ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095304

SALIENT CO-SPEECH GESTURE SYNTHESIZING WITH DISCRETE MOTION REPRESENTATION

Zijie Ye[1], *Jia Jia*[1,2,*], *Haozhe Wu*[1], *Shuo Huang*[1], *Shikun Sun*[1], *Junliang Xing*[1]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]Beijing National Research Center for Information Science and Technology
yzjscwy@gmail.com, jjia@tsinghua.edu.cn

ABSTRACT

Synthesizing co-speech gestures is challenging because the mapping from speech to gesticulation is inherently non-deterministic. When giving talks, people conduct not only gentle and rhythmic motions but also abrupt and salient gesticulations. Most previous research efforts, however, ignore this nature of co-speech gestures and synthesize deterministic results, producing over-smoothed movements with limited expressiveness. To address this issue, we propose a new co-speech gesture generation approach that produces high-quality salient gesticulations. Specifically, we build a discrete motion representation (DMR) space to bridge the speech-gesture mapping and the gesture generation stages. The incorporation of DMR enables random sampling in motion space and avoids the over-smooth problem in speech-gesture mapping. Based on DMR, we devise a novel multi-modal co-speech gesture synthesis model with temporal attention (MCGT). MCGT explicitly models DMR's categorical distribution conditioned on the speech context, which captures complex context patterns and produces more salient gesticulations in sync with the context. In addition, we construct a new benchmark for evaluating salient motion quality in co-speech gestures, containing a large-scale co-speech gesture dataset with salient gesticulations. We also introduce a new metric, referred to as *salient motion similarity*, to evaluate the salient motion quality. Experiments demonstrate superior results from our approach over several competing baselines.

Index Terms— Co-speech gesture, Discrete motion representation, Temporal attention

1. INTRODUCTION

The co-speech gesture is an essential form of non-verbal communication because it increases the credibility of the speech [1] and helps listeners to understand the speech. Synthesizing co-speech motions is vital to various applications like animation, virtual agents, and social robots. However, it is challenging because the mapping from the speech to the gesture is inherently non-deterministic. When giving talks, people conduct not only gentle, rhythmic motions tied with prosody but also abrupt and salient gesticulations to emphasize facts or convey particular messages. For example, as shown in Figure 1, in the first clip, the speaker waved his hand and put it on the desk rigorously when he emphasized "kill", while in the second clip, he just conducted very gentle motions when he talked about "kill". The same speaker can perform salient or gentle gestures even when saying the same word. As a result, deterministic approaches tend to produce the average of multiple plausible gesture motions, leading to over-smoothed results.

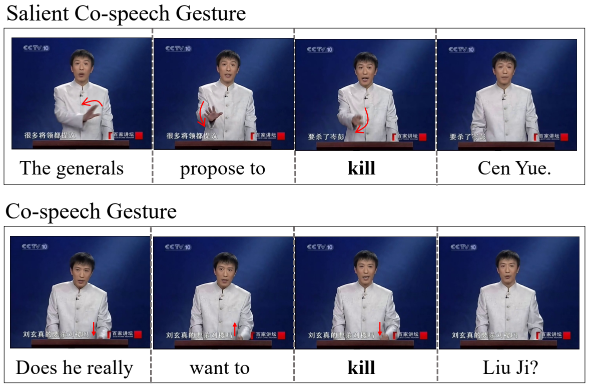

Fig. 1. Examples of the salient co-speech gestures (e.g., when saying "kill" in the top compared to an inconspicuous one in the bottom).

Previous research efforts have shown the rationality of co-speech gesture synthesis. Traditional co-speech gesture generation systems synthesize gestures in a rule-based manner [2, 3] showing limited capacity. Later, researchers propose data-driven approaches to generate co-speech gestures from audio and text transcriptions [4, 5, 6, 7]. However, most previous research efforts adopt CNN [4, 8] or RNN [5, 6, 7] in a deterministic way, producing over-smoothed results that lack expressiveness. Recently, Li et al. [9] propose synthesizing co-speech gestures with a conditional variational auto-encoder, but their model tends to stick to a constant latent code.

