shown in Figure 5. We can observe that having no text as input affects the model the most by making the acceleration smaller, which confirms the results from Table 3 and probably means that without text the model produces mainly beat gestures, which are having different characteristics than the other types of gestures. Having no audio also decrees the acceleration of the produced gestures, which might also indicate that we are modeling different gesture types. Removing PCA increases acceleration, making the distribution more similar to the ground truth. That is intuitive that training our model in the PCA space leads to reduced variability.

All those numerical evaluations are valuable, but they cannot tell us anything about people's perceptions of the produced gestures. To investigate the human perception of the gestures we conducted several user studies, which we describe in the following section.

7 PERCEPTUAL STUDIES

Two perceptual studies were conducted: we evaluated participants' perception of a virtual character's gestures as produced by the seven variants of our model as described in Table 2 (study 1), after which we compared the most preferred variant of our model from the first study was compared to the ground truth (study 2). The measures and experimental procedure, as described in this section, were identical for both studies. In this section, we describe the subjective measures, participants, experimental procedure, and report on and discuss the results for both studies.

7.1 Subjective Measures

We measured the perceived human-likeness of the virtual character's motion and how they related to the character's speech. Specifically, we adapted the questions from recent papers on co-speech gesture generation [12, 46], resulting in the following questions, "in which video...":

(1) ... are the character's movements most human-like?
(2) ... do the character's movements most reflect what the character says?
(3) ... do the character's movements most help to understand what the character says?
(4) ... are the character's voice and movement more in sync?

For each question, participants were able to chose which of the videos best corresponded to the question. Additionally they were able to indicate that they perceived the property to be equal in both videos for a particular question (e.g., "The character's movements help to understand the speech content equally in both videos").

7.2 Experimental setup

We compared each of the six variants of our model for which we manipulated a single component with the full model, resulting in six different comparisons. Participants were recruited using Amazon's Mechanical Turk (AMT) and were assigned to one of the six comparisons. Participants could complete the study only once, which was ensured by checking their AMT worker id and IP address. Participants were asked to evaluate 20 video pairs and rate them on the subjective measures. Six video pairs were added as attention tests, resulting in a total of 26 video pairs. 10 video pairs were randomly sampled from a pool of random segments (28 videos) and another 10 from a pool of semantic segments (20 videos). These video pairs were then randomly shuffled. Additionally, two video pairs in the random pool were used for the attention check as described below. To mitigate ordering effects, the video's placement (left or right)

