# Learn to Gesture: Let Your Body Speak

Tian Gan*
Shandong University
gantian@sdu.edu.cn

Zhixin Ma
Shandong University
zhixinma@sdu.edu.cn

Yuxiao Lu
Shandong University
luyuxiao0311@gmail.com

Xuemeng Song
Shandong University
songxuemeng@sdu.edu.cn

Liqiang Nie
Shandong University
nieliqiang@gmail.com

## ABSTRACT

Presentation is one of the most important and vivid methods to deliver information to audience. Apart from the content of presentation, how the speaker behaves during presentation makes a big difference. In other words, gestures, as part of the visual perception and synchronized with verbal information, express some subtle information that the voice or words alone cannot deliver. One of the most effective ways to improve presentation is to practice through feedback/suggestions by an expert. However, hiring human experts is expensive thus impractical most of the time. Towards this end, we propose a speech to gesture network (POSE) to generate exemplary body language given a vocal behavior speech as input. Specifically, we build an "expert" Speech-Gesture database based on the featured TED talk videos, and design a two-layer attentive recurrent encoder-decoder network to learn the translation from speech to gesture, as well as the hierarchical structure within gestures. Lastly, given a speech audio sequence, the appropriate gesture will be generated and visualized for a more effective communication. Both objective and subjective validation show the effectiveness of our proposed method.

**ACM Reference Format:**
Tian Gan, Zhixin Ma, Yuxiao Lu, Xuemeng Song, and Liqiang Nie. 2019. Learn to Gesture: Let Your Body Speak. In *ACM Multimedia Asia (MMAsia '19), December 15–18, 2019, Beijing, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3338533.3366602

## 1 INTRODUCTION

For students, scholars, corporates and professionals, oral presentation is one of the most common and effective ways to deliver ideas or knowledge to the audience. Studies have shown that effective oral presentation skills are crucial in diverse fields, including education, business and politics [6]. When a person speaks in public, the listeners will judge the speaker and his message based on what they see as well as what they hear. Appropriate body language can effectively enhance the delivered message [11]. However, most of the junior speakers are not good at body language. In the light of this, our goal is to develop a system to perform automatical body language suggestion for presentation, so that people can practice presentation through detailed instructions.

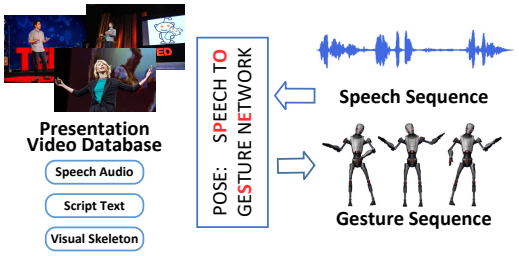


**Figure 1: Conceptual figure of our proposed system.**

Although it seems very desirable, performing automatical body language generation is a challenging task. First, despite the huge research and commercial interests, the lack of large-scale presentation datasets hinders the research progress. Second, accurate segmentation of gesture is difficult because of the unlimited types of gesture. To address the raised issues, we developed a speech to gesture network (POSE), a two-layer attentive recurrent encoder-decoder network, to "translate" speech to gestures. Given a speech audio signal as input, we first utilized the script structure and audio signal tokenization information to structure the speech into sequence-unit-frame hierarchy. We further extracted the acoustic features for each speech frame, and fed it into our proposed POSE. The translation from speech to gesture, speech-gesture interaction context, as well as the hierarchical structure inside the speech and gesture, are jointly encoded in the network.

In summary, the contributions of our proposed POSE for automatically generating body languages are as follows:

- We constructed a TED Speech-Gesture Presentation dataset, which consists of over 100k speaker presentation video segments (containing speech audio, gesture skeleton frames, and presentation transcripts) from 2,582 unique TED talk videos. The dataset is much larger than the ever-seen datasets for multimodal social interaction analysis, and can greatly benefit the future researchers in this community.
- We developed a two-layer attentive recurrent encoder-decoder framework, which successfully translates speech into gesture and encodes their hierarchical structure and speech-gesture interaction. Both objective and subjective evaluations show the effectiveness of our proposed method.
- We introduced a novel application that help people learn the exemplary body language at low cost.

*Tian Gan is the corresponding author.


*MMAsia '19, December 15–18, 2019, Beijing, China*

ACM ISBN 978-1-4503-6841-4/19/12…$15.00
https://doi.org/10.1145/3338533.3366602

## 2 RELATED WORK

**Presentation Assessment** Conventionally, the quality of a presentation is usually evaluated by painstaking manual analysis with experts. With the recent advancement of sensor technologies, research on automated presentation analysis has commenced. These studies applied various sensors to analyze different aspects of a presentation. Some focus only on the verbal aspects, such as fluency of speech, liveliness, speaking rate, and affective state of speech [1, 5]. Other studies focus on both verbal and non-verbal aspects of a presentation, and applied multi-modal analysis [7, 13], however, they only gave statistics about how the speaker behaves during a presentation, no further suggestions about the correct/suggested behaviors. Furthermore, due to the high cost of recording real-world data, the previous research all work on a small-scale dataset consists of recording with a few number of subjects.

