Language Agnostic Gesture Generation Model: A Case Study of Japanese Speakers' Gesture Generation Using English Text-to-Gesture Model

Genki Sakata[1], Naoshi Kaneko[2], Dai Hasegawa[3] and Shinichi Shirakawa[1][a]

[1] *Yokohama National University, Yokohama, Kanagawa, Japan*

[2] *Aoyama Gakuin University, Sagamihara, Kanagawa, Japan*

[3] *Hokkai Gakuen University, Sapporo, Hokkaido, Japan*

Keywords: Gesture Generation, Spoken Text, Multilingual Model, Neural Networks, Deep Learning, Human-Agent Interaction.

Abstract: Automatic gesture generation for speech audio or text can reduce the human effort required to manually create the gestures of embodied conversational agents. Currently, deep learning-based gesture generation models trained using a large-scale speech–gesture dataset are being investigated. Large-scale gesture datasets are currently limited to English speakers. Creating these large-scale datasets is difficult for other languages. We aim to realize a language-agnostic gesture generation model that produces gestures for a target language using a different-language gesture dataset for model training. The current study presents two simple methods that generate gestures for Japanese using only the text-to-gesture model trained on an English dataset. The first method translates Japanese speech text into English and uses the translated word sequence as input for the text-to-gesture model. The second method leverages a multilingual embedding model that embeds sentences in the same feature space regardless of language and generates gestures, enabling us to use the English text-to-gesture model to generate Japanese speech gestures. We evaluated the generated gestures for Japanese speech and showed that the gestures generated by our methods are comparable to the actual gestures in several cases, and the second method is promising compared to the first method.

1 INTRODUCTION

1.1 Background

Embodied conversational agents that interact with humans have become common with the progress of computing and artificial intelligence technologies. Humans read information from verbal as well as non-verbal cues, such as gestures and facial expressions in human–human communication. Therefore, embodied conversational agents, including virtual characters and humanoid robots, are required to implement human-like gestures and realize smooth human–computer interaction. However, manually creating gestures is time-consuming and labor intensive for content creators because it requires designing and implementing gesture motions according to speech content. Even if we record gesture motions using a motion capture device, facilities and actors would be required. Automatic gesture generation methods have been developed to automate the gesture creation process (Cassell et al., 2001; Levine et al., 2010; Chiu et al., 2015). Training a gesture generation model based on deep learning is a recent trend (Ginosar et al., 2019; Ahuja et al., 2020; Kucherenko et al., 2020; Yoon et al., 2020; Bhattacharya et al., 2021).

1.2 Related Work on Gesture Generation Models

Machine learning and deep learning techniques are often used to construct gesture generation models. In gesture generation models, speech audio or text is typically used as input. Output gestures are represented by a sequence of 2D or 3D coordinates/joint angles of human joint points. Specifically, audio features, such as Mel frequency cepstral coefficient (MFCC), and word embedding features, such as fastText (Bojanowski et al., 2017) and BERT (Devlin et al., 2019),

[a] https://orcid.org/0000-0002-4659-6108

Sakata, G., Kaneko, N., Hasegawa, D. and Shirakawa, S.
Language Agnostic Gesture Generation Model: A Case Study of Japanese Speakers' Gesture Generation Using English Text-to-Gesture Model.
DOI: 10.5220/0011643600003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 2: HUCAPP*, pages 47-54
ISBN: 978-989-758-634-7; ISSN: 2184-4321

are used as input of gesture generation models.

Hasegawa et al. (2018) constructed a gesture generation model that predicts 64 3D-joint coordinates from speech audio features. They used a bidirectional long-short-term memory (LSTM) network and a Japanese speech–gesture dataset of approximately five hours created by Takeuchi et al. (2017) using a motion capture device. Ginosar et al. (2019) created a large-scale English speech–gesture dataset from YouTube videos (144 hours in total) by using the OpenPose system (Cao et al., 2017; Simon et al., 2017) and trained the gesture generation model using U-Net (Ronneberger et al., 2015), in which the input and output of the model are speech audio features and 49 2D-joint coordinate sequences, respectively. Several studies have used speaker identity information as input to reflect personality and style in generated gestures (Ahuja et al., 2020; Bhattacharya et al., 2021; Yoon et al., 2020). Bhattacharya et al. (2021) constructed a text-to-gesture model based on Transformer (Vaswani et al., 2017), which used speech text and speaker attributes as input and predicted 23 3D-joint coordinates. Asakawa et al. (2022) evaluated text-to-gesture models using U-Net (Ronneberger et al., 2015). Their models take a word embedding feature sequence as input and output a sequence of 49 2D-joint coordinates. They further implied that the size and quality of gesture datasets affect the quality of gestures generated.