To address this problem, we propose synthesizing salient co-speech gestures to improve gesture expressiveness. The key idea of our approach is to leverage DMR as the bridge between the gesture generation stage and the speech-gesture mapping stage. We emphasize the following novelties of our framework: (1) The DMR space constructed by a vector quantized-variational auto-encoder (VQ-VAE) [10] on the gesture motion enables random sampling in motion space and lessens the burden of predicting over-long sequences for the synthesis model. (2) In the speech-gesture mapping stage, we devise a novel multi-modal co-speech gesture synthesis model with temporal attention (MCGT). MCGT captures complex patterns of the speech context, producing more salient gesticulations in sync with the context.

We construct a new benchmark to evaluate the quality of salient motions produced by different methods. It contains a large-scale co-speech gesture dataset with salient gesticulations. We also propose a new metric, salient motion similarity (SMS), to evaluate the quality of salient motions generated by different methods. We perform extensive experiments on the *TED Gesture* [7] dataset and our *LecGesture* dataset. Objective experiments demonstrate that our approach

[*]Corresponding author

generates co-speech gestures with higher SMS than baseline methods while achieving comparable Fréchet Gesture Distance (FGD). The subjective evaluation shows that, with the average opinion score of 17 participants, our approach outperforms baseline methods by 0.34 in expressiveness and 0.37 in speech-gesture synchronization. To summarize, our contributions are as three-fold:

- We propose to synthesize co-speech gestures using discrete motion representation (DMR). By learning a DMR space for gesture motions and modeling the distribution of DMR, our approach generates more high-quality salient motions.
- We devise a novel multi-modal co-speech gesture synthesis model with temporal attention (MCGT). MCGT captures complex patterns of the speech context, producing more salient gesticulations in sync with the context.
- We construct a new benchmark for evaluating salient motion quality in co-speech gesture synthesis. We first collect a large-scale co-speech gesture dataset with more salient gesticulations than existing datasets. Afterward, we introduce salient motion similarity (SMS) as a new metric to evaluate salient motion quality in co-speech gestures.

2. METHODOLOGY

This section describes our proposed co-speech gesture generation framework, as shown in Figure 2. We first construct a discrete motion representation (DMR) space with a VQ-VAE [10]. Afterward, we devise a novel multi-modal co-speech gesture synthesis model with temporal attention (MCGT) based on the DMR space. MCGT predicts the distribution of discrete gesture motion representations conditioned on the speech text, audio, and seed motion.

2.1. Discrete Representation Learning

van den Oord et al. [10] first propose to learn discrete representations without supervision. They incorporate ideas from vector quantization to create the Vector Quantized Variational Auto-Encoder (VQ-VAE). The VQ-VAE consists of an encoder network, a decoder network, and a latent embedding space. The encoder first encodes the input data. Afterward, the input to the decoder is calculated from the output of the encoder by a nearest-neighbor lookup in the latent embedding space. The discrete latent space learned by the VQ-VAE makes effective representations for the data space. Such discrete representations are a more natural fit for many modalities and complex reasoning, as suggested by van den Oord et al. [10].

2.2. DMR Auto-Encoder

To address the non-deterministic mapping issue, we learn a discrete motion representation (DMR) space for the gesture motion, which enables sampling in the motion space and helps to avoid over-smoothed results.

Specifically, we train a VQ-VAE [10] on the speech gesture motion clips. At training time, given a T-frame gesture sequence $\mathbf{X} \in \mathbb{R}^{M \times T}$, where M is the dimension of a gesture frame, we first encode it and perform the nearest neighbor lookup operation in the DMR space. The encoder is a ResNet [11] with a downsampling factor of 8. To achieve high expressiveness of the DMR space, we adopt multi-head categorical latent space following Richard et al. [12]. Specifically, we define our DMR space as $\mathbf{E} \in \mathbb{R}^{H \times D \times K}$, where H is the number of heads, D is the codebook size of each head, and K is the dimensionality of the latent vector. The nearest neighbor lookup is performed on the codebook of each head, respectively. The motion latent of $\mathbf{X}$ is denoted as:

$$\mathbf{C} = \mathrm{quantize}(\mathrm{encode}(\mathbf{X})) \in \mathbb{R}^{H \times K \times \frac{T}{8}}, \tag{1}$$

where $\mathrm{quantize}$ represents the nearest neighbor lookup over each head. Afterward, the decoder reconstructs $\hat{\mathbf{X}} \in \mathbb{R}^{M \times T}$ from $\mathbf{C}$.

