Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network

Conference Paper · November 2018

DOI: 10.1145/3267851.3267878

5 authors, including:

Dai Hasegawa
Hokkai-Gakuen University
52 PUBLICATIONS **393** CITATIONS

SEE PROFILE

Hiroshi Sakuta
National Institute for Materials Science
111 PUBLICATIONS **276** CITATIONS

SEE PROFILE

Kazuhiko Sumi
Aoyama Gakuin University
116 PUBLICATIONS **1,486** CITATIONS

SEE PROFILE

Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network

Dai Hasegawa
Hokkai Gakuen University
Sapporo, Japan
dhasegawa@hgu.jp

Naoshi Kaneko
Aoyama Gakuin University, Japan
Sagamihara, Japan
kaneko@it.aoyama.ac.jp

Shinichi Shirakawa
Yokohama National University
Yokohama, Japan
shirakawa-shinichi-bg@ynu.ac.jp

Hiroshi Sakuta
Aoyama Gakuin University, Japan
Sagamihara, Japan
sakuta@it.aoyama.ac.jp

Kazuhiko Sumi
Aoyama Gakuin University, Japan
Sagamihara, Japan
sumi@it.aoyama.ac.jp

ABSTRACT

We present a novel framework to automatically generate natural gesture motions accompanying speech from audio utterances. Based on a Bi-Directional LSTM Network, our deep network learns speech-gesture relationships with both backward and forward consistencies over a long period of time. Our network regresses a full 3D skeletal pose of a human from perceptual features extracted from the input audio in each time step. Then, we apply combined temporal filters to smooth out the generated pose sequences. We utilize a speech-gesture dataset recorded with a headset and marker-based motion capture to train our network. We validated our approach with a subjective evaluation and compared it against "original" human gestures and "mismatched" human gestures taken from a different utterance. The evaluation result shows that our generated gestures are significantly better than the "mismatched" gestures with respect to time consistency. The generated gesture also shows marginally significant improvement in terms of semantic consistency when compared to "mismatched" gestures.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Discourse, dialogue and pragmatics**; **Neural networks**;

KEYWORDS

gesture generation, deep learning, neural networks, long short-term memory

ACM Reference Format:
Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network. In *IVA '18: International Conference on Intelligent Virtual Agents (IVA '18), November 5–8, 2018, Sydney, NSW, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3267851.3267878

IVA '18, November 5–8, 2018, Sydney, NSW, Australia

ACM ISBN 978-1-4503-6013-5/18/11.
https://doi.org/10.1145/3267851.3267878

1 INTRODUCTION

In human-human communication, gestures (defined as *gesticulation* [18]) play an important role, such as complementing or emphasizing speech. And researchers have been introducing non-verbal expressions, including gesture, into computer systems with human-like appearances, called Embodied Conversational Agents (ECAs) [4]. It has also repeatedly been verified that gestures performed by ECAs or robots have positive effects in various applications [2, 12].

However, implementing meaningful gestures along with speech into ECAs costs time and effort. For example, one method requires animators to design motions themselves and apply them to a 3D model by hand with modeling software. There is also another way, which is working with motion capture systems. However, this requires expensive facilities and technical experts with the knowledge and skills needed to use the systems.

To date, several studies have been conducted to automatically generate gestures from speech or text. Early works tackled this problem using a rule-based approach [5, 6]. The rule-based approach has the advantage in that if we can prepare enough knowledge to represent a task and domain, the system will perform well. However, preparation takes a lot of effort. Thus, the domain is highly restricted.

To avoid this difficulty of the rule-based approach, a data-driven approach with machine learning has also been proposed [7, 8]. In the data-driven approach, we do not have to prepare elaborate domain knowledge. Rather, the "knowledge" is automatically acquired in the learning process, and this makes the data-driven approach be applicable to wider domains than the rule-based approach. However, generating gesture motions which are perfectly consistent with speech content still remains as a challenging problem. In [8], the authors solved the gesture generation problem as a classification task, thus the method still requires predefined gesture categories and handmade motion data. On the other hand, the work of [7] succeeded to generate the motions of *beat* gestures. However, this approach only used pitch and sound pressure of speech as inputs, so it is difficult to distinguish phonemes. Naturally, this leads to difficulty in speech recognition. Since the semantic content of speech is highly correlated to gestures, the difficulty in speech recognition causes a negative influence on the generated gestures.

In this paper, we take a deep neural network approach for automatic three dimensional (3D) gesture motion generation and also

use Mel-Frequency Cepstral Coefficients (MFCC) [9], which have been used for speech recognition. Thus, we expect that the network has the possibility to hold richer prosodic expression. In addition, we use Bi-Directional Recurrent Neural Networks (Bi-Directional RNN) [15, 24] with Bi-Directional Long Short-Term Memory (Bi-Directional LSTM) [13] as recurrent units, which have often been used in natural language processing. They can be trained in both the backward and forward direction over a long period of time, and match the bi-directional characteristics of language.

2 STATE OF THE ART

Cassell et al. [6] created the Behavior Expression Animation Toolkit (BEAT), which takes text as input to generate synthesized speech along with gestures and other nonverbal behaviors such as gaze and facial expression. The assignment is performed on the linguistic and contextual analysis of the input text, relying on rules predefined based on evidence from previous research on human conventional behavior.