In general, a large-scale dataset contributes to improving the performance of deep learning-based models. Several works collected large-scale datasets of English speech–gesture pairs (Ginosar et al., 2019; Ahuja et al., 2020; Yoon et al., 2020) using the OpenPose system (Cao et al., 2017; Simon et al., 2017), which estimates human skeletal information from videos.

1.3 Motivation and Contribution

Although existing methods have succeeded in generating smooth and human-comparable gesture motions, the used large-scale datasets are limited to English speakers. Therefore, current gesture generation studies are conducted primarily for English content. Collecting a large-scale speech–gesture dataset is difficult for other languages, even when using OpenPose. This is because we should prepare appropriate videos to collect high-quality gesture data by OpenPose, e.g., the angle and scale of view must be stable during gestures, and the person performing the gesture must fit on the screen. Such appropriate videos that provide accurate gesture motions through OpenPose are limited in other languages spoken by not so many people compared to English, such as Japanese. Moreover, collecting a large speaker dataset using a motion capture device is impractical owing to its high cost. Additionally, collecting speech–gesture datasets for many languages may be impractical.

This study aims to realize a language-agnostic gesture generation model that produces gestures for a target language using a different language's large-scale gesture dataset for model training. This would be valuable for constructing gesture generation models for languages spoken by not so many people because we can eliminate the need to collect a gesture dataset for a target language. To the best of our knowledge, gesture generation for a target language using only another language gesture datasets has not been examined. Therefore, we start with simple approaches toward a language-agnostic gesture generation model.

We present two simple methods for applying the text-to-gesture model trained using a specific language dataset for speakers of another language. In particular, this study considered generating gestures of Japanese speakers by leveraging an English speaker's gesture generation model, as a case study. In the first method, we simply translate Japanese speech text into English using a translation system and input the translated English word sequence into the English text-to-gesture model. The second method uses a multilingual embedding model that embeds English and Japanese sentences in the same feature space. We train the gesture generation model from multilingual embedding features using the English dataset and use it to produce gestures for Japanese speech text. Our methods do not require a Japanese speech–gesture dataset for model training. We evaluated the gestures generated for Japanese texts using the proposed methods through a quantitative evaluation and user study. The results show that the quality of several gestures generated by our methods is comparable to that of actual gestures, and the second method is better than the first one in several cases.

The contribution of this paper are summarized as follows:

- We tackled a novel problem setting for gesture generation, which generates gestures for Japanese texts using only the gesture generation model trained on an English dataset.
- We proposed and evaluated two simple methods that leverage the translation system or the multilingual model.

2 PROPOSED METHODS

This section introduces the two methods used to exploit the text-to-gesture model trained on the English dataset to generate gestures for Japanese speech texts. First, we formally describe the problem setting in 2.1. In 2.2, we explain the text-to-gesture model we used. Then, we describe the proposed methods in 2.3 and 2.4. Although we explain our gesture generation method in terms of Japanese speech, we note that our method is applicable for any language. An overview of our methods is illustrated in Fig. 1.