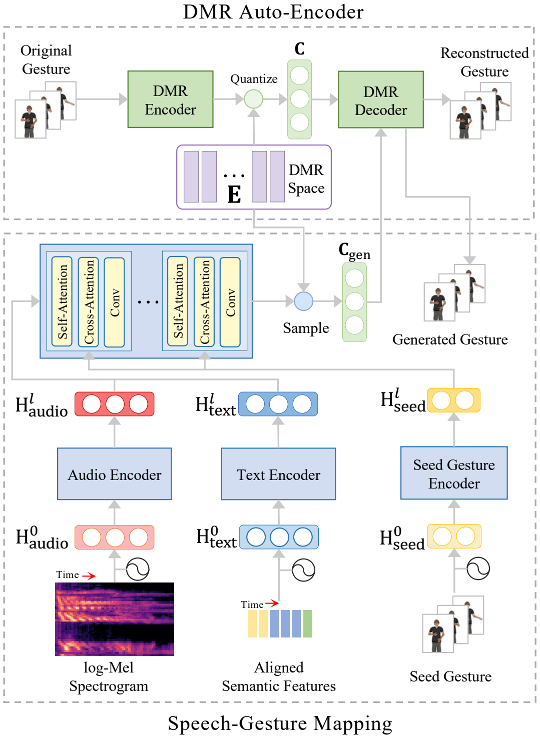

Fig. 2. The pipeline of our framework.

2.3. Speech-Gesture Mapping

Having obtained the DMR space for gesture motions, we devise a multi-modal co-speech gesture synthesis model with temporal attention (MCGT), which predicts the conditional distribution of DMR from the speech text, audio, and seed motion. Our model combines the multi-head attention mechanism [13] with convolutional feed-forward layers. This design enables it to capture complex patterns and long-range temporal relations of the speech context, producing more salient gesticulations in sync with the speech context.

Specifically, our model first encodes representations for each modality via multi-modal encoder networks. Then a cross-modal fusion network predicts co-speech gesture representations from the multi-modal representations.

Seed Gesture Encoder. The seed gesture encoder encodes seed gesture motion representations into latent representations to encourage consistent transition between seed gesture motions and the generated gestures. Given the DMR sequence of the seed gesture motion $\mathbf{C}_{\text{seed}} \in \mathbb{R}^{H \times K \times \frac{T_{\text{seed}}}{8}}$, the encoder first projects each frame of the DMR sequence into a D_{model}-dimensional space via a linear layer, where D_{model} is a hyper-parameter. We then inject sinusoidal positional encoding [13] into the projected frames to get $\mathbf{H}^0_{\text{seed}} \in \mathbb{R}^{D_{\text{model}} \times \frac{T_{\text{seed}}}{8}}$. Feeding $\mathbf{H}^0_{\text{seed}}$ through l encoder blocks, we get the seed gesture representation $\mathbf{H}^l_{\text{seed}}$.

Audio Encoder. The acoustic feature sequence $\mathbf{A} \in \mathbb{R}^{N \times T_{\text{gen}}}$ is the log-power Mel-spectrogram extracted from the speech audio.

We carefully set the window length and hop length so that the acoustic feature $\mathbf{A}$ and the target gesture motion latent sequence $\mathbf{X}_{\text{gen}}$ contain the same number of frames. Then we conduct the same projection and positional encoding injection on $\mathbf{A}$ to get $\mathbf{H}^0_{\text{audio}}$. Finally, after applying l encoder blocks, we obtain the audio representation, denoted as $\mathbf{H}^l_{\text{audio}} \in \mathbb{R}^{D_{\text{model}} \times T_{\text{gen}}}$.