Cassell et al. [5] later implemented a system in which *iconic* gestures for an embodied conversational agent were generated from text instructions for giving directions. The gestures to be generated were selected according to predefined rules, using abstract semantic representations of words such as size, characteristics, width, and height, which were attached beforehand.

In such rule-based approaches, relation of speech content and gesture can be described clearly by rules. Therefore, the process can produce gestures which are consistent with speech. However, it can only produce gestures prepared beforehand, and creating rules is very costly.

On the other hand, there is a data-driven approach that can produce gestures from audio prosodic features. Chiu et al. [8] used deep learning methods with the proposed Deep Conditional Neural Field (DCNF) model to predict gesture types from textual verbal content and prosodic features of a spoken utterance, defining fourteen kinds of gestural signs based on previous literature on gestures. The authors used six prosodic features, such as frequency of audio and Normalized Amplitude Quotient (NAQ) [1], transcriptions of speech, and parts of speech. The results of the research are promising, although the appearance of gestures is limited to the ones defined beforehand. And the output is not 3D motion, but one of the fourteen gesture categories. Thus, it has to be converted to a positional motion sequence when applied to ECAs.

Chiu et al. [7] also proposed a method to produce a 3D gesture motion sequence from audio input. This used the prosodic features, pitch and level, of speech together with motion data to train a gesture generator built based on Hierarchical Factored Conditional Restricted Boltzmann Machines (HFCRBM). These two prosodic features do not account for the actual semantic content of the speech, and thus the generator is designed to learn *beat* gestures, which are rhythmic motions usually related to prosody.

This approach is similar to ours. But we use Mel-Frequency Cepstral Coefficients (MFCC) [9] as the audio feature. Thus, we can expect that the network has the possibility to hold richer prosodic expression. In addition, we use Bi-Directional Recurrent Neural Networks (Bi-Directional RNN) [15, 24] with Bi-Directional Long Short-Term Memory (Bi-Directional LSTM) [13] as recurrent units, which have often been used in natural language processing. The system can be trained in both the backward and forward direction over a long period of time, and match the bi-directional characteristics of language.

3 OUTLINE OF PROPOSED METHOD

We propose a method to automatically generate gesture motions from speech audio. In this method, we will try to build a model to represent the relationship between speech content and gestures accompanying the speech by using a Bi-Directional LSTM Network which can take into account both backward and forward consistencies over a long period of time.

Figure 1 shows an outline of our proposed method. As shown, first, a speech audio of one sentence (.wav) is converted to MFCC [9] feature vectors for each time window. MFCC is a widespread audio feature designed for speech recognition, taking into account human perceptual tendency [11, 20]. Therefore, we believe that the MFCC feature can hold sufficient information for language.

Next, the MFCC feature vector is fed into a Bi-Directional LSTM Network which has five layers. Then, 3D positions of entire body joints are predicted by regression of the network. LSTM can hold previous inputs for a relatively long duration. Hence, it should be effective for specific data modeling such as a gesture which is related to a whole sentence or sometime beyond a sentence. Figure 2 shows the LSTM internal architecture.

The network is trained by a speech and motion paired dataset [26] collected in a semi organized interview style [10]. In the interview, the speaker's motion is recorded by an optical motion capture system.

Lastly, a smoothing process is applied to the network output by temporal filtering, because the output joint positions have small discontinuities between frames, originating from noise and prediction errors. And finally, we obtain time-series joint positions of a whole body from speech.

4 GESTURE GENERATION METHOD

In this section we will explain the proposed gesture generation method in detail. First, a speech audio input is divided into small segments of audio data of a certain duration. Then the divided audio segments are converted to MFCC features. Thus, we will get T length time-series MFCC feature vectors $\boldsymbol{X} = \{\boldsymbol{x}_t\},\ t = 1, ..., T$. Next, $\boldsymbol{x}_t$ is fed into the network along with MFCC feature vectors of S steps backward and forward. Thus, the actual input at time t will be the following: $\{\boldsymbol{x}_{t-S}, ..., \boldsymbol{x}_{t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}, ..., \boldsymbol{x}_{t+S}\}$. Also, if there is no audio data before/after a certain time step, we attach MFCC features of silent audio data. According to the growth point theory by McNeill [21], gesture is simultaneously generated with speech by interacting with each other, and they share its origin, a meaningful mental unit. Hence, we believe that when we decide on gestures, we need to hold information for a whole sentence or a phrase. We expect the Bi-Directional LSTM units to provide such functionality. In this paper, the time step t was 0.05 seconds (20 steps per second), the number of context steps S was 30 (1.5 seconds backward and forward, which is a 3.0 second time window), the MFCC time window was 0.125 seconds, and the MFCC feature had 26 dimensions.