**Co-Speech Gesture in Human Communication** Communication is a concept of information exchange incorporating visual perception, speech perception and the understanding of meaning. Gestures, as part of visual perception and synchronized with verbal information, representing an integral component of human communicative behavior are a key concept of human social interaction [12]. Gestures not only help the human speakers to illustrate what they express in speech, more crucially, they help to convey information which speech alone sometimes cannot provide, as in referential, spatial or iconic information [11]. At the same time, human listeners have been shown to be well-attentive to information conveyed via such non-verbal behaviors [8, 10]. In addition, providing multiple modalities helps to dissolve ambiguity that is common in the unimodal communication and, consequently, to increase robustness of communication. One of the most accepted classification of co-verbal gesture was introduced by McNeill [15], who distinguished four main types of co-verbal gestures: *beats*, *deictics*, *iconics*, and *metaphorics*.

**Gesture and Social Robotics** One of the main objectives of social robotics research is to design and develop robots that can engage in social environments in a way that is appealing and familiar to human interaction partners [20]. Adequate head movements and manual gestures have shown positive impacts for multimodal dialogue systems and social robotics [2]. Social robotics incorporate both speech and gesture synthesis; and in most cases the gestures are limited to the predefined repertoires of gestures [17, 20, 21]. A different line of research uses the data-driven method, by learning the relationship between speech and gesture. However, an accurate segmentation of gesture is difficult. These different lines of research advance knowledge in how robots might use gestures to improve human-robot interaction and peoples' perceptions of robots. Nevertheless, for robots to realize the full potential of using gestures, a better understanding of how human use different gestures, and how speech interacts with gestures are still needed.

## 3 PROPOSED METHOD

Communication is with multimodal strategies, and language is considered as being primarily governed by "combinations of discrete units", organized hierarchically and unfolding linearly [15]. People consciously or unconsciously use head motion, hand gestures, and facial expressions while speaking. Inspired by these, we model

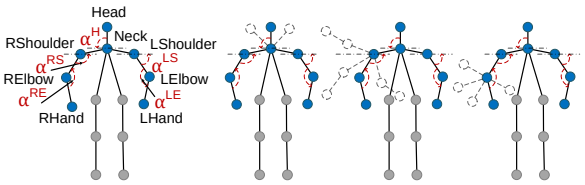


**Figure 2: Illustration of posture discretization.**

a presentation video as **speech** and **gesture** with hierarchical structure. The goal of our work is to automatically generate gesture frame sequences given speech frames. Formally, given a **speech unit sequence** $\mathbf{s} = \{\mathbf{s}_0, \ldots, \mathbf{s}_n\}$, our model translates it into a **gesture unit sequence** $\mathbf{g} = \{\mathbf{g}_0, \ldots, \mathbf{g}_m\}$, where $\mathbf{s}_i$ and $\mathbf{g}_j$ are the **speech units** and **gesture units**, respectively. To refer to individual **speech frames** or **skeleton frames** of speech or gesture unit sequences, we use the notation $s_{i,u}$ and $g_{j,v}$, which $u$ nad $v$ are the indexes for the speech frames and skeleton frames.

### 3.1 Speech and Gesture Representation

*3.1.1 Speech Feature Extraction.* Each speech frame $s_{i,u}$ is divided into $n_{i,u}$ 20ms audio chunks. Short time fourier transformations are applied on each of them, and the frequencies are analyzed in [0, 8K] Hz. This results in a 161-dimensional spectrum fingerprint vector, where each number in this vector represents how much energy is in each 50hz band. These $n_{i,u}$ audio chunks' spectrum fingerprint vector are further concatenated along temporal dimension, and resulted in a spectrogram $\mathbf{P} \in \mathbb{R}^{161 \times n_{i,u}}$.

To extract speech audio features, the Convolutional Neural Networks are used over each $P$ along both the frequency and time axes. The network structure consists of two convolutional layers, following by two fully-connected layers. All convolutional layers have 3x3 filter size and stride size set as one.The network has 64 feature maps in all convolutional layers, followed by a 3x1 max-pooling. ReLU is used as the activation function. The fully-connected layers all have 1,024 units, resulting in a 1024-dimensional representation $\mathbf{x}_{i,u}$ for speech frame $s_{i,u}$.

*3.1.2 Posture Discretization.* Instead of the use of predefined gestures, the sequences of posture are used to characterize human gesture. Particularly, the human posture space are quantified into 1200 states by enumerating the human skeleton joint orientation in 2D space. As shown in Figure 2, eight joints (colored in blue) are utilized to calculate five angles to define the gestures: $\alpha^{\mathrm{H}} \in [60°, 90°, 120°]$, the angle between "Head-Neck" and horizontal line; $\alpha^{\mathrm{RS}}, \alpha^{\mathrm{LS}} \in [60°, 112.5°, 157.5°, 210°]$, the angle between "RElbow-RShoulder" and horizontal line, and the angle between "LElbow-LShoulder" and horizontal line, respectively; $\alpha^{\mathrm{RE}}, \alpha^{\mathrm{LE}} \in [90°, 135°, 180°, 225°, 270°]$, the angle between "RHand-RElbow" and "RElbow-RShoulder", and the angle between "LHand-LElbow" and "LElbow-LShoulder", respectively.

