QPGesture: Quantization-Based and Phase-Guided Motion Matching for Natural Speech-Driven Gesture Generation

Sicheng Yang[1], Zhiyong Wu[1,4], Minglei Li[2], Zhensong Zhang[3], Lei Hao[3], Weihong Bao[1], Haolin Zhuang[1]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, China
[2] Huawei Cloud Computing Technologies Co., Ltd, China [3] Huawei Noah's Ark Lab, China
[4] The Chinese University of Hong Kong, Hong Kong SAR, China

Paper ID: 7523

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

1. Introduction

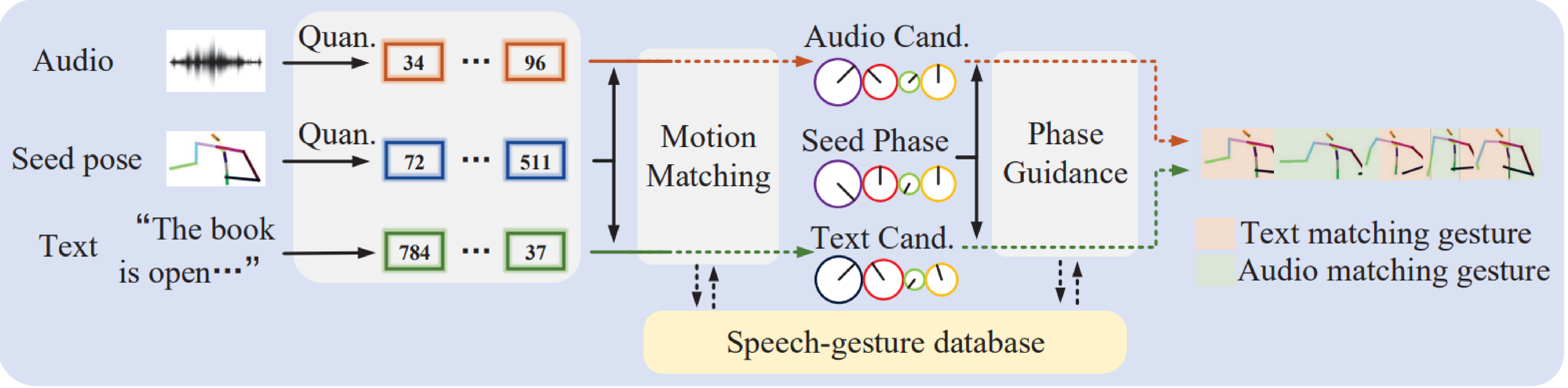
1.1 Motivation

- Problems :
  - Random jittering
  - Inherent asynchronicity with speech
- Goal:
  - ✓ Solve jittering problems, such as grabbing hands or pushing glasses
  - ✓ Better alignment of speech and gestures
  - ✓ Further improve the quality of gesture generation

1.2 Contribution

- Propose a novel quantization-based motion matching framework for speech-driven gesture generation
  - ✓ Address random jittering
  - ✓ Align diverse gestures with different speech using Levenshtein distance. Solve the issue of speech and gesture asynchrony and motion matching model inflexibility
  - ✓ A phase guidance strategy to select optimal audio and text candidates
  - ✓ Extensive experiments show that jittering and asynchronicity issues can be effectively alleviated by our framework

