Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation
Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, Bolei Zhou

Co-Speech Gesture Motivation

- **Task Definition**: Given a clip of speech information as input, we predict the 3D skeleton sequence that is aligned with the speech.
- **Key Observations**:
  1. Different co-speech gestures are related to distinct levels of audio. For example, the metaphorical gestures are associated with high-level speech semantics (e.g., when depicting a ravine, one would move two outstretched hands apart and saying gap”), while the low-level audio features of beat and volume lead to the rhythmic gestures.
  2. The dynamic patterns of different body parts are not the same, such as the flexible fingers and relatively still upper arms. Instead of holistically generating the whole skeleton, we should treat each part differently.

Our solution: Capture hierarchical audio-pose associations!

Framework

Overview

Key Components

- **Hierarchical Audio Learner**: Encode multi-level audio features to extract both rhythmic and semantic information.
- **Hierarchical Pose Inferer**: Infer gestures hierarchically and capture associations between multi-level audios and poses.

Our Approach

- **Hierarchical Audio Feature Extraction**: We take output of shallow, middle, deep encoder layers as low, middle, high level features \( f_{\text{low}}, f_{\text{mid}}, f_{\text{high}} \).
- **Contrastive Learning Strategy**: We take text feature \( f_t \) with high-level feature \( f_{\text{high}} \) as positive pairs; with low/mid level \( f_{\text{low}}, f_{\text{mid}} \) as negative pairs:

\[
\mathcal{L} = \mathbb{E}_{t \sim \mathcal{D}} \left[ -\log \frac{e^{\mathcal{S}(t, E_t)}}{e^{\mathcal{S}(t, E_t)} + e^{\mathcal{S}(t, E_{t\prime})}} \right],
\]

where \( \mathcal{S}(t, E_t) = \langle\mathbf{a}, E_t\rangle \) and \( \mathbf{a} = \mathcal{A}(t, \mathbf{a}) \).

- **Multi-Level Feature Blending**: Style coordinator \( \mathcal{C} \in \mathbb{R}^{3 \times l} \) controls ratio between hierarchical audio features and each level of motion hierarchy.

\[
f_h = \mathcal{C} \cdot \rho_h = \mathcal{C}[1, h] \cdot f_{\text{low}} + \mathcal{C}[2, h] \cdot f_{\text{mid}} + \mathcal{C}[3, h] \cdot f_{\text{high}} + \beta_h = \rho_h + \mathcal{W}_h \beta_h,
\]

where \( \rho_h = \mathcal{GRU}(\rho_{h-1}, \mathcal{A}(t, \mathcal{A}(t, \rho_{h-1})) \).

- **Coarse-to-Fine Pose Generation**: We design a level \( l = 6 \) level body hierarchy and predict from previous level’s inferred pose and current level’s audio. In this way, fine-grained gesture is learned in a coarse-to-fine manner.

Experiments

**Quantitative Comparisons**

<table>
<thead>
<tr>
<th>Method</th>
<th>TED Gesture</th>
<th>TED Expressive</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGD</td>
<td>6.588</td>
<td>27.122</td>
</tr>
<tr>
<td>BC</td>
<td>0.704</td>
<td>0.682</td>
</tr>
<tr>
<td>Diversity</td>
<td>171.482</td>
<td>168.223</td>
</tr>
</tbody>
</table>

**Ablation Study**

<table>
<thead>
<tr>
<th>Settings</th>
<th>TED Gesture</th>
<th>TED Expressive</th>
</tr>
</thead>
<tbody>
<tr>
<td>holistic w/o hand w/o body same ( f_h )</td>
<td>FGD 11.989</td>
<td>10.832</td>
</tr>
<tr>
<td>ASR HAZG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC</td>
<td>0.594</td>
<td>0.606</td>
</tr>
<tr>
<td>Diversity</td>
<td>156.079</td>
<td>158.823</td>
</tr>
</tbody>
</table>

**Conclusion with Github, Project Page**

In this paper, we propose a novel framework HAZG with Hierarchical Audio Learner and Hierarchical Pose Inferer for fine-grained co-speech gesture generation.