For each skeleton frame $g_{j,v}$, the value of each angle is calculated by the position of the coordinates, and skeleton frames are matched into one of these $|\alpha^{\mathrm{H}}| * |\alpha^{\mathrm{RS}}| * |\alpha^{\mathrm{LS}}| * |\alpha^{\mathrm{RE}}| * |\alpha^{\mathrm{RE}}| = 1200$ states to represent different human postures. At last, the onehot representation of human posture state is embedded into a 128 dimension vector $y_{j,v}$ to represent $g_{j,v}$.

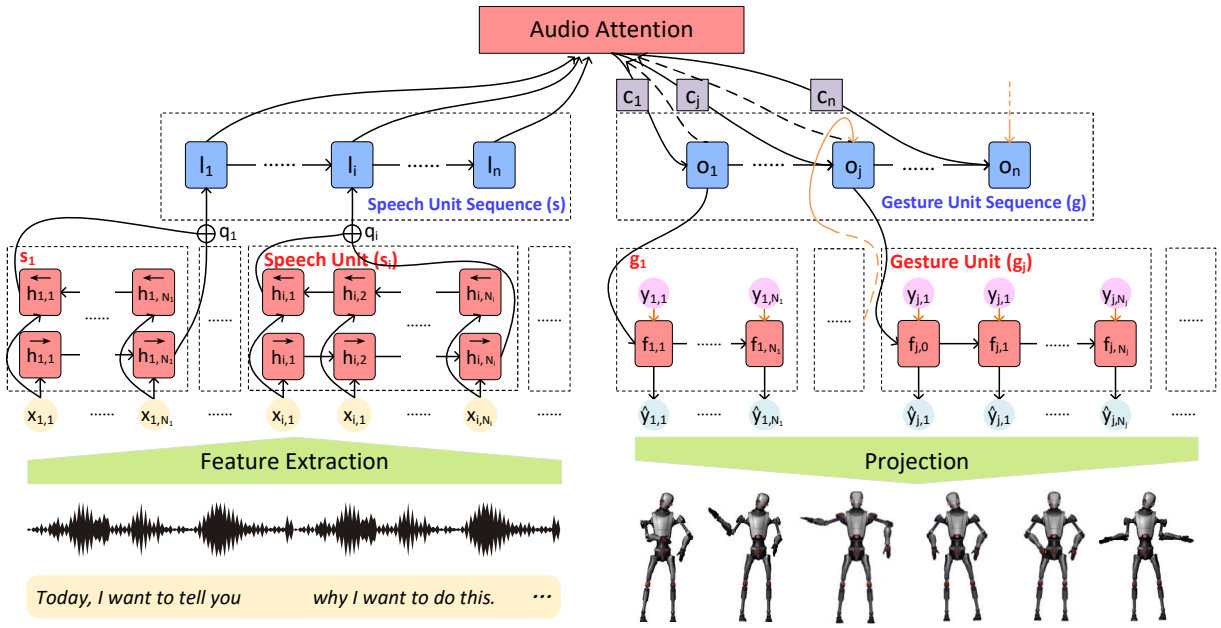


**Figure 3: Illustration of our proposed speech to gesture network (POSE) scheme.**

## 3.2 Long Short-Term Memory Unit

We chose to use Long Short-Term Memory (LSTM) unit as our non-linear transformation $f$. Let $\mathbf{x}_t$ denote the input vector, the specific parameterization of $f$ is given by:

$$\mathbf{i}_t = \sigma_g(\mathbf{W_i}\mathbf{x}_t + \mathbf{U_i}\mathbf{h}_{t-1} + \mathbf{b_i}), \quad (1)$$

$$\mathbf{f}_t = \sigma_g(\mathbf{W_f}\mathbf{x}_t + \mathbf{U_f}\mathbf{h}_{t-1} + \mathbf{b_f}), \quad (2)$$

$$\mathbf{o}_t = \sigma_g(\mathbf{W_o}\mathbf{x}_t + \mathbf{U_o}\mathbf{h}_{t-1} + \mathbf{b_o}), \quad (3)$$

$$\mathbf{c}_t = \mathbf{f}_t \cdot \mathbf{c}_{t-1} + \mathbf{i}_t \cdot \tanh(\mathbf{W_c}\mathbf{x}_t + \mathbf{U_c}\mathbf{h}_{t-1} + \mathbf{b_c}), \quad (4)$$