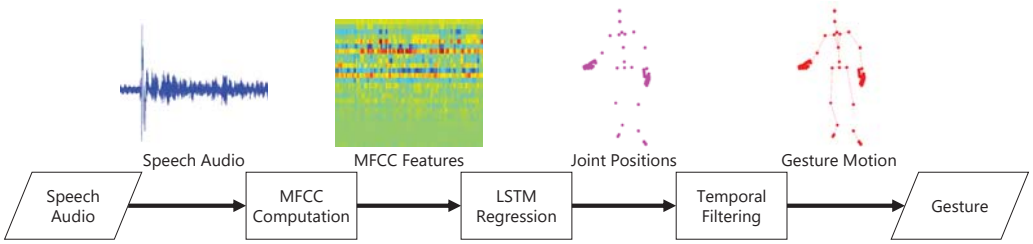


Figure 1: Outline of Proposed Method.

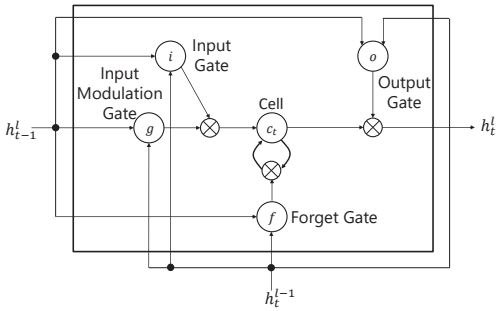


Figure 2: Internal Architecture of LSTM.

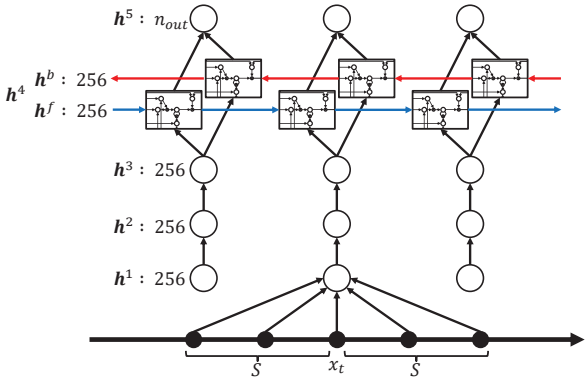


Figure 3: Bi-Directional LSTM Network Archtecture.

By feeding the MFCC feature vectors $\{\boldsymbol{x}_{t-S}, ..., \boldsymbol{x}_{t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{t+1}, ..., \boldsymbol{x}_{t+S}\}$ into the Bi-Directional LSTM Network, that outputs 3D joint positions $\boldsymbol{P}_t = \{\boldsymbol{p}_t^i\}$, $i = 1, ..., K$, where K is the number of joints of a whole body which we will predict. In our dataset [26], K was 64.

4.1 Network Architecture

Figure 3 shows the Bi-Directional LSTM Network architecture we used. The network consisted of five layers. Layers $\boldsymbol{h}^1 - \boldsymbol{h}^3$ and $\boldsymbol{h}^5$ were fully connected layers which are not recurrent and $\boldsymbol{h}^4$ was a Bi-Directional LSTM layer. In Figure 3, numbers on the left side show the number of feature dimensions for each layer. The number of output dimension n_{out} differs depending on the loss functions. We will explain about this later in Section 4.2. The network design was based on Deep Speech [15], which achieved one of the most successful speech recognition attempts.

$\boldsymbol{h}^1$, as described above, takes an MFCC feature vector $\boldsymbol{x}_t$ of current step t along with context MFCC feature vectors of S steps backward and forward. $\boldsymbol{h}^2$ and $\boldsymbol{h}^3$ take outputs from previous layers. Therefore, $\boldsymbol{h}^1 - \boldsymbol{h}^3$ of step t is calculated as

$$\boldsymbol{h}_t^l = g(\boldsymbol{W}^l \boldsymbol{h}_t^{l-1} + \boldsymbol{b}^l), \tag{1}$$

where $\boldsymbol{W}^l$ and $\boldsymbol{b}^l$ are the weight and bias parameters of $\boldsymbol{h}^l$ respectively, and g is an activation by Rectified Linear Unit (ReLU) [22].

$\boldsymbol{h}^4$ is a Bi-Directional LSTM layer which consists of forward LSTM units $\boldsymbol{h}^f$ calculating from $t = 1$ to $t = T$ and backward LSTM units $\boldsymbol{h}^b$ calculating backwards from step $t = T$ to $t = 1$ as shown below (where $\mathcal{M}$ is the calculation of the LSTM unit).

$$\boldsymbol{h}_t^f = g(\mathcal{M}(\boldsymbol{W}^4 \boldsymbol{h}_t^3 + \boldsymbol{W}^f \boldsymbol{h}_{t-1}^f + \boldsymbol{b}^4)), \tag{2}$$

$$\boldsymbol{h}_t^b = g(\mathcal{M}(\boldsymbol{W}^4 \boldsymbol{h}_t^3 + \boldsymbol{W}^b \boldsymbol{h}_{t+1}^b + \boldsymbol{b}^4)). \tag{3}$$

Lastly, the last layer $\boldsymbol{h}^5$ is calculated as below.

$$\boldsymbol{h}_t^5 = \boldsymbol{W}^5(\{\boldsymbol{h}_t^f, \boldsymbol{h}_t^b\}) + \boldsymbol{b}^5, \tag{4}$$

where the output of $\boldsymbol{h}^5$ are 3D joint positions. Therefore, no activation is applied. In addition, to avoid overfitting and to make learning stable, Batch Normalization [16] and 10% Dropout [25] are applied to $\boldsymbol{h}^1 - \boldsymbol{h}^4$.