2.1 Problem Setting

We specifically targeted the construction of a text-to-gesture model for Japanese without Japanese speech–gesture data for model training, while we can access a sufficient amount of English speech–gesture data. We considered using text information as input for the gesture generation models rather than audio input to leverage language translation systems and multilingual embedding models. We used the English gesture dataset collected by Ginosar et al. (2019) and its text information provided by Asakawa et al. (2022). This dataset contains 49 sets of 2D keypoint coordinates for the neck, shoulders, arms, elbows, and fingers obtained from video using OpenPose (Cao et al., 2017; Simon et al., 2017), and the spoken words in each frame. The frame per second of gesture motions was 15 FPS. We denote this English dataset as $\mathcal{D}_{\mathrm{EN}} = \{(\boldsymbol{x}_i, \boldsymbol{t}_i) | i = 1, \dots\}$, where $\boldsymbol{x}_i \in \mathcal{W}_{\mathrm{EN}}^{(N)}$ and $\boldsymbol{t}_i \in \mathbb{R}^{98 \times N}$ indicate the sequences of English words and 49 2D-keypoint coordinates representing gesture motion for the ith data, respectively, and N represents the sequence length (the number of frames). Note that to add the speech length information as input, words are duplicated over the corresponding frames while it is being pronounced. In addition, a special token *BLANK* is used, which represents no utterance. Our problem is to produce a text-to-gesture model for the Japanese speech text input $\boldsymbol{x}_{\mathrm{JP}} \in \mathcal{W}_{\mathrm{JP}}$ only using the English dataset $\mathcal{D}_{\mathrm{EN}}$, where $\mathcal{W}_{\mathrm{JP}}$ indicates the set of Japanese spoken word sequences.

2.2 Base Model of Text-to-Gesture Generation

We adopted the text-to-gesture model proposed in (Asakawa et al., 2022) as a baseline model. In the proposed methods, we first train the text-to-gesture model using the English dataset $\mathcal{D}_{\mathrm{EN}}$. The training setting was the same as that in (Asakawa et al., 2022), except for the word embedding method. Although fastText (Bojanowski et al., 2017) was used as the word embedding method in (Asakawa et al., 2022), we used LaBSE (Feng et al., 2022) for multilingual support. Word and sentence embedding have been widely used in natural language processing; it allows us to obtain a feature vector of a word or sentence. The bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019) is a representative model that obtains a word or sentence embedding. Language-agnostic BERT sentence embedding (LaBSE) (Feng et al., 2022) is a state-of-the-art multilingual sentence embedding model for translation based on BERT. In LaBSE, multilingual sentence embeddings are trained by combining several tasks; it supports 109 languages. The LaBSE model provides a multilingual embedding feature vector for a given sentence, and multilingual sentences are embedded in the same feature space. That is, similar meaning sentences among different languages are expected to be embedded in similar feature vectors.

Although LaBSE provides the embedded vector for a sentence, we input a word into the model and obtained the embedded vector for a word. This enables us to use the existing text-to-gesture model directly and embed similar words into the similar feature vector regardless of language. Our second proposed method exploits this LaBSE embedding property. We denote the embedded feature vectors of the word sequence $\boldsymbol{x}$ as $\boldsymbol{e} \in \mathbb{R}^{D \times N}$, where D is the dimension of the embedded feature. Each feature vector $e_j \in \mathbb{R}^D$ corresponding to the jth word x_j in $\boldsymbol{x}$ is given by $e_j = \mathcal{E}_{\mathrm{LaBSE}}(x_j)$. We denote the embedded vectors of $\boldsymbol{x}$ as $\boldsymbol{e} = (e_1, \dots, e_N)$. Note that $\mathcal{E}_{\mathrm{LaBSE}} : \mathcal{W} \to \mathbb{R}^D$ indicates the word embedding function by LaBSE, where $\mathcal{W}$ is the word space in any language. In our case, the dimension of the embedded vector was $D = 768$.

Then, the gesture generation function $\mathcal{F} : \mathbb{R}^{D \times N} \to \mathbb{R}^{98 \times N}$ was trained using the English dataset $\mathcal{D}_{\mathrm{EN}}$. Note that the input of the gesture generator is the embedded vectors denoted by $\boldsymbol{e} = \mathcal{E}_{\mathrm{LaBSE}}(\boldsymbol{x}) \in \mathbb{R}^{D \times N}$. The neural architecture of the gesture generator is the U-Net-based convolutional neural network (CNN) used in (Asakawa et al., 2022). In the architecture, the input vectors of $D \times N$ were downsampled to the size of $D \times N/32$ by 1D convolution operations; they were then transformed to a size of $98 \times N$ by upsampling and 1D convolution operations. There are skip connections between the downsampling and upsampling blocks. Note that any sequence length can be processed because this architecture is a fully convolutional neural network. Given the sequence of ground-truth motion coordinates as $\boldsymbol{t} \in \mathbb{R}^{98 \times N}$, the

