Speech Drives Templates: Co-Speech Gesture Synthesis with Learned Templates

Shenhan Qian*[1] Zhi Tu*[1] Yihao Zhi*[1] Wen Liu[1] Shenghua Gao†[1,2,3]

*Equal contribution †Corresponding author

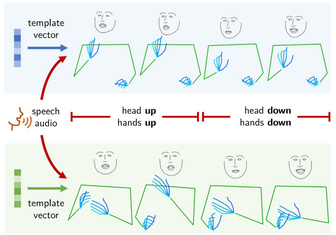
[1]ShanghaiTech University [2]Shanghai Engineering Research Center of Intelligent Vision and Imaging [3]Shanghai Engineering Research Center of Energy Efficient and Custom AI IC

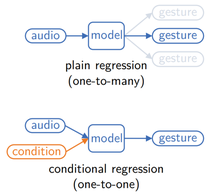
2021 ICCV OCTOBER 11-17 VIRTUAL

Code

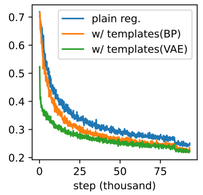
Introduction

Co-speech gesture synthesis





(a) Plain regression versus conditional regression.



(b) Regression loss curves on the training set.

Challenge: the mapping from speech audio to feasible gestures are not unique, leading to underfitting.

Solution: we relieve the ambiguity by adding conditions, *i.e.,* the template vectors in this paper. These template vectors are learned from the data and used to generate diverse gesture sequences. Further exploration of the template vectors validates the essence of the one-to-many mapping and unveils the roles of speech audio and our learned templates.

Contributions

- An audio-driven gesture synthesis model, with the learning of **template vectors**, relieving the ambiguity of co-speech gesture synthesis, enhancing the fidelity and variety without sacrificing synchronization quality;
- Two metrics for audio-driven gesture evaluation, **lip-sync error** as a proxy metric to evaluation gesture-syncing, and **Fréchet Template Distance (FTD)** to assess gesture fidelity and variety;
- Achieve superior performance on **Speech2Gesture** dataset and **two speakers we collect**;

Evaluation

Two metrics for audio-driven gesture evaluation

Lip-sync error:

$$\mathcal{E}_{\text{lip}} = \frac{\frac{1}{F}\sum_{i=1}^{F} \|d^{(i)} - \hat{d}^{(i)}\|_2}{\max_{1\le n\le F} \hat{d}^{(n)}}$$

where $d^{(i)}$ is the distance between the center keypoints of upper and lower lip in the i-th frame of the generated gesture sequence $\mathbf{G}$, and $\hat{d}^{(i)}$ is the corresponding distance for ground-truth gesture sequence $\hat{\mathbf{G}}$.

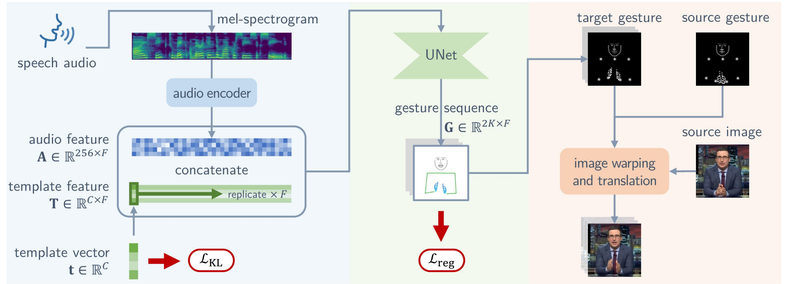
Fréchet Template Distance (FTD):

$$\text{FTD} = |\mu_{\mathbf{t}} - \mu_{\hat{\mathbf{t}}}|^2 + \text{tr}\left(\Sigma_{\mathbf{t}} + \Sigma_{\hat{\mathbf{t}}} - 2\left(\Sigma_{\mathbf{t}}\Sigma_{\hat{\mathbf{t}}}\right)^{1/2}\right)$$

where $\mu_{\mathbf{t}}$ and $\Sigma_{\mathbf{t}}$ are the mean vector and covariance matrix of the template vectors $[\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_N]$, encoded from synthesized gesture sequences $[\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_N]$ across the test set with the VAE.