4.2 Training

The network was trained with a dataset created by a semi-structured interview style where speech audio and motion captured data are paired [26]. We used mean squared error for the loss function. The loss function L was defined as

$$L(\boldsymbol{v}, \boldsymbol{y}) = \frac{1}{T \times K} \sum_{t=1}^{T} \sum_{i=1}^{K} \|\boldsymbol{y}_t^i - \boldsymbol{v}_t^i\|^2, \tag{5}$$

where $\boldsymbol{y}$ was the output of the Bi-Directional LSTM Network and $\boldsymbol{v}$ was the Ground Truth.

When calculating the loss, it is easy to use 3D positions. However, human posture at step t highly depends on the posture at the previous step $t - 1$. Thus, we use velocity as a additional feature, which is the difference between positions at step t and step $t - 1$. That is, $\boldsymbol{y}_t$ will be $\boldsymbol{y}_t = \{\boldsymbol{P}_t, \boldsymbol{V}_t\}$ where $\boldsymbol{P}_t$ is a 3D position of a joint and $\boldsymbol{V}_t$ is its 3D velocity. $\boldsymbol{v}$ will also become $n_{out} = 6 \times K$.

Table 1: Gestures in Dataset [26].

Type of Gesture	Total Number	Percentage
Iconic	202	4.76
Metaphoric	2,906	68.41
Deictic	132	3.11
Beat	1,008	23.73

We used Adam [19] for an optimization algorithm where $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and the batch size was 2,056. We trained the network for 500 epochs which was enough to converge the learning.

4.3 Smoothing by Temporal Filter

The output of the network has small discontinuities between frames originating from noise in the input data or prediction errors. It is very disturbing for humans to see someone gesturing with such discontinuities. Therefore, we will address the problem in post processing by using two kinds of temporal filtering. One kind is a 1€ filter [3], and the other is a Moving Average (MA) filter.

The 1€ filter can eliminate small shimmering when motion speed is not fast by filtering out low frequencies. As the motion speed accelerates, the filter makes the cutoff frequency higher. Thus, the filter can preserve a human like high speed movement at the same time. However, the 1€ filter does not filter relatively large discontinuities. On the other hand, the MA filter will smooth this with an averaging window which averages the values over a certain range of steps.

In our method, we first filter instances of small shimmering in the network's output with a 1€ Filter, then we use an MA filter to reduce larger discontinuities. The filter parameters were experimentally decided and we used $f_{c_{min}} = 0.1$ and $\beta = 0.08$ for the 1€ filter, and five steps (0.25 seconds) for the MA filter's averaging window.

5 QUANTITATIVE EVALUATION

5.1 Dataset and Experiment

We conducted an experiment with a dataset created by Takeuchi et al. [26]. The dataset consists of mp3-encoded speech audio content paired with motion captured motion data represented in Biovision Hierarchy (BVH) format. Each speech audio was divided into single sentences.

The speech audio and motion were recorded by using a headset microphone and an optical motion capture system built with eight OptiTrack Prime 41 cameras. The gestures in the dataset were counted based on a categorization scheme by McNeill [21]. The results are shown in Table 1. The most frequently appearing gesture category was *metaphoric*, and the second was *beat*.

The dataset has 1,049 sentences (196 minutes of speech in total). In our experiment, we used 767 sentences for training, 192 sentences for validation, and 90 sentences for testing. To avoid overfitting, we also performed data augmentation to double the training data by compounding white noise. The Euler joints in BVH data were converted to joint positions where the origin was located at the hip.

Thus, the relationship between speech audio and the 64 joint positions were modeled by a total of 1,534 training sentences.

Table 2: Prediction Accuracy of Proposed Method.

Method	APE [cm]
Position	8.59
+ Acceleration	8.44
+ Velocity	8.14
+ Velocity + 1€	8.08
+ Velocity + MA	8.08
+ Velocity + 1€ + MA (Proposed)	**8.03**

5.2 Results

We use Average Position Error (APE) as the evaluation measure. APE compares the predicted positions with positions that originally accompanied the speech, and it calculates Euclidean distance as described below.

$$APE = \frac{1}{T \times K} \sum_{t=1}^{T} \sum_{i=1}^{K} \|\boldsymbol{y}_t^i - \boldsymbol{v}_t^i\|_2. \quad (6)$$

Table 2 shows the APE for gesture prediction. As shown, the APE of the proposed architecture was 8.03. This meant for each joint there was 8.03 cm difference in an average of 64 joints. The proposed method showed a lower APE compared with the case of using only position for the loss function ("Position" in Table 2), the case of using acceleration instead of velocity ("+ Acceleration" in Table 2), and the case of using only one of the temporal filtering methods ("+ Velocity + 1€" and "+ Velocity + MA" in Table 2).

To see the prediction accuracy of each joint, we show APEs for 14 representative joints which were also used in other posture analyses [17, 27] in Figure 4. The result was, regarding gesture, right and left wrists had larger errors.