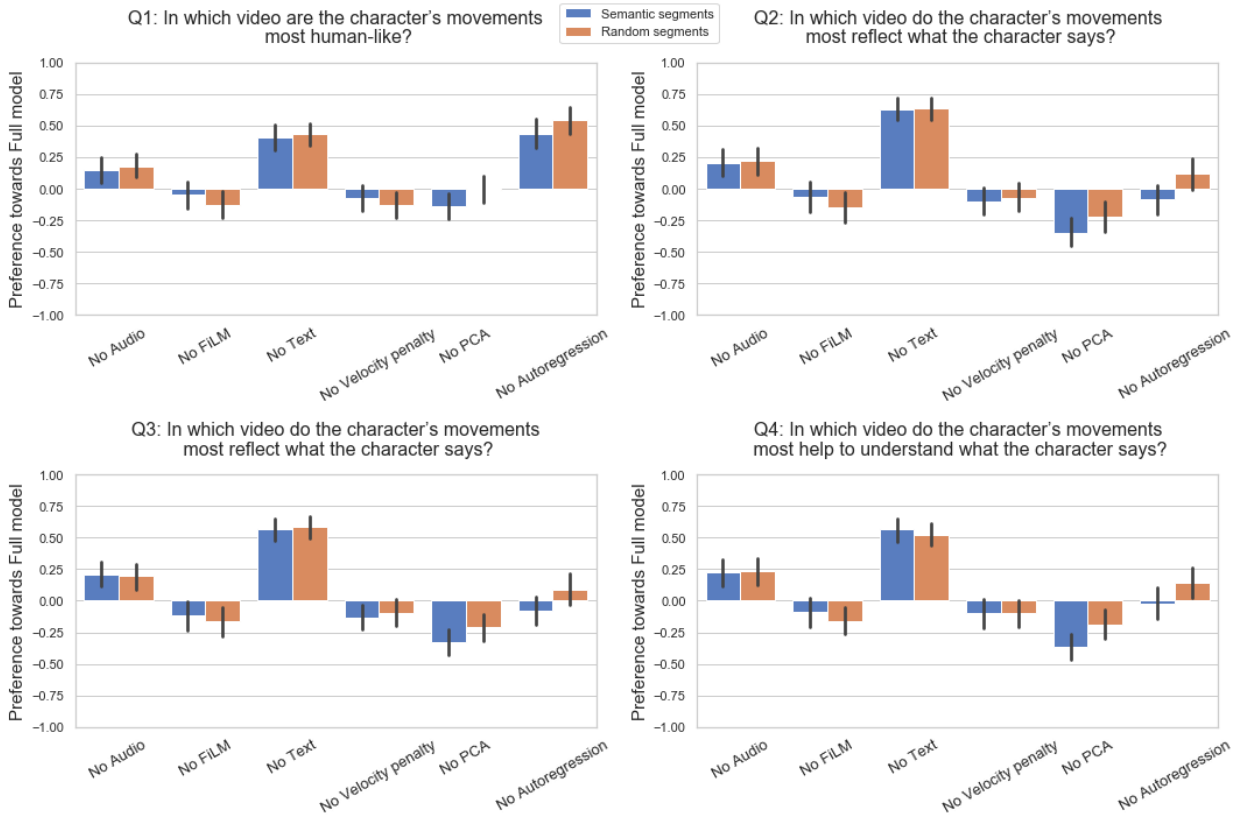

Figure 6: The first user study results: comparing different version of out model by a pairwise preference choice. Four questions were asked about each pair of videos as illustrated above the figures. Preference towards full model was calculated and 95% confidence intervals are shown.

was determined randomly. Video samples for all the model variants were uploaded anonymously to Vimeo and can be found at vimeo.com/showcase/6577888.

7.3 Participants

7.3.1 Study 1. 123 participants (mean age = 41.8 ± 12.3, 52 male, 70 female, 1 other) remained after exclusion of 477 participants who failed to complete the attention checks, experienced technical issues, or stopped the study prematurely. The majority was American (N = 120), and reported to have moderate experience with technology (mean = 3.68 ± 0.86).

7.3.2 Study 2. 20 participants (mean age = 39.1 ± 8.4, 9 male, 11 female) remained, after exclusion of 31 participants who failed to complete the attention checks, experienced technical issues, or stopped the study prematurely. The majority was American (N = 18) and reported to have moderate experience with technology (mean = 3.79 ± 0.76).

7.4 Experimental Procedure

Participants were presented with two videos side by side, which they could replay as often as needed. For each video pair, we asked participants to submit their answer to the four questions as described in Section 7.1.

We included six items that served as attention checks throughout the trial. For each participant, the audio of one random video was heavily distorted in the second and seventeenth video pairs, one of the videos was heavily distorted in the seventh and twenty-first video pairs, and two of the same videos were presented in the thirteenth and twenty-fourth video pairs. Participants were asked to report video pairs for which there were issues with the video or audio and for which video. Participants who had failed two attention checks where the video or audio had been manipulated were automatically excluded from the study.

After a participant accepted the human intelligent task (HIT), they were presented with our webpage, embedded in the HIT's interface, which would show the questions. First, we collected participants' demographics. On the next page, the task instructions were given.

After reading the instructions, participants went through a training phase to familiarize themselves with the interface and task. The training phase consisted of five items which were not included in the analysis, with video segments not present in the study showcasing gestures of different quality.

7.5 Results

7.5.1 Comparison of Different Model Variants. We conducted a binomial test with Holm-Bonferroni correction to analyze the data with binomial options being either Full model or one of the model variants, thus excluding ties. The analysis was done in a double-blind fashion such that the conditions were obfuscated during analysis and revealed after the statistical tests had been performed. 24 comparisons were excluded due to being reported as having technical issues. The results are shown in Figure 6.

We can see from the evaluation of the "No Text" model that removing semantic information from the model drastically decreases the perceived human-likeness of the produced gestures and how much they are linked to speech: participants preferred the Full Model over the model without Text in all the four questions asked with p<.0001. This confirms that semantics are important for appropriate gesture generation.

The "No Audio" model is unlikely to generate beats, and might not follow an appropriate speech rhythm when used with a speaking avatar. Results in Figure 6 confirm that: we can see that removing acoustic information from the model decreases its quality. In terms of human-likeness (Q1) there was a statistical difference only for random samples with p< .01. For the Q2 and Q3, the "Full model" was also preferred model with p<.01. For the last question participants preferred the full model with p<.05.

Removing autoregression from the model affects negatively only perceived naturalness (p<.0001), as can been seen in Figure 6. It confirms what we established by the numerical evaluation: the proposed model without autoregression produces jerky gestures, which do not look natural, but the jerkiness does not influence whether the gestures are semantically linked to the speech content.

Removing FiLM or velocity penalty did not make a statistical difference to the gestures perception. That indicates that those components of the model are not critical for the model.