2. Model Architecture



3. Methodology

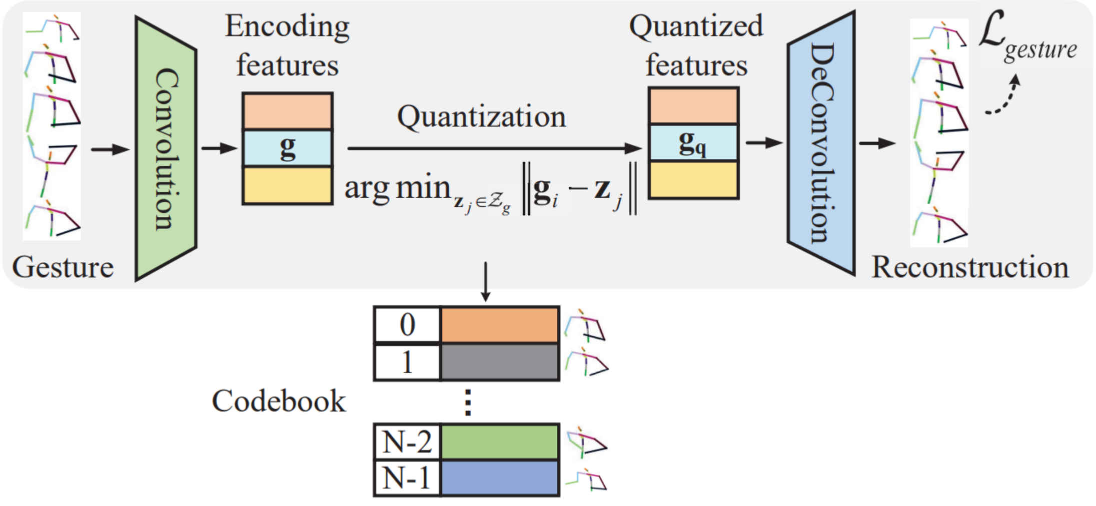
3.1 Learning a discrete latent space representation

- vq-wav2vec
- Gesture VQ-VAE
  - Encode the joint sequence $\mathbf{G}$

$$\mathbf{g} = E_g(\mathbf{G})$$

  - Decoder

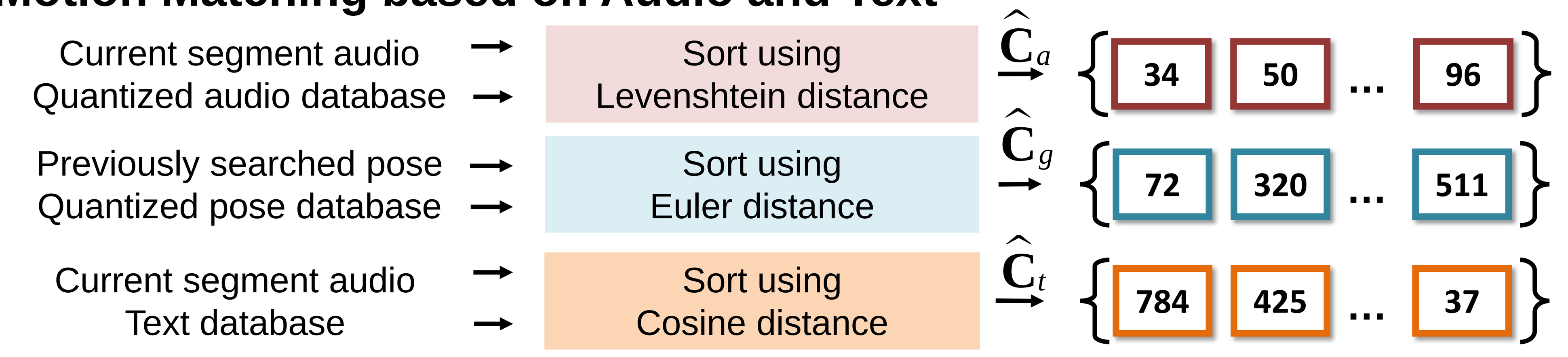
$$\widehat{\mathbf{G}}_1 = D_g\left(\mathbf{g}_q\right) = D_g(\mathbf{q}(E_g(\mathbf{G})))$$



  - The encoder, decoder and codebook can be trained by optimizing:

$$\mathcal{L}_{gesture(E_g, D_g, \mathcal{Z}_g)} = \|\widehat{\mathbf{G}}_1 - \mathbf{G}_1\|_1 + \alpha_1 \left\|\widehat{\mathbf{G}}_1' - \mathbf{G}_1'\right\|_1 + \alpha_2 \left\|\widehat{\mathbf{G}}_1'' - \mathbf{G}_1''\right\|_1 + \left\|\mathrm{sg}[\mathbf{g}] - \mathbf{g_q}\right\| + \beta\left\|\mathbf{g} - \mathrm{sg}\left[\mathbf{g_q}\right]\right\|$$