$$\mathbf{h}_t = \mathbf{o}_t \cdot \sigma_h(\mathbf{c}_t), \quad (5)$$

where $\sigma_g$ is the logistic sigmoid function, $\sigma_h$ is the tanh function, $\cdot$ represents the element-wise scalar product between vectors, $\mathbf{i}$,$\mathbf{f}$,$\mathbf{o}$, and $\mathbf{c}$ are respectively the *input*, *forget*, and *output* gate, and *cell* activation vectors, $\mathbf{W_i}$,$\mathbf{W_f}$,$\mathbf{W_o}$,$\mathbf{W_c} \in \mathbb{R}^{d_u \times |\mathbf{x}|}$, $\mathbf{U_i}$,$\mathbf{U_f}$,$\mathbf{U_o}$,$\mathbf{U_c} \in \mathbb{R}^{d_u \times d_u}$, and $\mathbf{b_i}$,$\mathbf{b_f}$,$\mathbf{b_o}$$\mathbf{b_c}$ are in $\mathbb{R}^{d_u}$, $d_u$ is the length of hidden state vector. The $\mathbf{W}$ matrices encode the input $\mathbf{x}_t$ while the $\mathbf{U}$ matrices specialize in retaining or forgetting the information in $\mathbf{h}_{t-1}$. Hereafter, this function will be noted as LSTM($\mathbf{h}_{t-1}, \mathbf{x}_t$).

## 3.3 Architecture

Our proposed speech to gesture network (POSE), as shown in Figure 3, is an attentive two-layer (unit-level and sequence-level) recurrent encoder-decoder network. The intuition of the hierarchy is that rather than learning to predict a single skeleton frame each time, our model predicts the skeleton frame sequence as a whole. In this way, the transition between different skeleton frames, and the transition between different gesture units are learnt jointly.

*3.3.1 Unit-Level Encoding.* A shortcoming of conventional RNN is that only historic context can be exploited. In our work where the whole speech unit is processed at once, it is helpful to exploit future context as well.

For the audio frame representation sequence $\mathbf{x}_i = \{x_{i,1:N_i}\}$, which are extracted from speech unit $\mathbf{s}_i$, we used Bidirectional LSTM (BiLSTM) [22] to encode the general and contextual speech information. In the training session, it first computes the forward and hidden sequence $\vec{\mathbf{h}}_i$ and $\mathbf{h}_i$ respectively, and then combines these two into a fixed-length vector $\mathbf{q}_i$:

$$\vec{\mathbf{h}}_{i,u} = \text{LSTM}_{enc}^{unit}(\vec{\mathbf{h}}_{i,u-1}, x_{i,u}) \quad (6)$$

$$\mathbf{h}_{i,u} = \text{LSTM}_{enc}^{unit}(\mathbf{h}_{i,u+1}, x_{i,u}) \quad (7)$$

$$\mathbf{q}_i = (\mathbf{h}_{i,1} + \vec{\mathbf{h}}_{i,N_i})/2, \quad (8)$$

where $u = 1, \ldots, N_i$, $\vec{\mathbf{h}}_{i,u}$ and $\mathbf{h}_{i,u}$ represent forward and backward hidden state at time u respectively, and $\vec{\mathbf{h}}_{i,0}$ and $\mathbf{h}_{i,N_i+1}$ are initialized with $\vec{\mathbf{1}}$.

In summary, the unit-level BiLSTM encoder maps speech unit's variable-length speech frames to a fixed-length vector. Its parameters are shared across the unit-Level encoding. Therefore, the obtained representation $\mathbf{q}_i$ is a general, contextual representation of speech unit $\mathbf{s}_i$.

*3.3.2 Sequence-Level Encoding.* The sequence-level RNN takes the output of the unit-level encoding $\mathbf{q}_i$ as the input, and computes the sequence of sequence-level recurrent states. For the sequence-level RNN, we also use the LSTM function:

$$\mathbf{l}_v = \text{LSTM}_{enc}^{seq}(\mathbf{l}_{v-1}, \mathbf{q}_v), \; v = 1, \ldots, N_v, \quad (9)$$

where $\mathbf{l}_v \in \mathbb{R}^{d_h}$ is the sequence-level recurrent state and $\mathbf{l}_0 = \vec{\mathbf{1}}$, $d_h$ is the length of hidden state vector. The same as unit-level encoder, the RNN shares parameters across the sequence-level encoding.

*3.3.3 Attention-based Sequence Alignment.* Co-speech gestures interact with speeches in terms of content and timing, and the interaction between gesture and speech influence the exchange of

information [15]. In order to model the interaction between speech and gesture, we introduced the attention mechanism when translate the speech modality into gesture modality. With the attention mechanism, our generated co-verbal gesture frames is not a simple one-to-one replacement of speech frames, the decoder "attend" to different parts of the source speech units with the help of the context vector.