Our model without PCA provided unexpected results. We can observe from the videos that removing PCA resulted in higher variability of gestures. When it comes to human-likeness, there was no statistical difference. For the Q2 and Q3 (see Figure 6) the model without PCA was significantly better (p<.001). For Q4 there was statistical difference only for "semantic" sample with p<.001, since they probably require more sophisticated gesticulation. Taking it all together, participants preferred the model without PCA, so it was chosen as our final model for the following comparisons.

7.5.2 Comparing to the ground truth gestures. After analyzing different variants of our model and choosing the best one ("No PCA") we compared it to the ground truth gestures using the same procedure as before. The Binomial test was used to analyze the differences, followed by Holm-Bonferroni correction of the *p*-values.

The results are given in Figure 7. We can clearly see that it was easy for the participants to find which condition was the ground truth. When we look at the videos, we can see that the generated gestures have very little motion in the spine and shoulder. It is not surprising that the model could not connect whole spine movements with speech since they are random and not correlated with the speech. That is why all the models predict average values for the spine and shoulder positions, resulting in the avatar looking somewhat stiff. We believe that was the main reason why it was so easy for the participants to identify the ground truth.

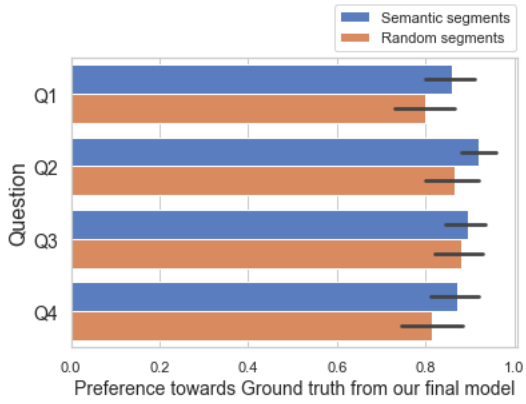

Figure 7: The second user study results: comparing the best model with the ground truth gestures using the same questions as before. Preference towards Ground Truth was calculated and 95% confidence intervals are shown.

8 CONCLUSIONS

We present a new machine learning-based model[2] for co-speech gesture generation. To the best of our knowledge, this is the first model capable of producing continuous gestures, while using both acoustics and semantics of the speech. We evaluate different architecture choices using both objective and subjective measures. Our experiments indicate that:

(1) Using both modalities of the speech – audio and text – can enhance gesture-generation models and enable them to generate both beat gestures and semantically-linked gestures, such as metaphoric gestures.
(2) Autoregressive connections can enforce continuity of the gestures, without vanishing-gradient issues and with few parameters to learn.
(3) Applying PCA to the motion space can harm the variability of the produced gestures, since it might restrict the model by removing perceptually-important variation from the data.

The main limitation of our work is that it requires an annotated dataset (with text transcriptions), which is time-consuming and labor-intensive. To overcome this we might train our model directly on the transcriptions for the Automatic Speech Recognition.

We see several possible directions for future work:

- Firstly, the model can be made probabilistic, so that, instead of a single pose, it predicts a probability distribution over "plausible" poses for the next frame. This should solve issues related to the MSE loss promoting "mean-collapsed" gestures close to to the average pose.
- Secondly, our gesture generation model could be applied to a humanoid robot, such as NAO (as in [46]).
- Finally, it would be very interesting to test how those gestures may influence human-agent interaction

[2]Code will be made publicly available upon acceptance to enhance reproducibility.

REFERENCES

[1] 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 (GeSpIn 2011)*.

[2] Timothy W Bickmore. 2004. Unspoken rules of spoken interaction. *Commun. ACM* 47, 4 (2004), 38–44.

[3] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. Beat: The behavior expression animation toolkit. In *Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*. ACM.

[4] Gabriel Castillo and Michael Neff. 2019. What do we express without knowing?: Emotion in Gesture. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 702–710.

[5] Chung-Cheng Chiu and Stacy Marsella. 2011. How to train your avatar: A data driven approach to gesture generation. In *International Workshop on Intelligent Virtual Agents (IVA'11)*. Springer, 127–140.

[6] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: A deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents (IVA '15)*. Springer.

[7] Hang Chu, Daiqing Li, and Sanja Fidler. 2018. A Face-to-Face Neural Conversation Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7113–7121.

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

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Gaëlle Ferré. 2010. Timing relationships between speech and co-verbal gestures in spontaneous French. In *Language Resources and Evaluation, Workshop on Multimodal Corpora*, Vol. 6. 86–91.

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

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

[13] Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences* 3, 11 (1999), 419–429.

[14] Maria Graziano and Marianne Gullberg. 2018. When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in psychology* 9 (2018), 879.

[15] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *International Conference on Intelligent Virtual Agents (IVA '18)*. ACM, 79–86.

