Supplementary Material: A Motion Matching-based Framework for Controllable Gesture Synthesis from Speech

IKHSANUL HABIBIE, MOHAMED ELGHARIB, and KRIPASHINDU SARKAR, MPI Informatics, Germany
AHSAN ABDULLAH, SIMBARASHE NYATSANGA, and MICHAEL NEFF, UC Davis, USA
CHRISTIAN THEOBALT, MPI Informatics, Germany

1 PROPOSED K-NN ALGORITHM

Additional details on constructing the audio and pose features.
As described in the main document, we use audio feature \( f \) and pose feature \( p \) to measure feature similarity between the input sequence at frame \( t \) and the sequences in the Matching Database. Each audio feature at frame \( t \) stores the relevant audio information at time \( t, t+2, ..., t+14 \) (at 15 fps). For the audio, we use the first 13 coefficients of the Mel-frequency cepstral coefficients (MFCC) as well as the log mean energy of the input audio \( m_0 \in \mathbb{R}^{112} \) (112 = 14 features x 8 frames). Similarly, every pose feature frame at time \( t \) stores information about the 3D pose coordinate of both left and right wrists, elbows, index finger root, and little finger root at time \( t, t+2, t+4, \) and \( t+6 \). We combine the 3D positions of all 8 joints into a single vector \( p_t \in \mathbb{R}^{96} \) (96 = 2 hands x 4 joints x 3 dims x 4 frames). This is equivalent to storing the hand trajectory within the next 0.5 seconds. The output pose at frame \( t \) contains the 3D coordinate of 13 body joints (pelvis-relative), and 21 joints for each hand, which we combine into a 1-D vector \( g_t \in \mathbb{R}^{165} \) (165 = 55 joints x 3 dims).

Pseudocode. A more detailed and procedural description of our proposed speech-gesture k-NN approach is shown in Algorithm 1.

2 NETWORK ARCHITECTURE

Figure 1 shows a detailed representation of the cGAN-based resynchronization network architecture. We use a similar architecture to Habibie et al. [2021] with a modification to accommodate the input motion from the k-NN. The input channel of the first layer of the generator consists of 193 (165 parameters for body+hand and 28 parameters for audio) instead of only 28 parameters. Since it does not predict facial expressions, the cGAN uses only one decoder which produces 165 parameters as output. The incorporation of a motion matching precursor has not been previously explored.

We employ a standard WGAN-GP formulation to train the method. To this end, we also remove the last sigmoid layer of the discriminator. The generator is updated after every 5 iterations to ensure that the average of the combined real and fake critic training curve fluctuates around 0.

3 ADDITIONAL EVALUATION OF CONTROL QUALITY

Figure 2 shows additional behavior of the synthesis methods under various types of control signals. Similar to the results mentioned in the main document, our method can generally follow the given input constraint. Our method produce a higher variation over the output sequence compared to MoGlow, which is crucial to improve the realism of the synthesis quality. The qualitative comparison shown

**Algorithm 1**: audio-to-gesture k-NN search

Data:
- list of audio feat. sequence \( F = [\mathbf{f}_0, \mathbf{f}_1, ..., \mathbf{F}_M] \)
- list of pose feat. sequence \( P = [\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{P}_M] \)
- list of gesture sequence \( G = [\mathbf{g}_0, \mathbf{g}_1, ..., \mathbf{G}_M] \)
- \( \mathbf{F} = [\mathbf{f}_0, \mathbf{f}_1, ..., \mathbf{f}_{T_{match}-1}] \)
- \( \mathbf{P} = [\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_{T_{match}-1}] \)
- \( \mathbf{G} = [\mathbf{g}_0, \mathbf{g}_1, ..., \mathbf{g}_{T_{match}-1}] \)

Input: \( k \in \mathbb{Z} \), the desired k-best neighbors, audio feat. sequence \( F = [\mathbf{f}_0, \mathbf{f}_1, ..., \mathbf{F}_T] \), control \( C = [c_0, c_1, ..., c_{T-1}] \), initial pose feat. \( \mathbf{p}_{r-1} \), \( f \in \mathbb{R}^{112}, \mathbf{p} \in \mathbb{R}^{96} \), \( c \in \{0, 1\} \)

Output: gesture sequence \( G = [\mathbf{g}_0, \mathbf{g}_1, ..., \mathbf{g}_T] \)

\( \mathbf{g} \in \mathbb{R}^{165} \)