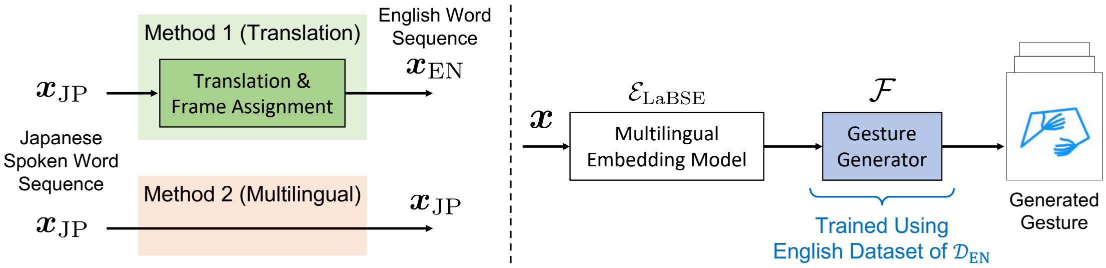

Figure 1: Overview of Our Methods.

loss function is defined by

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{t})\in\mathcal{D}_{\mathrm{EN}}}[\|\boldsymbol{t}-\mathcal{F}(\mathcal{E}_{\mathrm{LaBSE}}(\boldsymbol{x}))\|_1] \\ +\mathbb{E}_{(\boldsymbol{x},\boldsymbol{t})\in\mathcal{D}_{\mathrm{EN}}}[\|T(\boldsymbol{t})-T(\mathcal{F}(\mathcal{E}_{\mathrm{LaBSE}}(\boldsymbol{x})))\|_1] \quad , \quad (1)$$

where $T(\boldsymbol{t}) = (t_2 - t_1, \ldots, t_N - t_{N-1})$ is the temporal difference vector of the keypoint coordinates. The first term is the L1 loss between the keypoint coordinates of the ground truth and generated gesture, and the second term is the L1 loss between the velocities of the ground truth and generated gesture motions. The details of the architecture and training hyperparameters can be found in (Asakawa et al., 2022). We used the publicly available code from (Asakawa et al., 2022).

In the proposed methods, we considered providing gesture motions for Japanese speech text $\boldsymbol{x}_{\mathrm{JP}} \in \mathcal{W}_{\mathrm{JP}}$ using the text-to-gesture function $\mathcal{F}$ trained on the English dataset $\mathcal{D}_{\mathrm{EN}}$.

2.3 Gesture Generation by Translation (Method 1)

In the first method, we translated Japanese speech text pronounced in N' frames into English text using a language translation system. In the experiment, we used DeepL,[1] a neural machine translation service. Note that we cannot know which English word should be assigned to each frame because the translated text does not have information on the pronunciation length or timing of each word. Moreover, the number of frames to be input into the gesture generator should be N' to generate a gesture of the same length as the Japanese speech. To address this problem, we simply assigned the same number of frames for each word such that the total sequence length was N'. That is, the number of frames for each English word was approximately N'/N_{words}, where N_{words} is the number of words in the translated English text. We denote the English word sequence with the N' frames given by this process as $\boldsymbol{x}_{\mathrm{EN}} \in \mathcal{W}_{\mathrm{EN}}^{(N')}$. Then, we input the English word sequence into the word embedding and gesture generator to obtain the generated gesture motion as $\mathcal{F}(\mathcal{E}_{\mathrm{LaBSE}}(\boldsymbol{x}_{\mathrm{EN}}))$. We call this method “Method 1 (Translation).”

The order and number of words in sentences between Japanese and English differ, even if they express the same meanings. The text-to-gesture model provides the gesture motion based on the input English word order and number of frames for each English word. Therefore, the generated gestures can mismatch with the original Japanese speech text, may leading to unnatural gestures. However, we note that human gestures are not so rigorously time-aligned with spoken words. Namely, the timing of a gesture and a spoken word could be misaligned in human communication, as indicated in (McNeill, 1996). Therefore, we experimentally evaluate the performance of this simple method.