Specifically, the context vector $\mathbf{c}_i$ depends on speech unit sequence level hidden state $\{\mathbf{l}_1, \ldots, \mathbf{l}_{|s|}\}$. Each $\mathbf{c}_i$ contains information about the whole speech unit sequence $s$ with a strong focus on the parts surrounding the $i$-th speech unit of the input sequence. It is computed as a weighted sum of these hidden states $\{\mathbf{l}_{1:|s|}\}$:

$$\mathbf{c}_i = \sum_j \alpha_{i,j} \mathbf{l}_j. \tag{10}$$

The weight $\alpha_{i,j}$ of each hidden state $\mathbf{l}_j$ is computed by:

$$\alpha_{i,j} = exp(e_{i,j}) / \sum_k exp(e_{i,k}). \tag{11}$$

$$e_{i,j} = \mathbf{w_a}^\top \tanh(\mathbf{W_a o}_{i-1} + \mathbf{U_a l}_j + \mathbf{b_a}) \tag{12}$$

is an alignment model which scores how well the inputs around position $j$ and the output at position $i$ match, and $\mathbf{w_a}$,$\mathbf{b_a} \in \mathbb{R}^{d_{att}}$, $\mathbf{W_a}, \mathbf{U_a} \in \mathbb{R}^{d_{att} \times d_h}$.

*3.3.4 Sequence-Level Decoding.* The sequence-level hidden state $\mathbf{o}_j$ is computed by:

$$\mathbf{o}_j = \text{LSTM}^{seq}_{dec}(\mathbf{o}_{j-1}, f_{j-1,N_{j-1}}, \mathbf{c}_j), \tag{13}$$

where $\mathbf{o}_{j-1}$ is the previous gesture unit hidden state, $f_{j-1,N_{j-1}}$ is the last hidden state of the unit-level decoder that takes $\mathbf{o}_{j-1}$ as input, $\mathbf{c}_j$ is the attentive context vector.

*3.3.5 Unit-Level Decoding.* We take $\mathbf{o}_j$ as the start vector of corresponding unit-level decoder, the decoder will predict the output frame on the condition of $\mathbf{o}_j$ and previous frames $f_{1:j-1}$. The probability is defined by:

$$P(f_{j,k}|y_{j,k}, \mathbf{o}_j) = \prod_{k \in [1,N_j]} (f_{j,k}|f_{j,1:k-1}, y_{j,k}, \mathbf{o}_j), \tag{14}$$

where $f_{j,k}$ and $y_{j,k}$ are respectively the prediction output and real output skeleton frame representation of the corresponding input speech frame representation $x_{j,k}$.

*3.3.6 Skeleton Frame Prediction.* For each prediction $f_{j,k}$, we project it back to the posture state representation vector, and apply softmax function on it to find out the probability of the posture state. The predicted skeleton frame is the posture state with the largest probability.

*3.3.7 Training.* The model needs to adjust the parameters of four LSTM functions $\text{BiLSTM}^{unit}_{enc}$, $\text{LSTM}^{seq}_{enc}$, $\text{LSTM}^{seq}_{dec}$, $\text{LSTM}^{unit}_{dec}$, a speech feature projection and a posture embedding matrices, and the parameters for attention mechanism. The model is trained to minimize the cross-entropy loss:

$$L = \frac{-1}{N_s} \sum_{i,j} y_{i,j} \log P(\hat{y}_{i,j}) + (1 - y_{i,j}) \log P(1 - \hat{y}_{i,j}), \tag{15}$$

where $N_s$ is the total number of training samples.

# 4 DATASET

TED is a media organization which posts videos of presentations online for free distribution. The rich multimodal data on TED website have been utilized in multiple aspects, e.g., video recommendation [18], acoustic modeling [16], machine translation [24], etc. However, none of the research work has utilized the TED talks to explore the presentation skills. Therefore, we crawled the **TED Speech-Gesture Presentation** dataset and gathered 2,582 unique videos published online from June 2006 to March 2018.

The videos uploaded on the TED website have been carefully edited (full of shot changes among shots like close-up shot, medium shot, audience shot, etc.), thus cannot be directly used to learn the body language of the speaker. We applied shot detection to the original video, and conducted face detection to select the segments with speaker shown in the scene. At last, we obtained 109,762 video segments after shot detection, and selected 35,658 useful video segments (in total 6,249 minutes) for our experiment.

# 5 EXPERIMENTS AND RESULTS

## 5.1 Data Preparation

*5.1.1 Text-based Structure.* Since the transcripts on TED website have the exact timing for each sentence, we further conducted text parsing to break the video into sentence-level components. Each sentence corresponds to the **sequence** in our work.

*5.1.2 Audio-based Tokenization.* Audio signal is modeled as being composed of a sequence of audio tokens, which correspond to the speech and gesture units in our proposed model. An audio signal tokenization tool auditok[1] is applied to cut the input audio signal into multiple tokens, which corresponds to the speech and gesture **unit** in our proposed model.

*5.1.3 Skeleton Frame Extraction.* A full-body pose estimator OpenPose [3] is used at 5 fps to provide the joints of a speaker's skeleton[2]. Because most of the videos contain only the upper body of the speaker, we select 8 key joints *(Nose, Neck, LShoulder, RShoulder, LElbow, RElbow, LWrist, RWrist)* to characterize speaker's gesture (as shown in Figure 2).