\( t = 0 \)

initialize \( G = [], P = [\mathbf{p}_{r-1}] \)

while \( t < T \) do

\( \mathbf{P} = [], \mathbf{F} = [], \mathbf{G} = [] \)

for \( m = 0 \) to \( M - 1 \) do

\( r = 0 \)

\( \text{pdist} = \infty \)

for \( s = 1 \) to \( T_{match}-1 \) do

if \( c_s == 1 \)

if \( d(\hat{\mathbf{p}}^m_s, \mathbf{p}_{r-1}) < \text{pdist} \) then

\( \text{pdist} = d(\hat{\mathbf{p}}^m_s, \mathbf{p}_{r-1}) \)

\( r = s \)

end

end

append(\( \mathbf{P}, \hat{\mathbf{p}}^m_{r(\mathbf{p}_{r-1})} \))

append(\( \mathbf{F}, \mathbf{f}_{T_{match}-1} \))

append(\( \mathbf{G}, \mathbf{g}_{T_{match}-1} \))

end

\( \hat{\mathbf{P}} = [\mathbf{p}_0(0, N-1), \mathbf{p}_1(0, N-1), ..., \mathbf{P}_M(0, N-1)] \)

\( \hat{\mathbf{F}} = [\mathbf{f}_0(0, N-1), \mathbf{f}_1(0, N-1), ..., \mathbf{F}_M(0, N-1)] \)

\( \hat{\mathbf{G}} = [\mathbf{g}_0(0, N-1), \mathbf{g}_1(0, N-1), ..., \mathbf{G}_M(0, N-1)] \)

\( R_{audio} = \text{relrank}\{d(\mathbf{f}_t, \mathbf{f}_j), d(\mathbf{f}_j, \mathbf{f}_k), \ldots\} \)

\( R_{pose} = \text{relrank}\{d(\mathbf{p}_t, \mathbf{p}_j), d(\mathbf{p}_j, \mathbf{p}_k), \ldots\} \)

\( R_{combined} = R_{audio} + R_{pose} \) (elem. wise)

sort \( R_{combined} \); sort its indices into \( i_{combined} \)

\( i = i_{combined}(k) \)

append(\( \mathbf{G}, \mathbf{g}_i(0, N-1) \))

append(\( \hat{\mathbf{P}}, \mathbf{p}_i(0, N-1) \))

\( t = t + N \)
in Figure 3 demonstrates the efficacy of our methods to follow the provided control input.

Table 1. Summary of the search database for the “Oliver” sequences. The data is recorded from an “in-the-wild” setting, and thus contain various types of speech gestures unseen in other studio-captured dataset.

<table>
<thead>
<tr>
<th>Total duration</th>
<th>11.4 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total unique videos</td>
<td>105 videos</td>
</tr>
<tr>
<td>Total unique clips</td>
<td>9624 clips</td>
</tr>
<tr>
<td>Duration per clip</td>
<td>64 frames @ 15 fps</td>
</tr>
</tbody>
</table>

5 USER STUDY

Before each study, each user is presented with a set of instructions describing the objective and procedure of the experiment. Below is an excerpt of the instruction page in one of our studies:

"The purpose of this user study is to measure the synthesis quality of AI-based methods that automatically generate the 3D body and hand gestures of a virtual character from speech input. For each audio clip, you will see 8 different animations that consist mostly of synthetic videos but also some direct copies of the actual performance. After watching each clip, you will be asked to provide responses to two..."
Fig. 2. Controlled synthesis result comparison of slow left hand speed (a) and low hand height (b) between the k-NN (ours, blue), k-NN+cGAN (ours, orange), and MoGlow ([Alexanderson et al. 2020], green) over a test sequence.

Fig. 3. Qualitative comparison of the controlled synthesis of k-NN with low left hand signal (a), k-NN with high hand signal (b), k-NN + cGAN with low hand signal (c), and k-NN + cGAN with high hand signal (d) over a test sequence.
prompts using seven point. The prompts are: The clip appears natural and the gesture follows the speaking style of the speaker. The gesture and the audio are well synchronized. Please ignore the quality of the facial expression of the virtual character. Each synthesis method is person-specific and they try to mimic the gesturing style of the speaker. We will show a video example of the speaker alongside their virtual character to demonstrate their actual speech gesture characteristics. This user study will take about 10 minutes to complete. Thank you for your participation.

REFERENCES