2.4 Gesture Generation Using Multilingual Embedding (Method 2)

The second method exploits the property of multilingual embedding models for gesture generation in different languages. We prepared a Japanese speech text pronounced in N' frames, $\boldsymbol{x}_{\mathrm{JP}} \in \mathcal{W}_{\mathrm{JP}}^{(N')}$, which includes the number of pronounced frames for each Japanese word. This speech text information was obtained using the speech-to-text system in the experiment. We applied the LaBSE model directly to $\boldsymbol{x}_{\mathrm{JP}}$ to obtain the embedded feature vectors as $\boldsymbol{e} = \mathcal{E}_{\mathrm{LaBSE}}(\boldsymbol{x}_{\mathrm{JP}})$. Subsequently, we input the embedded feature vectors into the gesture generator, obtaining the gesture motion as $\mathcal{F}(\mathcal{E}_{\mathrm{LaBSE}}(\boldsymbol{x}_{\mathrm{JP}}))$. We call this method “Method 2 (Multilingual).”

The LaBSE multilingual language model provides the same feature space between different languages, and the text-to-gesture model was trained to generate gesture motions from embedded features. Therefore, we expected that the gesture generator would work properly for Japanese speech text, even if it was trained on an English text-gesture dataset. In this method, the order of words and number of frames for each word in the input of the gesture generator were

[1] https://www.deepl.com/

Table 1: Mean absolute error (MAE) between the coordinates of the generated gesture and ground truth motion.

Video	Method 1	Method 2
ID	(Translation)	(Multilingual)
1	79.20	**69.58**
2	179.08	**135.75**
3	**102.23**	123.81
4	78.52	**42.38**
5	154.39	**154.12**

Table 2: Standard deviation (STD) of keypoint coordinates of generated gestures.

Video	Ground	Method 1	Method 2
ID	Truth	(Translation)	(Multilingual)
1	27.3	**34.3**	39.4
2	102.7	**135.7**	53.0
3	64.4	**48.6**	96.4
4	8.9	90.5	**35.3**
5	47.4	104.6	**34.9**

Table 3: Percentage of correct keypoints (PCK) between the keypoints of the generated gesture and ground truth motion.

Video	Method 1	Method 2
ID	(Translation)	(Multilingual)
1	0.13	**0.17**
2	0.08	**0.10**
3	**0.20**	0.17
4	0.36	**0.55**
5	**0.06**	0.05

maintained the same as in the original Japanese text. Therefore, we expected this method to generate gesture motions based on the original Japanese text information, leading to more natural gesture motions than Method 1 (Translation).

3 EXPERIMENT AND RESULTS

3.1 General Settings

As described in Section 2.1, we used the English speech-gesture dataset provided by Ginosar et al. (2019),[2] and its text information was provided by Asakawa et al. (2022) [3] as $\mathcal{D}_{\text{EN}}$. We chose speaker "Oliver" to train the text-to-gesture model owing to the large size and high quality of motions extracted by OpenPose in Oliver's dataset. Although the dataset contains face keypoints, they were unused in the gesture generator. However, the face keypoints are used when displaying the gesture motions in the user study. The number of frames in the training data was $N = 64$ (≈ 4.2 s).

We used the pretrained LaBSE model.[4] To tokenize an input word, we used the code from (Yang et al., 2021) following the instructions of the LaBSE model. A zero vector was assigned for the token *BLANK*, indicating that the frame does not contain a word. The pretrained model for word embedding is not updated during the training of the gesture generator.

We prepared the Japanese speech data to compare the generated gestures using our methods. We collected data by trimming five YouTube videos of Japanese speakers.[5] As for the English dataset, we extracted 49 2D keypoint coordinates and face keypoints using OpenPose; the spoken words in each frame were extracted using the Google Cloud Speech-to-Text API. The speakers in Videos 1 to 4 are male, whereas the speaker in Video 5 is female. The length of Video 1 is 64 frames ($\approx$ 4.2 s), that of Videos 2 and 3 is 128 frames ($\approx$ 8.5s), and that of Videos 4 and 5 is 192 frames (12.8s).