Text Encoder. For the semantic features, we first apply a pretrained BERT [14] model released by Cui et al. [15] to get a feature vector for each word in the sentence and align them with the audio. After the alignment operation, the semantic feature sequence $\mathbf{S}$ will contain the same number of frames as $\mathbf{A}$. Then we conduct the same projection and positional encoding injection on $\mathbf{A}$ to get $\mathbf{H}^0_{\text{text}}$. After applying l encoder blocks, the resulted representation is denoted as $\mathbf{H}^l_{\text{text}} \in \mathbb{R}^{D_{\text{model}} \times T_{\text{gen}}}$.

Cross-Modal Fusion Network. Having obtained the encoded multi-modal context representations $\mathbf{H}^l_{\text{seed}}$, $\mathbf{H}^l_{\text{audio}}$ and $\mathbf{H}^l_{\text{text}}$, we predict the target DMR sequence $\mathbf{C}_{\text{gen}}$ with a cross-modal fusion network. The cross-modal fusion network is mainly made up of several decoder blocks. Each decoder block consists of a convolutional layer, a multi-head self-attention layer, and a multi-head cross-attention layer. We treat the sum of $\mathbf{H}^l_{\text{audio}}$ and $\mathbf{H}^l_{\text{text}}$ as the input sequence of the decoder block, while taking $\mathbf{H}^l_{\text{seed}}$ as the *query* of the cross-attention layer in the decoder block. In this way, the cross-modal fusion network can generate co-speech gestures closely related to the acoustic and semantic representations while maintaining a smooth and natural transition from the seed gesture sequence to the target gesture sequence.

Finally, the output of the cross-modal fusion network is transformed into categorical distribution probabilities using an MLP and a softmax function over each motion latent head, $\hat{\mathbf{C}}_{\text{gen}} \in \mathbb{R}^{H \times D \times \frac{T_{\text{gen}}}{8}}$. During training, we minimize the cross-entropy loss between $\hat{\mathbf{C}}_{\text{gen}}$ and the motion latent given by the latent encoder:

$$\mathcal{L}_{\text{cross}} = -\sum_{h=1}^{H} \sum_{t=1}^{\frac{T_{\text{gen}}}{8}} \log(\hat{\mathbf{C}}^{h,k,t}_{\text{gen}}), \tag{2}$$

where $k = \text{argmin}_j(||\mathbf{E}^{h,j} - \text{encode}(\mathbf{X}_{\text{gen}})^{h,t}||_2)$. During inference, we first sample DMR from the conditional categorical distributions given by MCGT. Afterward, we decode the gesture motions from the sampled DMR with the DMR decoder.

3. EXPERIMENT

3.1. Benchmark Building

Although previous works [4, 16, 7] have provided a few co-speech gesture datasets, there are limitations, such as lack of text transcriptions [4] or no sufficient salient gesture motions [7]. On the other hand, there is no currently widely-accepted metric for evaluating the salience of gestures. Observing these limitations, we build a new benchmark to evaluate the quality of salient motions produced by different methods.

Dataset. Our *LecGesture* dataset contains sufficient salient gesture motions. We construct our dataset from a famous Chinese lecture program[1] with the help of pose estimators [17, 18, 19] and a forced aligner [20]. In total, the *LecGesture* dataset contains 4,240 clips of lectures with 5 speakers for 24 hours.

Having obtained the dataset, we visualize the end-effector velocity and acceleration distribution in the *LecGesture* dataset and the *TED Gesture* [7] dataset. As shown in Figure 3, in the *TED Gesture* dataset, the majority of motion velocity samples are located in the low-speed ($0.0 \sim 0.03$) interval, while in our dataset, samples are mainly located in the medium-speed ($0.03 \sim 0.15$) interval. As the velocity and acceleration of most salient gesture motions are faster than those of gentle gesture motions, we can conclude from the distribution that the *Speec2Gesture* dataset mainly consists of gentle gesture motions, while our *LecGesture* dataset contains more salient gesticulations. In this work, we also conduct experiments on the *TED Gesture* [7] dataset for fair comparison.

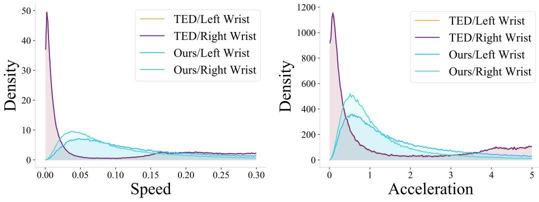

Fig. 3. Results of dataset observation.