3.2 Motion Matching based on Audio and Text



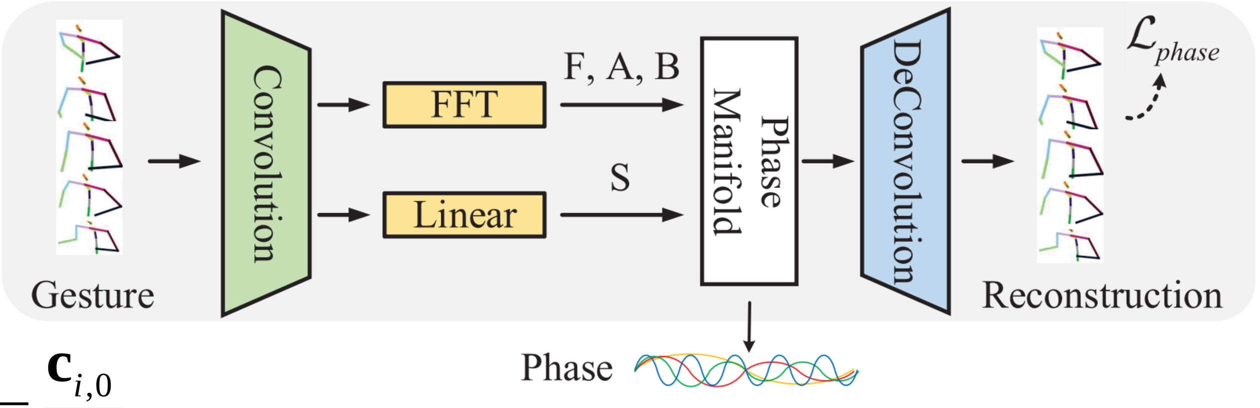
$\widehat{\mathbf{C}}_a + \widehat{\mathbf{C}}_g \xrightarrow[\text{weighting}]{\text{Ranking}}$ Audio candidate $\mathbf{C}_a$ $\quad$ $\widehat{\mathbf{C}}_t + \widehat{\mathbf{C}}_g \xrightarrow[\text{weighting}]{\text{Ranking}}$ Text candidate $\mathbf{C}_t$

3.3 Phase-Guided Gesture Generation

- Encode the joint sequence

$$\mathbf{L} = E_p(\mathbf{G})$$

- Periodic parameters



$$\mathbf{A}_i = \sqrt{\frac{2}{T}\sum_{j=1}^{K}\mathbf{p}_{i,j}}, \quad \mathbf{F}_i = \frac{\sum_{j=1}^{K}\left(\mathbf{f}_j \cdot \mathbf{p}_{i,j}\right)}{\sum_{j=1}^{K}\mathbf{p}_{i,j}}, \quad \mathbf{B}_i = \frac{\mathbf{c}_{i,0}}{T},$$

$$\left(s_x, s_y\right) = FC\left(\mathbf{L}_i\right), \quad \mathbf{S}_i = \mathrm{atan}\,2\left(s_y, s_x\right)$$

$$\hat{\mathbf{L}} = f(\mathcal{T}; \mathbf{A}, \mathbf{F}, \mathbf{B}, \mathbf{S}) = \mathbf{A} \cdot \sin(2\pi \cdot (\mathbf{F} \cdot \mathcal{T} - \mathrm{S})) + \mathbf{B}$$

- Loss Function

$$\mathcal{L}_{phase} = \mathcal{L}_{phase-recon}(\mathbf{G}, h(\widehat{\mathbf{L}}))$$

4. Experiments

4.1 Dataset

- BEAT dataset; 15 joints corresponding to the upper body
- 8:1:1 by training, validation, and testing