3.2 Quantitative Evaluation of Generated Gestures

We report the mean absolute error (MAE) between the coordinates of the generated gesture and ground-truth motion, and the standard deviation (STD) of the keypoint coordinates as quantitative metrics. The MAE is a measure of how similar the generated gesture is to the actual motion, and the STD is a measure of the scale of the gesture. The coordinates were standardized when calculating these metrics, and the STD was averaged over 98 coordinates. Tables 1 and 2 show the MAE and STD for each video, respectively. We observed that Method 2 (Multilingual) can generate gestures with a smaller MAE than Method 1 (Translation), except for Video 3. The STD values of Method 1 (Translation) are closer to the ground truth than those of Method 2 (Multilingual) in Videos 1, 2, and 3, whereas those of Method 2 (Multilingual) are closer in Videos 4 and 5.

We also report the probability of correct keypoints (PCK) (Yang and Ramanan, 2013), a widely used metric for pose detection, between the keypoints of

[2] https://github.com/amirbar/speech2gesture

[3] https://github.com/GestureGeneration/text2gesture_cnn

[4] https://tfhub.dev/google/LaBSE/2

[5] The number of videos for evaluation is small because collecting clean Japanese speakers' gesture data by OpenPose is difficult due to the small scale of videos compared with English, as in our motivation. We aim to evaluate the concept of our methods and demonstrate the possibility of Japanese gesture generation using only the English dataset.

Table 4: Questionnaire used in the user study.

Q1 (Naturalness)	Which gesture looks natural?
Q2 (Smoothness)	Which gesture looks smooth?
Q3 (Human-Likeness)	Which gesture looks like a human movement?
Q4 (Voice Match)	Which gesture matches the speech voice?
Q5 (Content Match)	Which gesture matches the speech content?
Q6 (Understandability)	Which gesture promotes understanding of the speech content?

Table 5: Evaluation results from the user study. Each value indicates the rate answered that the generated gesture is equal to or better than the ground truth. The bold font indicates the higher value between Methods 1 and 2, and the underline indicates that a significant difference exists between the rates of Methods 1 and 2 at a significance level of 5% by Fisher's exact test.

Video ID	Method	Q1	Q2	Q3	Q4	Q5	Q6
1	Method 1 (Translation)	**0.25**	0.3125	**0.4375**	**0.4375**	0.375	**0.6875**
	Method 2 (Multilingual)	**0.25**	**0.375**	0.3125	0.1875	**0.5**	0.5
2	Method 1 (Translation)	0.125	0.1875	0.125	0.25	0.375	0.4375
	Method 2 (Multilingual)	**0.6875**	**0.5625**	**0.6875**	**0.4375**	**0.5**	**0.5625**
3	Method 1 (Translation)	**0.1875**	**0.375**	**0.3125**	**0.1875**	**0.1875**	**0.125**
	Method 2 (Multilingual)	0.0625	0	0.25	0.0625	0.125	0.0625
4	Method 1 (Translation)	0.5625	0.5625	0.5	0.6875	0.6875	0.8125
	Method 2 (Multilingual)	**0.6875**	**0.9375**	**0.5625**	**0.75**	**0.875**	**1**
5	Method 1 (Translation)	0.25	0.1875	0.3125	0.125	0.125	0.25
	Method 2 (Multilingual)	**0.5**	**0.4375**	**0.4375**	**0.1875**	**0.375**	**0.375**

the generated gesture and ground truth motion. The PCK is the accuracy given by comparing the keypoints between the generated and ground truth motion. As done in (Ginosar et al., 2019), the averaged PCK values over $\alpha = 0.1, 0.2$ are reported, where α is a parameter determining acceptable errors between predicted and ground truth keypoints. Table 3 shows the PCK values for each video. The tendency of the result is similar to that of MAE. That is, Method 2 (Multilingual) is superior to Method 1 (Translation) for three videos. Because there is generally no unique correct gesture for a given speech, the quality of gestures generated should be evaluated in a user study.