Metrics. Although the Fréchet Gesture Distance (FGD) reflects the overall quality of generated co-speech gestures, it does not measure the pairwise similarity between each generated clip and the ground-truth clip. Previous works adopt different types of metrics to measure the pairwise similarity, such as mean velocity difference (MVD) [21] and mean absolute error of joint coordinates (MAEJ) [7]. However, because the co-speech gesture is highly non-deterministic and the speaker conducts gentle gesticulations most of the time, directly comparing the velocity or joint location frame by frame does not reflect the similarity. For example, as we show in our experimental results, even a static mean-pose baseline outperforms the SOTA in terms of MVD. Observing the limitations of existing metrics, we propose a new metric, salient motion similarity (SMS), to measure the quality of salient motions. Specifically, given a gesture sequence $\mathbf{X} \in \mathbb{R}^{M \times T_{\text{gen}}}$, we denote the j-th joint's velocity in the k-th frame as $\mathbf{V}^{j,k} \in \mathbf{R}^3$. The SMS between the ground-truth clip $\mathbf{X}$ and the generated clip $\mathbf{X}_{\text{gen}}$ is defined as:

$$\text{SMS}(\mathbf{X}, \mathbf{X}_{\text{gen}}) = \frac{1}{||\mathcal{J}||} \sum_{j \in \mathcal{J}} \max_{m \in [k-w, k+w]}(\frac{\mathbf{V}^{j,k} \cdot \mathbf{V}^{j,m}_{\text{gen}}}{||\mathbf{V}^{j,k}||^2_2}), \tag{3}$$

where $k = \text{argmax}_p(||\mathbf{V}^{j,p}||_2)$, $w = 8$, $\mathbf{V}_{\text{gen}}$ represents the velocity of generated clips, and $\mathcal{J}$ represents the set of end-effectors. We measure the velocity similarity in the time window $[t-w, t+w]$ because we do not expect motions exactly the same as the ground-truth. We also use FGD to evaluate the overall gesture quality.

We train our model on the proposed benchmark. We set $D_{\text{model}} = 128$, $l = 7$, $T_{\text{seed}} = 64$ and $T_{\text{gen}} = 128$. We use the Adam optimizer [22] with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$. We train the DMR auto-encoder for 950k steps and the MCGT for 420k steps with a batch size of 64 and a learning rate of 10^{-4} until convergence. We select 3816 clips from the *LecGesture* dataset for training and hold out 424 for testing.

3.2. Quantitative Results

Table 1 shows the comparison results between the proposed MCGT and the baselines. To measure the overall gesture quality, we first calculate the FGD between the generated gesture sequences and the ground-truth sequences. FGD measures how close the distribution of generated clips is to the real ones. The experiment results demonstrate that our approach achieves comparable FGD with Trimodal [7] and outperforms the other baselines. The FGD values confirm that our approach generates overall realistic gesture motions.

[1] https://tv.cctv.com/lm/bjjt/

Table 1. Objective comparison of different methods and the ablated model. The *Mean Pose* baseline outputs the mean pose of the training set regardless of the input.

Methods	*LecGesture* dataset			*TED* Dataset	
	FGD↓	SMS↑	MVD↓	FGD↓	SMS↑
Trimodal [7]	0.042	0.258	0.101	3.729	0.267
S2G [4]	0.524	0.218	0.058	19.254	0.224
A2B [23]	0.580	0.382	0.074	22.861	0.373
Mean Pose	4.214	0	0.080	182.712	0
Ours	0.061	**0.403**	0.124	3.942	0.391
Ours w/o DMR	**0.011**	0.303	0.108	3.451	0.315

Table 2. The mean opinion scores (MOS) of different methods, with 95% confidence interval.