Furthermore, because of the difficulty in the task of tracking skeleton of the speaker, some joints are lost (the coordinate become zeros) during the tracking. We replace these outliers by smoothly interpolating the coordinates with the historical tracking data of these lost joints. At last, our dataset contains 1,874,755 skeleton frames, 306,734 gesture units, and 106,053 Gesture Unit Sequences.

## 5.2 Baselines

**Rand** is a simple alternative method for synthesizing body language in real time. It has previously been used to animate facial expressions [19] and to add details to coarsely defined motion [14]. We utilize the uniform distribution to generate a 5-fps posture state sequence for each speech audio sequence.

**Seq2Seq** network [23] is an effective tool to model a generative and temporal problem. We built an LSTM-based Seq2Seq network with one hidden layer (with 128 hidden states). The input of the network

[1] https://github.com/amsehili/auditok

[2] These joints are: Nose, Neck, RShoulder, RElbow, RWrist, RWrist, LShoulder, LElbow, LWrist, RHip, RKnee, RAnkle, LHip, LKnee, LAnkle, REye, LEye, REar, LEar.

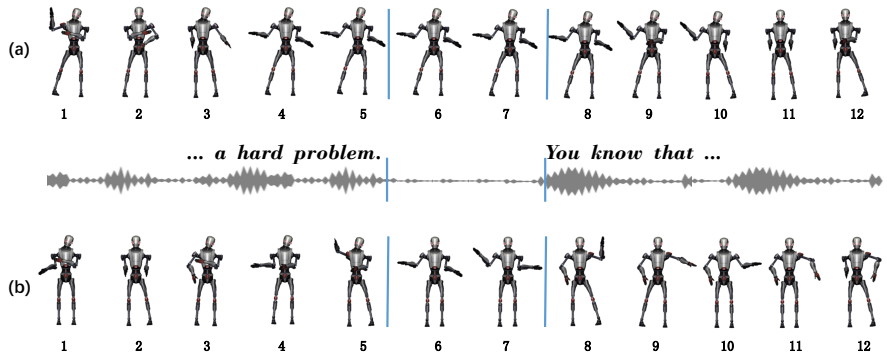


**Figure 4: An example of three speech-gesture units ([1:5;6,7;8:12]) of *"...a hard problem. You know that..."*. Sequence (a) and (b) show the visualization of posture sequence generated by $\text{POSE}_{Att}$ and $\text{POSE}_{Att}^{w/o}$, respectively.**

**Table 1: Results of the Objective Evaluation**

| Method | $S_{QoM}$ | $S_{v_{hand}}$ |
|---|---|---|
| RAND | 0.6151 | 0.6941 |
| Seq2Seq | 0.7127 | 0.8920 |
| $\text{POSE}_{Att}$ | **0.7976** | 0.8917 |
| $\text{POSE}_{Att}^{w/o}$ | 0.7720 | **0.9029** |

are sequences of audio frame feature (without the hierarchical structure). Similar with other baselines, the network will output a sequence of posture states with 5 fps.

$\textbf{POSE}_{\textbf{Att}}^{\textbf{w/o}}$ is our proposed POSE without attention module, as a comparison method to validate the effectiveness of the attention mechanism on gesture generation. We denote our original attention-based two-layer recurrent encoder-decoder network as $\textbf{POSE}_{\textbf{Att}}$.

## 5.3 Model Configuration

The model exploits stochastic gradient descent algorithm for optimization. Dataset is divided into 8:1:1 for training, testing, and validation. Dropout (with dropout rate of 0.5) is applied on hidden layers to avoid overfitting.

## 5.4 Evaluation

*5.4.1 Objective Evaluation.* The objective measurement consists of three main aspects of gesture quality: *gesture quantity*, *gesture speed*, and *gesture fluidity*. We utilized Quantity of Motion (QoM) [4] over $n_{\text{QoM}} = 5$ skeleton frames to measure the amount of detected motion to characterize gesture quantity and gesture speed. The QoM is computed with a technique based on Silhouette Motion Images (SMIs). Specifically, the SMI at frame t is generated by adding together the silhouettes extracted in the previous $n_{\text{QoM}}$ frames and then subtracting the silhouette at frame t:

$$\text{SMI}[t] = \{ \sum_{i=0}^{n_{\text{QoM}}} \text{Silhouette}[t-i] \} - \text{Silhouette}[t]. \quad (16)$$

QoM is computed as the area (i.e., number of pixels) of a SMI, normalized by the area of silhouette:

$$\text{QoM} = Area(\text{SMI}[t,n])/Area(\text{Silhouette}[t]) \quad (17)$$

For the fluidity of gesture, we calculated the speed of hand's barycentre in the 2D plane:

$$v_{\text{hand}} = ||\dot{P_{\text{lh}}}(\psi)|| + ||\dot{P_{\text{rh}}}(\psi)||, \quad (18)$$

where $||P_{\text{lh}}(\psi)||$ and $||P_{\text{rh}}(\psi)||$ denote the position of left and right hands, respectively.