3.3 User Study for Generated Gestures

In the user study, each gesture generated for the Japanese speech was compared to the corresponding ground truth motion extracted by OpenPose to evaluate the generated gestures. A total of 32 native Japanese speakers, 26 men and six women between the ages of 18 and 51, participated. We followed the user study conducted in (Asakawa et al., 2022). Participants watched the generated gesture and its ground-truth motion videos placed one above the other and then answered six questions on gesture quality. The position of the generated gesture and ground-truth motion was randomized. The face keypoints extracted by OpenPose and speech audio of the original video were also displayed with the gesture motions. The questionnaire for the participants is shown in 4. Each participant watched the videos and selected the answers from “Upside,” “Downside,” and “Same level.” The user study was conducted using a Google form. Participants answered to either the experiment for Method 1 or 2, i.e., participants did not score both Methods 1 and 2. We collected the answers from 16 participants for each method.

If a participant selected the answer corresponding to the generated gesture or that of “Same level,” the generated gesture can be regarded equal to or better than the ground-truth gesture. We computed the rate answered that the generated gesture is equal to or better than the ground-truth for each question. That is, the high value of this rate indicates a better gesture. Table 5 shows the results of the user study, where the bold font indicates the higher value between Methods 1 and 2, and the underline indicates that a statistical significance exists between Methods 1 and 2.

We observe that all values for Video 4 by Methods 1 and 2 and that the most values for Video 2 by Method 2 are greater than 0.5, implying that the gestures generated are comparable to ground-truth motions. The generated gestures obtaining high scores in Video 4 may be because the ground-truth motion in Video 4 has fewer movements, whereas the generated gestures have more movements, as shown in Fig. 2. Comparing Methods 1 and 2, Method 2 shows better results in Videos 2, 4, and 5 although the scores of Method 2 are inferior to Method 1 in Video 3. We

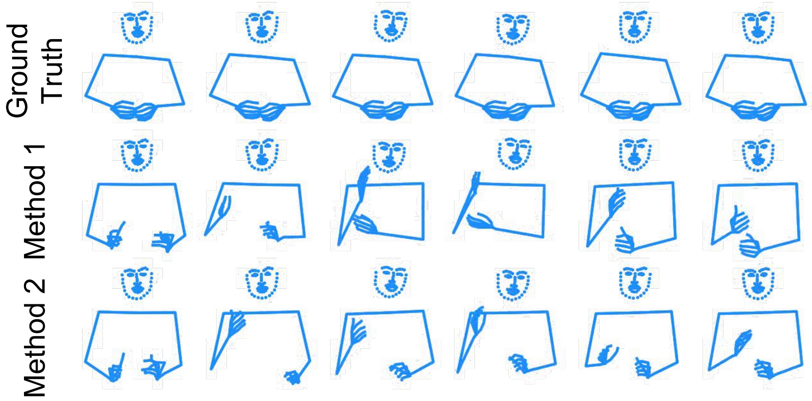

Figure 2: Example of the generated and ground-truth gestures for Videos 4. Each image is 0.5 seconds apart.

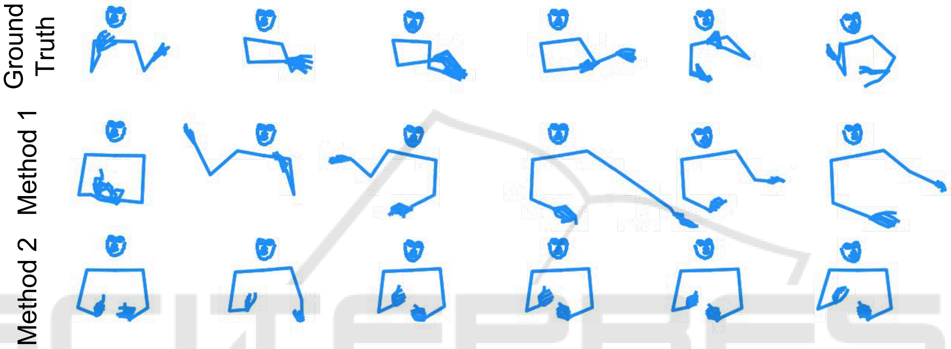

Figure 3: Example of the generated and ground-truth gestures for Videos 2. Each image is 0.5 seconds apart.

observe from Fig. 3 that the gesture by Method 2 is more smooth and natural than that of Method 1 and even that of the ground truth motion. In this case, the hand motion in the gesture generated by Method 1 appeared to be mismatching the speech content and unnatural. Checking gestures for Video 3, the gesture generated by Method 2 included mismatch and unnatural movements, as observed in Video 2 by Method 1.