Methods	Expressiveness	Content Matching
Ground-Truth	4.35 ± 0.32	4.32 ± 0.52
S2G [4]	2.05 ± 0.46	2.01 ± 0.60
A2B [23]	2.66 ± 0.48	2.86 ± 0.47
Trimodal [7]	3.21 ± 0.28	3.25 ± 0.30
Ours	$\mathbf{3.55 \pm 0.31}$	$\mathbf{3.62 \pm 0.29}$
Ours w/o DMR	3.38 ± 0.35	3.46 ± 0.32

We then measure the pairwise similarity between each generated sequence and the ground truth. As the co-speech gesture is highly non-deterministic, we do not expect models to generate the same motions as the ground truth. Instead, we measure the velocity differences between different methods. We first calculate the mean velocity difference (MVD), which measures the velocity difference frame by frame. We argue that this frame-by-frame difference makes MVD unsuitable for evaluating co-speech gestures. Because the speaker conducts gentle gesticulations most of the time, predicting a static mean pose could achieve lower MVD. As shown in Table 1, even the Mean Pose baseline achieves lower MVD than Trimodal [7]. Audio2Body [23] and Speech2Gesture [4] also reach lower MVD even though they exhibit poor FGD. Observing this limitation, we further calculate salient motion similarity (SMS) for our approach and the baseline methods. Compared with Trimodal [7] and Speech2Gesture [4], our approach generates co-speech gesture motions with higher SMS. Although motions generated by Audio2Body [23] also exhibit high SMS, they show noisy dynamics and poor temporal continuity, as indicated by the high FGD value.

3.3. Qualitative Results

Figure 4 shows qualitative results generated from one sample in the *LecGesture* dataset. Speech2Gesture [4] tends to generate only slow and gentle gesture motions. Motions generated by Audio2Body [23] exhibit noisy dynamics but generates no salient motions because it is difficult to directly translate audio to co-speech gesture motions with a single-layer RNN. Trimodal [7] shows high overall gesture quality but suffers from the over-smooth problem because it assumes a deterministic mapping from the speech to the gesture. Meanwhile, equipped with the discrete motion representation, our approach generates more high-quality salient gesticulations

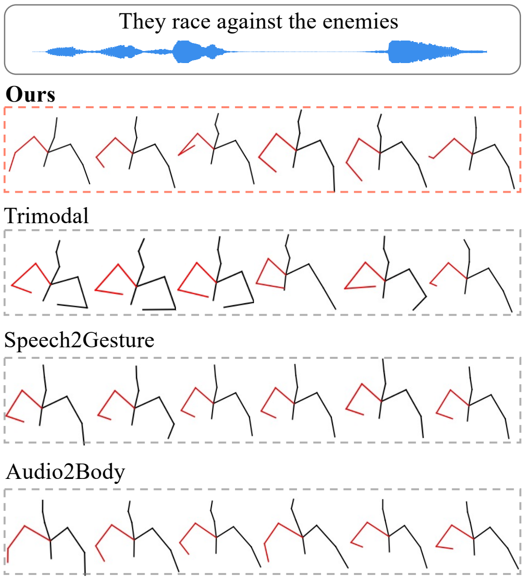

Fig. 4. Co-speech gestures generated by different methods.

3.4. Subjective Evaluation

To compare the perceptual quality of the co-speech gesture generated by different methods, we conduct a user study. We ask participants to rate the mean opinion score (MOS) for the generated gestures in 1∼5 (higher scores denote better quality). Specifically, for each video, the participants rate (1) the expressiveness of the co-speech gestures and (2) the content matching degree between the co-speech gestures and the speech content.

We recruited 17 participants. Each participant rated 20 videos. The results of the user study are shown in Table 2. Our approach achieves higher MOS both in expressiveness and content matching, outperforming Trimodal [7] by 0.34 in terms of expressiveness and 0.37 in terms of content matching. The results of the user study confirm that compared with baselines, our model generates co-speech gestures with better perceptual quality.

3.5. Ablation Study

In order to understand the function of the DMR space in our model, we conduct an ablation study. Specifically, we remove the DMR space from our approach and train the MCGT to regress the joint coordinates in co-speech gesture motions directly. As shown in Table 1 and Table 2, although removing DMR space leads to lower FGD, it decreases SMS and perceptual quality. We also notice that the ablated model outperforms Trimodal [7] in terms of both FGD and SMS, confirming that the convolutional architecture with temporal attention improves both overall gesture quality and gesture salience.