[16] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. 2019. MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *arXiv preprint arXiv:1905.06598* (2019).

[17] Chien-Ming Huang and Bilge Mutlu. 2012. Robot behavior toolkit: Generating effective social behaviors for robots. In *International Conference on Human Robot Interaction (HRI '12)*. ACM/IEEE.

[18] Carlos T Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. 2018. A Speech-Driven Hand Gesture Generation Method and Evaluation in Android Robots. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3757–3764.

[19] Ryo Ishii, Taichi Katayama, Ryuichiro Higashinaka, and Junji Tomita. 2018. Generating body motions using spoken language in dialogue. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA '18)*. ACM, 87–92.

[20] Jana M Iverson and Esther Thelen. 1999. Hand, mouth and brain. The dynamic emergence of speech and gesture. *Journal of Consciousness Studies* 6, 11-12 (1999), 19–40.

[21] Patrik Jonell, Taras Kucherenko, Erik Ekstedt, and Jonas Beskow. 2019. Learning Non-verbal Behavior for a Social Robot from YouTube Videos.. In *ICDL-EPIROB 2019 Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions*. Oslo, Norway.

[22] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR '15)*.

[23] Stefan Kopp, Hannes Rieser, Ipke Wachsmuth, Kirsten Bergmann, and Andy Lücking. 2007. Speech-gesture alignment. In *3rd Conference of the International Society for Gesture Studies*.

[24] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing Input and Output Representations for Speech-Driven Gesture Generation. In *Proc. IVA*, Vol. 19. ACM, Paris, France, 97–104. https://doi.org/10.1145/3308532.3329472

[25] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118* (2018).

[26] Daniel P Loehr. 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology* 3, 1 (2012), 71–89.

[27] Samuel Mascarenhas, Manuel Guimarães, Rui Prada, João Dias, Pedro A Santos, Kam Star, Ben Hirsh, Ellis Spice, and Rob Kommeren. 2018. A Virtual Agent Toolkit for Serious Games Developers. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 1–7.

[28] David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

[29] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[30] Shannon Monahan, Emmanuel Johnson, Gale Lucas, James Finch, and Jonathan Gratch. 2018. Autonomous agent that provides automated feedback improves negotiation skills. In *International Conference on Artificial Intelligence in Education*. Springer, 225–229.

[31] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)* 27, 1 (2008), 5.

[32] Miriam A Novack, Elizabeth M Wakefield, and Susan Goldin-Meadow. 2016. What makes a movement a gesture? *Cognition* 146 (2016), 339–348.

[33] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[34] Wim Pouw and James A. Dixon. 11.-13.09.2019. Quantifying gesture-speech synchrony. In *6 th Gesture and Speech in Interaction (GESPIN '19)*. Paderborn : Universitätsbibliothek, 76–80.

[35] Wim Pouw, Steven J Harrison, and James A Dixon. 2019. Gesture–speech physics: The biomechanical basis for the emergence of gesture–speech synchrony. *Journal of Experimental Psychology: General* (2019).

[36] Lazlo Ring, Timothy Bickmore, and Paola Pedrelli. 2016. Real-Time Tailoring of Depression Counseling by Conversational Agent. *Iproceedings* 2, 1 (2016), e27.

[37] Najmeh Sadoughi and Carlos Busso. 2018. Expressive speech-driven lip movements with multitask learning. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 409–415.

[38] Najmeh Sadoughi and Carlos Busso. 2019. Speech-driven animation with meaningful behaviors. *Speech Communication* 110 (2019), 90–100.

[39] Maha Salem, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2011. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *2011 Ro-Man*. IEEE, 247–252.

[40] Giampiero Salvi, Jonas Beskow, Samer Al Moubayed, and Björn Granström. 2009. SynFace: Speech-driven facial animation for virtual speech-reading support. *EURASIP Journal on Audio, Speech, and Music Processing* (2009), 3.

[41] William R Swartout, Jonathan Gratch, Randall W Hill Jr, Eduard Hovy, Stacy Marsella, Jeff Rickel, and David Traum. 2006. Toward virtual humans. *AI Magazine* 27, 2 (2006), 96–96.

[42] Benigno Uria, Iain Murray, Steve Renals, Cassia Valentini-Botinhao, and John Bridle. 2015. Modelling acoustic feature dependencies with artificial neural networks: Trajectory-RNADE. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4465–4469.

[43] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2019. End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 37–40.

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

[45] Xin Wang, Shinji Takaki, and Junichi Yamagishi. 2018. Autoregressive neural f0 model for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 8 (2018), 1406–1419.

[46] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *International Conference on Robotics and Automation (ICRA '19)*. IEEE.