We calculated the QoM and $v_{hand}$ for all the methods and the original skeleton tracking sequence (denoted as GT) over the whole dataset. To evaluate different methods, we estimated and quantified the similarity ($S$) of the distribution for QoM and $v_{hand}$. We adopted the similarity measure [9] which uses the first-order statistics to explain the similarity $S$ between the distribution of the results with the GT's distribution. The higher the value of $S$ is, the more similar the two distributions will be.

Table 1 shows the objective evaluation results. We can see from the table that RAND performs the worst among all these methods. This is understandable because the random selection method does not consider the transition between consecutive skeleton frames, which may lead to an unnatural constant-speed transition between different postures. As for the gesture quantity ($S_{QoM}$), Seq2Seq performs worse than our POSE w/ or w/o attention mechanism. This better performance may due to the hierarchical structure in our proposed POSE, which encodes the transition between skeleton frames and between gesture units. As for the gesture fluidity ($S_{v_{hand}}$), our proposed POSE performs similar with Seq2Seq. $\text{POSE}_{Att}^{w/o}$ has even a slightly better performance than $\text{POSE}_{Att}$ for gesture fluidity, which may due to the sacrifice for the attention-based sequence alignment.

*5.4.2 Subjective Evaluation.* We also performed a user study and asked 40 subjects (20 females and 20 males) to assess the quality of generated gestures. We selected 5 pieces of speech (with 15s in average, covering speaker of both men and women ), and generated the gesture sequences with RAND, Seq2Seq, $\text{POSE}_{Att}$, $\text{POSE}_{Att}^{w/o}$, and GT. We presented all these 25 pieces of video to the subjects, and used the same evaluation metric in [20] (for communicative robot gesture evaluation) to ask the subjects to give a rating from 1 to 5 for the measurement of *gesture quantity*, *gesture speed*, *gesture fluidity*, *speech-gesture content*, *speech-gesture timing*, and *naturalness*.

We observed that generally RAND showed continuous random movements of the skeleton, while Seq2Seq showed similar behavior

**Table 2: Results of the Subjective Evaluation**

| Measure | Scale | RAND | Seq2Seq | POSE$_{Att}$ | POSE$^{w/o}_{Att}$ | GT |
|---|---|---|---|---|---|---|
| M1: Gesture Quantity | 1 = too few, 5 = too many | 3.95 | 3.40 | 2.55 | 2.75 | 1.88 |
| M2: Gesture Speed | 1 = too slow, 5 = too fast | 4.03 | 3.53 | 3.03 | 3.30 | 2.10 |
| M3: Gesture Fluidity | 1 = not appropriate, 5 = very appropriate | 2.93 | 4.03 | 4.20 | 4.03 | 4.25 |
| M4: Speech-Gesture Content | 1 = not appropriate, 5 = very appropriate | 1.53 | 1.98 | 4.38 | 4.13 | 4.33 |
| M5: Speech-Gesture Timing | 1 = not appropriate, 5 = very appropriate | 2.18 | 2.45 | 4.13 | 3.85 | 4.25 |
| M6: Naturalness | 1 = artificial, 5 = natural | 1.98 | 2.58 | 4.38 | 3.33 | 4.20 |

but with smoother transition between different postures. POSE$_{Att}$ and POSE$^{w/o}_{Att}$ showed clear grouping effect where one movement pattern is followed by another. The GT, showed least movement in the visualization. This is mainly because of the discretization of posture state space, which eliminates the slight motion movement.

The user study results shown in Table 2 validate our observations with the visualization: RAND is considered as having the least amount of gestures with the fastest speed, and Seq2Seq performs slightly better than RAND. Our proposed POSE$_{Att}$ and POSE$^{w/o}_{Att}$ perform similarly in gesture quantity and speed, with GT is slightly slower and contains fewer gestures.

We note that our proposed POSE$_{Att}$ outperforms POSE$^{w/o}_{Att}$ in terms of speech-gesture content, timing, and naturalness. After further investigation, we found out that in general POSE$_{Att}$ contains more "silent" gesture than POSE$^{w/o}_{Att}$. Figure 4 shows a concrete example of this difference. It is an example of three speech-gesture units ([1:5;6,7;8:12]) inside a sequence *"... a hard problem. You know that..."*. For the silent speech unit ([6,7]), POSE$_{Att}$ generates more identical postures (frame 4 to 8 in sequence a) compared to POSE$^{w/o}_{Att}$. Also, the boundary of the identical postures is not exactly the same with the unit boundary. This will lead to a certain "delay" effect which will improve the quality of speech-gesture interaction, and ultimately naturalness.