4. CONCLUSION

This paper proposes a novel co-speech gesture generation approach to improve the gesture salience in co-speech gesture synthesis. Our approach adopts the discrete motion representation to bridge the speech-gesture mapping stage and the gesture generation stage. To better evaluate the effectiveness of our approach, we construct a new benchmark for evaluating salient motion quality in co-speech gesture synthesis. Experiments demonstrate that our approach produces high-quality salient gesticulations while keeping the overall gesture motion realistic.

5. ACKNOWLEDGEMENTS

This work is supported by the National Key R&D Program of China under Grant No.2021QY1500, and the state key program of the National Natural Science Foundation of China (NSFC) (No.61831022).

6. REFERENCES

[1] Robert M Krauss, Yihsiu Chen, and Rebecca F Gotfexnum, “Lexical gestures and lexical access: A process model,” *Language and gesture*, vol. 2, pp. 261, 2000.

[2] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy W. Bickmore, “BEAT: the behavior expression animation toolkit,” in *the 28th Annual Conference on Computer Graphics and Interactive Techniques*, Lynn Pocock, Ed. 2001, pp. 477–486, ACM.

[3] Chien-Ming Huang and Bilge Mutlu, “Robot behavior toolkit: Generating effective social behaviors for robots,” in *International Conference on Human-Robot Interaction*. 2012, pp. 25–32, ACM.

[4] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik, “Learning individual styles of conversational gesture,” in *IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3497–3506, Computer Vision Foundation / IEEE.

[5] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow, “Style-controllable speech-driven gesture synthesis using normalising flows,” *Comput. Graph. Forum*, vol. 39, no. 2, pp. 487–496, 2020.

[6] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström, “Gesticulator: A framework for semantically-aware speech-driven gesture generation,” in *International Conference on Multimodal Interaction*. 2020, pp. 242–250, ACM.

[7] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee, “Speech gesture generation from the trimodal context of text, audio, and speaker identity,” *ACM Trans. Graph.*, vol. 39, no. 6, pp. 222:1–222:16, 2020.

[8] JinHong Lu, Tianhang Liu, Shuzhuang Xu, and Hiroshi Shimodaira, “Double-dcccae: Estimation of body gestures from speech waveform,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2021, pp. 900–904, IEEE.

[9] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao, “Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders,” in *IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11273–11282, IEEE.

[10] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, “Neural discrete representation learning,” in *Annual Conference on Neural Information Processing Systems*, 2017, pp. 6306–6315.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778, IEEE Computer Society.

[12] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh, “Meshtalk: 3d face animation from speech using cross-modality disentanglement,” in *IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1153–1162, IEEE.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Annual Conference on Neural Information Processing Systems*, 2017, pp. 5998–6008.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2019, pp. 4171–4186, ACL.

[15] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu, “Revisiting pre-trained models for chinese natural language processing,” in *Findings of the Association for Computational Linguistics: EMNLP*. 2020, vol. EMNLP 2020, pp. 657–668, ACL.

[16] Ylva Ferstl and Rachel McDonnell, “Investigating the use of recurrent motion modelling for speech gesture generation,” in *International Conference on Intelligent Virtual Agents*. 2018, pp. 93–98, ACM.

[17] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li, “Faceboxes: A CPU real-time face detector with high accuracy,” in *IEEE International Joint Conference on Biometrics*. 2017, pp. 1–9, IEEE.

[18] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu, “RMPE: regional multi-person pose estimation,” in *IEEE International Conference on Computer Vision*. 2017, pp. 2353–2362, IEEE Computer Society.

[19] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli, “3d human pose estimation in video with temporal convolutions and semi-supervised training,” in *IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 7753–7762, Computer Vision Foundation / IEEE.

[20] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using kaldi,” in *Annual Conference of the International Speech Communication Association*. 2017, pp. 498–502, ISCA.

[21] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li, “Makeittalk: Speaker-aware talking-head animation,” *ACM Trans. Graph.*, vol. 39, no. 6, pp. 221:1–221:15, 2020.

[22] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[23] Eli Shlizerman, Lucio M. Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman, “Audio to body dynamics,” in *IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7574–7583, Computer Vision Foundation / IEEE Computer Society.