4 CONCLUSION

This study presented two simple methods to generate gestures for a target language without its gesture dataset. We demonstrated gesture generation of Japanese speech text using the text-to-gesture model trained on an English dataset. The experimental evaluation showed that our methods could generate gestures comparable to actual gestures in several cases. In addition, we observed that Method 2 (Multilingual) is better than Method 1 (Translation) in several cases. An extensive user study with more Japanese speeches and participants should be conducted to fully understand the effect of our methods. In particular, investigating when Method 2 is superior to Method 1 will be useful for future research. Although we used only text information as model input, adding audio information to the gesture generator is a possible focus for future work.

Our methods can be straightforwardly applied to other languages other than Japanese. It can be realized by translating the target language sentence to English using a translation system in Method 1 (Translation). For Method 2 (Multilingual), we can embed other language sentences into the same feature space by the multilingual model of LaBSE and generate gestures. Extending our case study to other languages is an interesting direction. Finally, fine-tuning the gesture generation model using a small dataset of the target language is a possible future work.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI, Japan Grant Number JP21K12160.

REFERENCES

Ahuja, C., Lee, D. W., Nakano, Y. I., and Morency, L. (2020). Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *European Conference on Computer Vision (ECCV)*, volume 12363 of *LNCS*, pages 248–265. Springer International Publishing.

Asakawa, E., Kaneko, N., Hasegawa, D., and Shirakawa, S. (2022). Evaluation of text-to-gesture generation model using convolutional neural network. *Neural Networks*, 151:365–375.

Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., and Manocha, D. (2021). Text2Gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310.

Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2001). BEAT: The behavior expression animation toolkit. In *28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*, pages 477–486. Association for Computing Machinery.

Chiu, C.-C., Morency, L.-P., and Marsella, S. (2015). Predicting co-verbal gestures: A deep and temporal modeling approach. In *15th International Conference on Intelligent Virtual Agents (IVA)*, pages 152–166. Springer International Publishing.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186. Association for Computational Linguistics.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT sentence embedding. In *60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 878–891. Association for Computational Linguistics.

Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., and Malik, J. (2019). Learning individual styles of conversational gesture. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3492–3501.

Hasegawa, D., Kaneko, N., Shirakawa, S., Sakuta, H., and Sumi, K. (2018). Evaluation of speech-to-gesture generation using bi-directional lstm network. In *18th International Conference on Intelligent Virtual Agents (IVA)*, pages 79–86. Association for Computing Machinery.

Kucherenko, T., Jonell, P., van Waveren, S., Henter, G. E., Alexandersson, S., Leite, I., and Kjellström, H. (2020). Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *2020 International Conference on Multimodal Interaction (ICMI)*, pages 242–250. Association for Computing Machinery.

Levine, S., Krähenbühl, P., Thrun, S., and Koltun, V. (2010). Gesture controllers. *ACM Transactions on Graphic*, 29(4).

McNeill, D. (1996). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer International Publishing.

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653.

Takeuchi, K., Kubota, S., Suzuki, K., Hasegawa, D., and Sakuta, H. (2017). Creating a gesture-speech dataset for speech-based automatic gesture generation. In *HCI International 2017 – Posters' Extended Abstracts*, pages 198–202. Springer International Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yang, Y. and Ramanan, D. (2013). Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890.

Yang, Z., Yang, Y., Cer, D., Law, J., and Darve, E. (2021). Universal sentence representation learning with conditional masked language model. In *2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6216–6228. Association for Computational Linguistics.

Yoon, Y., Cha, B., Lee, J.-H., Jang, M., Lee, J., Kim, J., and Lee, G. (2020). Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, 39(6).