Approach

Pipeline



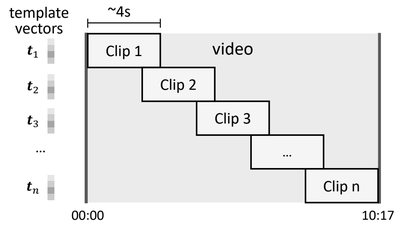
$$\mathcal{L} = \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} \qquad \mathcal{L}_{\text{reg}} = \frac{1}{F}\sum_{i=1}^{F}\|\mathbf{G}^{(i)} - \hat{\mathbf{G}}^{(i)}\|_1 \qquad \mathcal{L}_{\text{KL}} = D_{\text{KL}}\left(\mathcal{N}\left(\mu_{\mathbf{t}}, \sigma_{\mathbf{t}}^2\right) \| \mathcal{N}\left(0, 1\right)\right)$$

Template vector learning

Assignment: Each clip in the dataset is paired with a template vector.

Learning: We provide two ways to learn the template vectors:

1. template-BP: optimize template vectors with the back-propagated gradients of the regression loss.
2. template-VAE: train a VAE to reconstruct all gesture clips and take the encoding of each clip as its template vector.

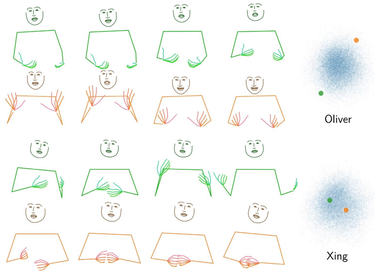


Examples of synthesized images

Two English speakers from the Speech2Gesture dataset and two Mandarin speakers from the Internet.

Experiments

Visualization of template gestures



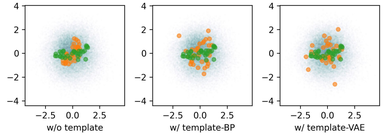
For a better interpretation of the template space, we sample characteristic template vectors (green) and feed them into the decoder of our trained VAE to obtain a gesture sequence. The gestures of opposite template vectors exhibit clear semantic symmetry including head orientations, arm poses, and hand behaviors.

Different settings of template learning

	template type	$\mathcal{E}_{\text{lip}}\downarrow$	FTD ↓
w/o template	-	**0.17**	1.66
w/ template-BP	frame-wise	0.21	0.78
w/ template-BP	clip-wise	**0.17**	1.26
w/ template-VAE	clip-wise	**0.17**	**0.92**

Our method with clip-wise template vectors achieve the best balance between variety (lower FTD) and synchronization(lower lip-sync error).

Visualize gestures in the template space



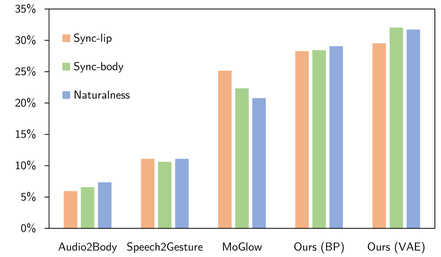
We plot the encoding of ground-truth (green) and generated (orange) gestures with the encoder of the trained VAE. Our results from template vectors span a larger space, indicating higher diversity.

Comparison with baselines on two English speakers and two Mandarin speakers

Our results achieve lower FTD (higher variety and fidelity) while maintaining a low lip-sync error. Higher L_2 distances of our results also indicate a larger diversity.

	Oliver			Kubinec			Xing			Luo		
	L_2 dist.	$\mathcal{E}_{\text{lip}}\downarrow$	FTD ↓	L_2 dist.	$\mathcal{E}_{\text{lip}}\downarrow$	FTD ↓	L_2 dist.	$\mathcal{E}_{\text{lip}}\downarrow$	FTD ↓	L_2 dist.	$\mathcal{E}_{\text{lip}}\downarrow$	FTD ↓
Audio to Body [25]	49.7	0.19	3.48	70.9	0.17	4.51	50.9	0.18	4.75	48.4	**0.16**	2.70
Speech2Gesture [11]	53.5	0.23	8.30	64.9	0.20	4.53	48.0	0.19	4.49	63.7	0.20	3.10
MoGlow [2]	50.6	0.20	2.28	78.1	0.16	2.49	48.4	0.18	4.94	54.8	0.18	1.47
Ours (w/ template-BP)	50.6	**0.17**	1.26	83.7	**0.15**	1.98	50.0	**0.17**	2.72	51.5	**0.16**	1.21
Ours (w/ template-VAE)	62.4	**0.17**	**0.92**	100.7	**0.15**	**1.07**	57.8	0.18	**1.72**	80.8	0.17	**0.69**

Human study



Ablation study on the normalization operation and body representation

	Hierarchical	$\mathcal{E}_{\text{lip}}\downarrow$	FTD ↓
BN		0.20	7.01
IN		0.19	1.54
IN*		0.19	**1.53**
IN*	✓	**0.17**	1.66

Hierarchical means that gesture keypoints are split into four parts, each part rebased to its local root node. **IN*** denotes instance normalization on the dimension of channels.