6 USER STUDY

When a neural network or statistical method developed in Artificial Intelligence is applied to Human-Computer Interaction related tasks, quantitative measurement is not sufficient for evaluation [14]. It requires user study. We conducted an empirical study to see how generated gestures were perceived by human participants. We explain the evaluation below.

6.1 Method

We conducted a 1×3 factorial within-participant design experiment controlling a gesture factor: original vs. predicted vs. mismatched. The original gestures were the gestures that originally accompanied the speech, the predicted gestures were the gestures produced by the proposed method, and the mismatched gestures were chosen from the dataset, so they were actually performed by a human but they accompanied different speech. The motion for the mismatched condition was selected from longer instances than the original speech and was cut at the end to match in length.

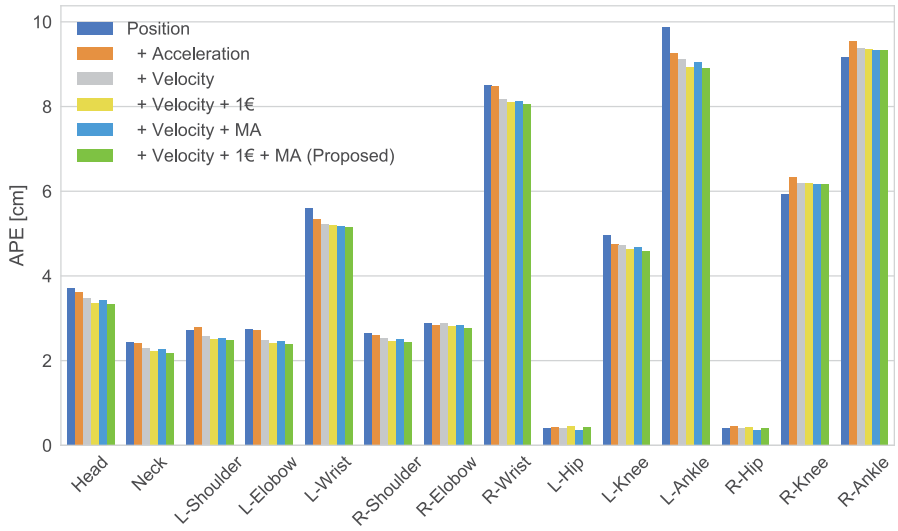


Figure 4: APEs of 14 Body Joints.

Table 3: Evaluation Scales and Items of Questionnaire.

Scale	Item (translated from Japanese)	Cronbach's α
Naturalness	Gesture was natural.	0.93
	Gesture was smooth.	
	Gesture was comfortable.	
Time Consistency	Gesture timing was matched to speech.	0.93
	Gesture speed was matched to speech.	
	Gesture pace was matched to speech.	
Semantic Consistency	Gesture was matched to speech content.	0.95
	Gesture well described speech content.	
	Gesture helped me to understand speech content.	

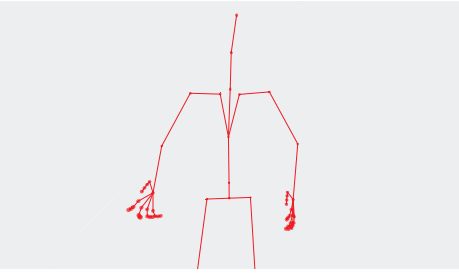


Figure 5: An Example of Gesture Presentation.

The speeches we used in the evaluation were 10 speeches selected from the test dataset of 90 sentences based on the criterion gesture per second (g/s). g/s represents the occurrence rate of gesture, that is the number of gestures that appeared per second. We calculated g/s for all 90 speeches and selected ten as follows.

- The three highest g/s speeches (1.58 g/s, 0.68 g/s, 0.68 g/s).
- Four speeches from the middle g/s range (all 0.39 g/s).
- The three lowest g/s speeches (0.19 g/s, 0.18 g/s, 0.16 g/s).

We used a questionnaire for measurement as shown in Table 3. The measured scales were naturalness, time consistency, and semantic consistency. Each scale was measured by three items with a seven level Likert scale anchored from 1: strongly disagree to 7: strongly agree.

The participants were asked to access a web page built for the experiment, view a single video of synchronized speech and gesture, and then answer the questionnaire. The order of speech was fixed, but the gesture conditions (original vs. predicted. vs. mismatched) were counter-balanced. Figure 5 shows an example video presentation. There were ten speeches × three kinds of gesture which resulted in thirty videos total.

6.2 Results

Thirty undergraduate participants (twenty five males, five females) were recruited by email announcements and took part in the study. For the questionnaire items, Cronbach's αs were all above 0.9 as

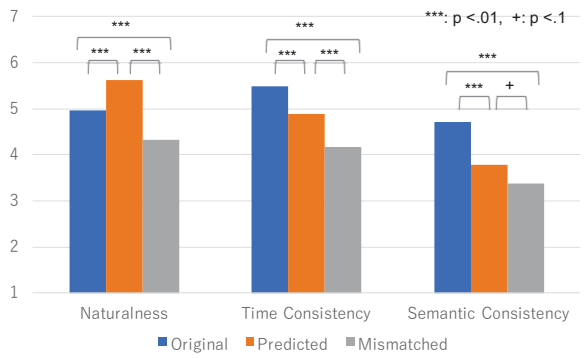


Figure 6: Results of Questionnaire.