## 6 CONCLUSION AND FUTURE WORK

In this paper, we present a gesture generation model which translates speech audio signal into sequence of gestures by an attentive two-layer RNN. Both objective and subjective evaluation show the effectiveness of our proposed method. In addition, TED Speech-Gesture Presentation dataset is established, which can benefit future researchers in this domain. In the future, we would like to extend our body language to facial expression, and eye contact, which are also important non-verbal behaviors during communication. Besides, we will further investigate novel methods for gesture and speech coordination.

## 7 ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China, No.: 61772310, No.: 61702300, No.: 61702302, No.:61802231, No.:U1836216; the Project of Thousand Youth Talents 2016; the Shandong Provincial Natural Science and Foundation, No.: ZR2019JQ23, No.: ZR2019QF001; the Future Talents Research Funds of Shandong University, No.: 2018WLJH63; the Fundamental Research Funds of Shandong University (No. 2017HW001).

## REFERENCES

[1] Paul Boersma. 2002. Praat, a system for doing phonetics by computer. *Glot international* 5 (2002), 341–345.
[2] Elif Bozkurt, Yücel Yemez, and Engin Erzin. 2016. Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. *Speech Communication* 85 (2016), 29–42.
[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *IEEE CVPR*. 1302–1310.
[4] Ginevra Castellano, Santiago D. Villalba, and Antonio Camurri. 2007. Recognising Human Emotions from Body Movement and Gesture Dynamics. In *International Conference on Affective Computing and Intelligent Interaction*. 71–82.
[5] Lei Chen, Chee Wee Leong, Gary Feng, and Chong Min Lee. 2014. Using Multimodal Cues to Analyze MLA'14 Oral Presentation Quality Corpus: Presentation Delivery and Slides Quality. In *ACM MLA Workshop*. 45–52.
[6] Norah E Dunbar, Catherine F Brooks, and Tara Kubicka-Miller. 2006. Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education* 31, 2 (2006), 115.
[7] Tian Gan, Yongkang Wong, Bappaditya Mandal, Vijay Chandrasekhar, and Mohan S. Kankanhalli. 2015. Multi-sensor Self-Quantification of Presentations. In *ACM MM*. 601–610.
[8] Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences* (1999), 419–429.
[9] Luis Gonzalez-Abril, Jose M Gavilan, and Francisco Velasco Morente. 2014. Three Similarity Measures between One-Dimensional DataSets. *Revista Colombiana de Estadística* 37, 1 (2014), 79–94.
[10] Autumn B Hostetter. 2011. When do gestures communicate? A meta-analysis. *Psychological bulletin & review* 137, 2 (2011), 297.
[11] Autumn B Hostetter and Martha W Alibali. 2008. Visible embodiment: Gestures as simulated action. *Psychonomic bulletin & review* 15, 3 (2008), 495–514.
[12] Adam Kendon. 1986. Current issues in the study of gesture. *The biological foundations of gestures: Motor and semiotic aspects* 1 (1986), 23–47.
[13] Junnan Li, Yongkang Wong, and Mohan S. Kankanhalli. 2016. Multi-stream Deep Learning Framework for Automated Presentation Assessment. In *IEEE ISM*. 222–225.
[14] Yan Li, Tianshu Wang, and Heung-Yeung Shum. 2002. Motion texture: a two-level statistical model for character motion synthesis. *ACM ToG*. 21, 3 (2002), 465–472.
[15] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago Press.
[16] Yajie Miao, Hao Zhang, and Florian Metze. 2015. Speaker Adaptive Training of Deep Neural Network Acoustic Models Using I-Vectors. *IEEE/ACM ToASLP* 23, 11 (2015), 1938–1949. https://doi.org/10.1109/TASLP.2015.2457612
[17] Izidor Mlakar, Zdravko Kačič, and Matej Rojc. 2013. TTS-driven Synthetic Behaviour-generation Model for Artificial Bodies. *International Journal of Advanced Robotic Systems* 10, 10 (2013), 344.
[18] Nikolaos Pappas and Andrei Popescu-Belis. 2013. Sentiment analysis of user comments for one-class collaborative filtering over ted talks. In *SIGIR*. 773–776.
[19] Ken Perlin. 1997. Layered compositing of facial expression. In *ACM SIGGRAPH*. 226–227.
[20] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina J. Rohlfing, and Frank Joublin. 2012. Generation and Evaluation of Communicative Robot Gesture. *International Journal of Social Robotics* 4, 2 (2012), 201–217.
[21] Mehmet Emre Sargin, Oya Aran, Alexey Karpov, Ferda Ofli, Yelena Yasinnik, Stephen Wilson, Engin Erzin, Yücel Yemez, and A. Murat Tekalp. 2006. Combined Gesture-Speech Analysis and Speech Driven Gesture Synthesis. In *IEEE ICME*. 893–896.
[22] Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE ToSP* 45, 11 (1997), 2673–2681.
[23] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *NIPS*. 3104–3112.
[24] Hainan Xu and Philipp Koehn. 2017. Zipporah: a Fast and Scalable Data Cleaning System for Noisy Web-Crawled Parallel Corpora. In *EMNLP*. 2945–2950.