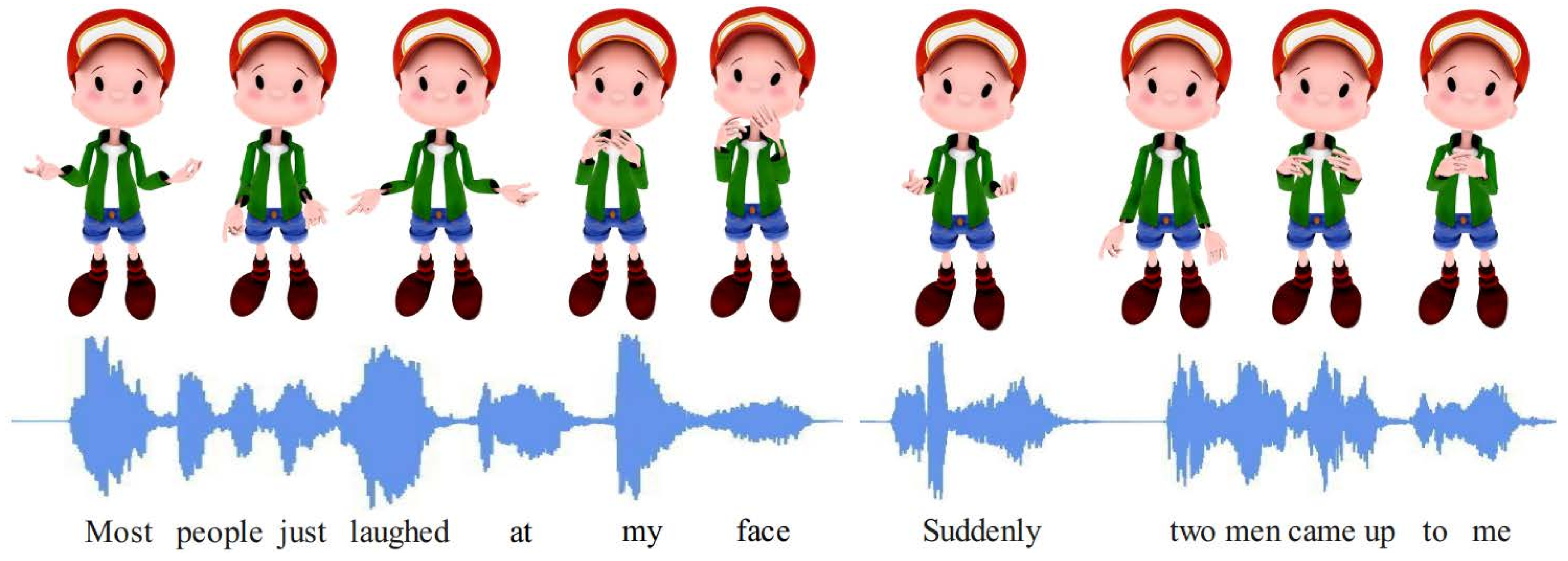
4.2 Comparison to Existing Methods

Name	Objective evaluation			Subjective evaluation	
	Hellinger distance average ↓	FGD on feature space ↓	FGD on raw data space ↓	Human-likeness	Appropriateness
Ground Truth (GT)	0.0	0.0	0.0	3.79 ± 0.19	3.62 ± 0.21
End2End [47]	0.146	64.990	16739.978	3.64 ± 0.11	3.23 ± 0.14
Trimodal [46]	0.155	48.322	12869.98	3.31 ± 0.17	3.20 ± 0.19
StyleGestures [5]	**0.136**	35.842	9846.927	3.66 ± 0.08	3.30 ± 0.11
KNN [17]	0.364	43.030	12470.061	2.38 ± 0.10	2.35 ± 0.13
CaMN [31]	0.149	52.496	10549.455	3.65 ± 0.16	3.29 ± 0.15
Ours	**0.136**	**19.921**	**5742.281**	**4.00 ± 0.14**	**3.66 ± 0.23**

4.3 Ablation Studies

Name	Objective evaluation			Subjective evaluation	
	Hellinger distance average ↓	FGD on feature space ↓	FGD on raw data space ↓	Human-likeness	Appropriateness
w/o wavvq + WavLM	0.151	19.943	6009.859	3.87 ± 0.21	3.64 ± 0.21
w/o audio	0.134	20.401	5871.044	3.87 ± 0.21	3.63 ± 0.20
w/o text	**0.118**	23.929	6389.866	3.57 ± 0.29	3.41 ± 0.23
w/o phase	0.138	**19.195**	5759.167	3.90 ± 0.11	3.65 ± 0.17
w/o motion matching (GRU + codebook)	0.140	30.404	11642.641	3.78 ± 0.14	3.43 ± 0.16
Ours	0.136	19.921	**5742.281**	**4.07 ± 0.15**	**3.77 ± 0.21**

Reference



Project page

[1] *Bailando: 3D Dance Generation by Actor-Critic GPT with Choreographic Memory.*
[2] *A Motion Matching-based Framework for Controllable Gesture Synthesis from Speech.*
[3] *DeepPhase: periodic autoencoders for learning motion phase manifolds.*