Table 4: Comparison Among Conditions on AJs.

Type of Gesture	AJ
Original	0.37
Predicted	**0.06**
Mismatched	0.37

shown in Table 3. Thus, we will summarize them into scales for later analysis.

We conducted one-way ANOVA on each scale for each speech, and used the Bonferroni method for post-hoc analysis. Regarding the results of the ANOVAs, we found significant main effects in all three scales: naturalness ($F(2, 58) = 2.407$), time consistency ($F(2, 58) = 1.259$), and semantic consistency ($F(2, 58) = 2.533$). The results of our post-hoc analyses are shown in Figure 6.

7 DISCUSSION

7.1 Naturalness

The predicted gestures were perceived as more natural than not only the mismatched gestures, but also the original gestures. Those results were unexpected. Therefore, we investigated the dataset and found that the dataset had a small amount of recording noise originating from tracking errors in the motion capture system. On the contrary, the predicted gestures were smoothed by temporal filtering and they had no discontinuities in motion. Thus, gestures in the prediction condition looked smoother than gestures in the original condition.

To evaluate the discontinuities in time for each condition, we calculated the Average Jerk (AJ) [23]. Jerk is a temporal differentiation of acceleration. Thus, the more discontinuities a gesture has, the larger the AJ. As shown in Table 4, the AJ of the predicted gestures was one sixth of the AJ of the original gestures and the mismatched gestures which were raw motion data from the database.

Therefore, we should evaluate gestures in terms of naturalness in which we control discontinuities in gestures for all condition by temporal filtering in the future.

7.2 Time Consistency

The predicted gestures were rated better than the mismatched gestures in terms of time consistency. However, in our experiment, the predicted gestures did not outperform the original gestures. One of the reasons was that the predicted gestures moved too smoothly. It seemed that the predicted gestures moved with a steady speed when compared to the original gestures. Generally, the initial movements of gestures are faster than following movements, and usually synchronize to the beginning of words or phrases in speech.

Figure 9 shows histograms of moving distance per frame of both the original gestures and the predicted gestures. As can be seen, the original gestures had higher speed movements compared to the predicted gestures, and the predicted gestures without filtering had similar characteristics to the original gestures. On the other hand, the filtered predicted gestures had lower speed movements compared to the original gestures. Therefore, it is possible that we can improve perceived time consistency with improved filtering methods.

In addition, Chiu et al. [7] predicted 3D positions from prosodic feature, pitch and sound pressure. And they confirmed that prediction performed *beat* gestures were as natural as original gestures. Therefore, it is possible for us to have more explicit features along with MFCC features.

7.3 Semantic Consistency

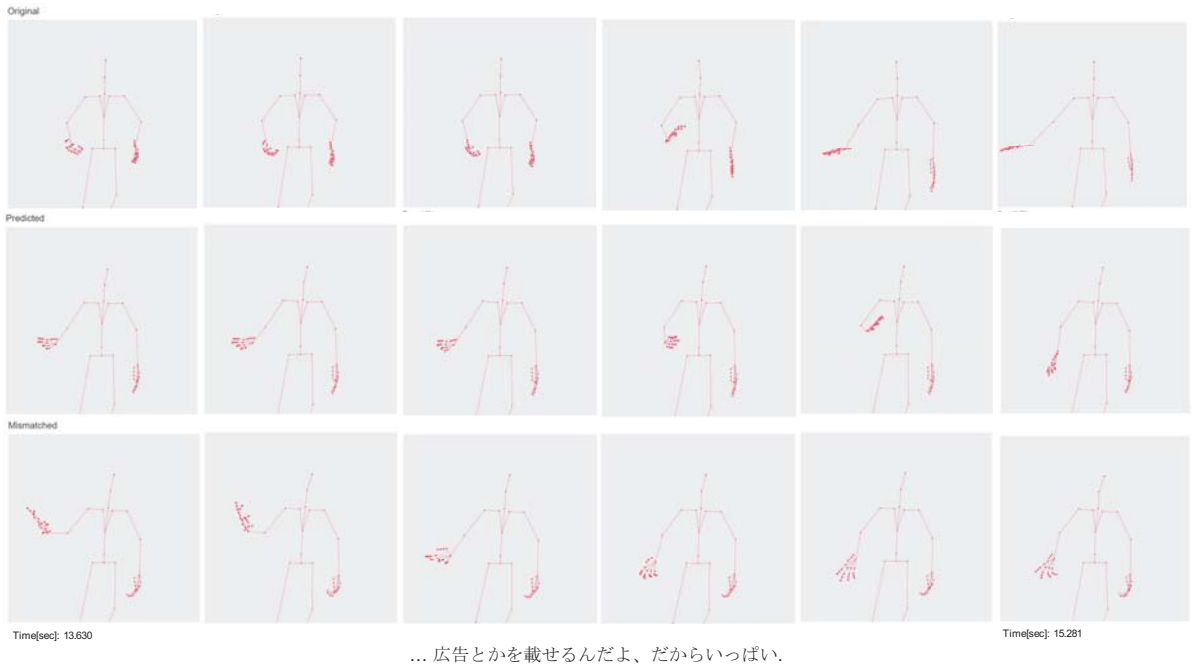
In the evaluation of semantic consistency, the predicted gestures were rated lower than the original gestures. We knew that it was a rather difficult goal to achieve, more so than naturalness or time consistency. However, if we take a close look, we were able to find gestures in the prediction which can be interpreted as semantically similar to the original gestures.

For example, Figure 7 shows a part of one of the test speeches, saying "...(you have to) post advertisements, and I mean a lot." and is accompanied by gestures in the three conditions. The original gestures performed a motion in the later part which meant "putting something on." Likewise, the predicted gestures also performed a motion which can be interpreted as similar.

There is also another example shown in Figure 8. It shows a gesture accompanying this speech instance: "... there are jobs which are easy and pay more, and there are jobs which are hard and pay less ... " The original gesture put the first job on the left side, and then next job on the right. On the other hand, it seemed that the predicted gesture put the first one on the right side, and the next one on the left side.

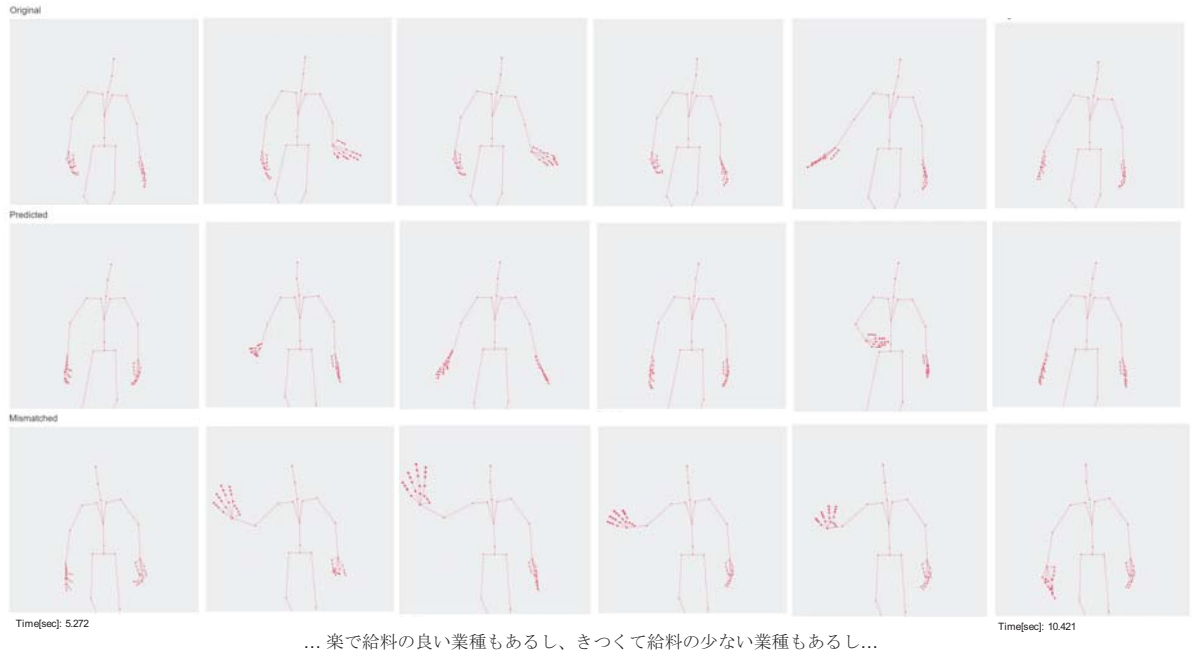
However, those similarities were probably coincidence. We believe that predicting semantically valid gesture is the ultimate goal of this research. And there might be several possible courses of action we can take to further the research to the next step. We believe that creating a dataset is the most important thing to do. First of all, we are not trying to produce all gestures. We aim for only relatively abstract gestures such as *metaphoric* gestures as defined in [21].

Gestures, usually movements of the hands and arms, which spontaneously occur in accompaniment with speech, are classified into four types: *iconic*, *metaphoric*, *beat*, and *deictic*. *Iconic* gestures present images of concrete objects or actions, and *deictic* gestures are used



… 広告とかを載せるんだよ、だからいっぱい.
… (you have to) post advertisements, and I mean a lot.

Figure 7: Speech 1107 (https://youtu.be/4Tkhh-dzpTk).



… 楽で給料の良い業種もあるし、きつくて給料の少ない業種もあるし…
… there are jobs which are easy and pay more, and there are jobs which are hard and pay less …

Figure 8: Speech 1112 (https://youtu.be/06cinEAPPfI).

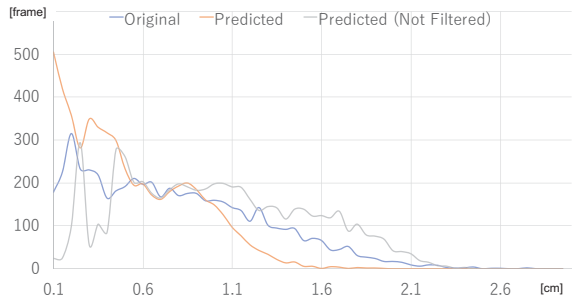


Figure 9: Histogram of Moving Distance per Frame.

to indicate environmental objects around the speaker. *Iconic* and *deictic* may be called gestures of the concrete. *Beat* gestures are movements with the rhythmical pulsations of speech which index the accompanied word or phrase as being significant. Contrary to the two concrete types, *iconic* and *deictic*, in *metaphoric* gestures, abstract meaning is presented as if it had form by utilizing space. So, *metaphoric* and *beat* gestures may be called gestures of the abstract. We believe that we can automatically generate *metaphoric* and *beat* gestures. However, our dataset has still only 196 minutes of speech. We will need more data to create a better model to produce *metaphoric* and *beat* gestures.

8 CONCLUSIONS

We proposed a method to produce gesture motion as 3D position sequences from speech audio input. Speech audio data were converted to MFCC features and then fed into a Bi-Directional LSTM Network. The network was trained by a speech-motion paired dataset of 1,534 sentences in total. The network output 3D position sequences of sixty four joints of a whole body. The output was smoothed by two kinds of temporal filtering. The proposed method was evaluated by a user study of thirty participants. The participants compared the gestures predicted by the proposed method with the original gestures and the mismatched gestures. Regarding the results, we found that the participants perceived that the predicted gestures were more time consistent than the mismatched gestures. Also, the generated gesture showed marginally significant improvement in terms of semantic consistency when compared to the mismatched gestures.

ACKNOWLEDGMENTS

This work was supported by JSPS Grant-in-Aid for Young Scientists (B) Grant Number 17K18075.

REFERENCES

[1] Paavo Alku, Tom Bäckström, and Erkki Vilkman. 2002. Normalized amplitude quotient for parametrization of the glottal flow. *The Journal of the Acoustical Society of America* 112, 2 (2002), 701–710.
[2] Timothy W Bickmore, Laura M Pfeifer, Donna Byron, Shaula Forsythe, Lori E Henault, Brian W Jack, Rebecca Silliman, and Michael K Paasche-Orlow. 2010. Usability of conversational agents by patients with inadequate health literacy: Evidence from two clinical trials. *Journal of Health Communication* 15, S2 (2010), 197–210.
[3] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 2012. 1€ filter: A simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 2527–2530.
[4] Justine Cassell. 2000. *Embodied conversational agents*. MIT press.
[5] Justine Cassell, Stefan Kopp, Paul Tepper, Kim Ferriman, and Kristina Striegnitz. 2007. Trading spaces: How humans and humanoids use speech and gesture to give directions. *Conversational Informatics* (2007), 133–160.
[6] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. BEAT: the behavior expression animation toolkit. In *Proceedings of the SIGGRAPH Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. 477–486.
[7] Chung-Cheng Chiu and Stacy Marsella. 2011. How to train your avatar: A data driven approach to gesture generation. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA)*. 127–140.
[8] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: a deep and temporal modeling approach. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA)*. 152–166.
[9] Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28, 4 (1980), 357–366.
[10] Rosalind Edwards and Janet Holland. 2013. *What is qualitative interviewing?* Bloomsbury Academic.
[11] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* 44, 3 (2011), 572–587.
[12] Rui Fang, Malcolm Doering, and Joyce Y Chai. 2015. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 271–278.
[13] Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5 (2005), 602–610.
[14] Cheolho Han, Sang-Woo Lee, Yujung Heo, Wooyoung Kang, Jaehyun Jun, and Byoung-Tak Zhang. 2017. Criteria for human-compatible AI in two-player vision-language tasks. In *Proceedings of the Linguistic And Cognitive Approaches To Dialog Agents (LaCATODA), Workshop of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)*. 28–33.
[15] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567.
[16] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*. 448–456.
[17] Sam Johnson and Mark Everingham. 2010. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference (BMVC)*. 12.1–12.11. doi:10.5244/C.24.12.
[18] Adam Kendon. 1980. Gesticulation and speech: Two aspects of the process of utterance. *The relationship of verbal and nonverbal communication* 25, 1980 (1980), 207–227.
[19] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
[20] A Lawson, Pavel Vabishchevich, M Huggins, P Ardis, Brandon Battles, and A Stauffer. 2011. Survey and evaluation of acoustic features for speaker recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5444–5447.
[21] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.
[22] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the International Conference on Machine Learning (ICML)*. 807–814.
[23] Brandon Rohrer, Susan Fasoli, Hermano Igo Krebs, Richard Hughes, Bruce Volpe, Walter R Frontera, Joel Stein, and Neville Hogan. 2002. Movement smoothness changes during stroke recovery. *Journal of Neuroscience* 22, 18 (2002), 8297–8304.
[24] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.
[25] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.
[26] Kenta Takeuchi, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa, and Hiroshi Sakuta. 2017. Creating a gesture-speech dataset for speech-based automatic gesture generation. In *Proceedings of the International Conference on Human-Computer Interaction (HCI)*. 198–202.
[27] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. 2015. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 648–656.