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 $\mathbf{f}_t$ and pose feature $\mathbf{p}_t$ 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 $\mathbf{m}_t \in \mathbb{R}^{14}$. We stack these together into a 1-D vector $\{\mathbf{m}_t, \mathbf{m}_{t+2}, ..., \mathbf{m}_{t+14}\} = \mathbf{f}_t \in \mathbb{R}^{112}$ (112 = 14 features × 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 $\mathbf{p}_t \in \mathbb{R}^{96}$ (96 = 2 hands × 4 joints × 3 dims × 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 $\mathbf{g}_t \in \mathbb{R}^{165}$ (165 = 55 joints × 3 dims).

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

ALGORITHM 1: audio-to-gesture k-NN search

Data:
list of audio feat. sequence $\mathcal{F} = [\tilde{\mathbf{F}}^0, \tilde{\mathbf{F}}^1, ..., \tilde{\mathbf{F}}^{M-1}]$,
list of pose feat. sequence $\mathcal{P} = [\tilde{\mathbf{P}}^0, \tilde{\mathbf{P}}^1, ..., \tilde{\mathbf{P}}^{M-1}]$,
list of gesture sequence $\mathcal{G} = [\tilde{\mathbf{G}}^0, \tilde{\mathbf{G}}^1, ..., \tilde{\mathbf{G}}^{M-1}]$,
$\tilde{\mathbf{F}} = [\tilde{\mathbf{f}}_0, \tilde{\mathbf{f}}_1, ..., \tilde{\mathbf{f}}_{T_{match}-1}]$, $\tilde{\mathbf{P}} = [\tilde{\mathbf{p}}_0, \tilde{\mathbf{p}}_1, ..., \tilde{\mathbf{p}}_{T_{match}-1}]$,
$\tilde{\mathbf{G}} = [\tilde{\mathbf{g}}_0, \tilde{\mathbf{g}}_1, ..., \tilde{\mathbf{g}}_{T_{match}-1}]$,
$\tilde{\mathbf{f}} \in \mathbb{R}^{112}$, $\tilde{\mathbf{p}} \in \mathbb{R}^{96}$, $\tilde{\mathbf{g}} \in \mathbb{R}^{165}$
Input : $k \in \mathbb{Z}$, the desired k-best neighbors,
audio feat. sequence $\mathbf{F} = [\mathbf{f}_0, \mathbf{f}_1, ..., \mathbf{f}_{T-1}]$,
control $\mathbf{C} = [\mathbf{c}_0, \mathbf{c}_1, ..., \mathbf{c}_{T-1}]$,
initial pose feat. $\mathbf{p}_{(-1)}$,
$\mathbf{f} \in \mathbb{R}^{112}$, $\mathbf{p} \in \mathbb{R}^{96}$, $\mathbf{c} \in \{0, 1\}$
Output : gesture sequence $\mathbf{G} = [\mathbf{g}_0, \mathbf{g}_1, ..., \mathbf{g}_{T-1}]$,
$\mathbf{g} \in \mathbb{R}^{165}$
t = 0;
initialize $\mathbf{G} = [\,]$, $\mathbf{P} = [\mathbf{p}_{(-1)}]$;
while $t < T$ **do**
  $\hat{\mathbf{P}} = [\,]$, $\hat{\mathbf{F}} = [\,]$, $\hat{\mathbf{G}} = [\,]$;
  for $m \leftarrow 0$ **to** $M-1$ **do**
    $r = 0$;
    $pdist = \infty$;
    for $s \leftarrow 1$ **to** $T_{match} - 1$ **do**
      if $\mathbf{c}_s == 1$ **then**
        if $d(\hat{\mathbf{p}}_s^m \in \tilde{\mathbf{P}}^m, \mathbf{p}_{t-1}) < pdist$ **then**
          $pdist = d(\hat{\mathbf{p}}_s^m \in \tilde{\mathbf{P}}^m, \mathbf{p}_{t-1})$;
          $r = s$;
        end
      end
    end
    $append(\hat{\mathbf{P}}, \hat{\mathbf{p}}^m_{r:(r+N-1)})$;
    $append(\hat{\mathbf{F}}, \hat{\mathbf{f}}^m_{r:(r+N-1)})$;
    $append(\hat{\mathbf{G}}, \hat{\mathbf{g}}^m_{r:(r+N-1)})$;
  end
  $\hat{\mathbf{P}} = \{\hat{\mathbf{p}}^0_{0:(N-1)}, \hat{\mathbf{p}}^1_{0:(N-1)}, ..., \hat{\mathbf{p}}^{M-1}_{0:(N-1)}\}$;
  $\hat{\mathbf{F}} = \{\hat{\mathbf{f}}^0_{0:(N-1)}, \hat{\mathbf{f}}^1_{0:(N-1)}, ..., \hat{\mathbf{f}}^{M-1}_{0:(N-1)}\}$;
  $\hat{\mathbf{G}} = \{\hat{\mathbf{g}}^0_{0:(N-1)}, \hat{\mathbf{g}}^1_{0:(N-1)}, ..., \hat{\mathbf{g}}^{M-1}_{0:(N-1)}\}$;
  $R_{audio} = relrank[d(\hat{\mathbf{f}}^0_0, \mathbf{f}_t), d(\hat{\mathbf{f}}^1_0, \mathbf{f}_t), ...]$;
  $R_{pose} = relrank[d(\hat{\mathbf{p}}^0_0, \mathbf{p}_{t-1}), d(\hat{\mathbf{p}}^1_0, \mathbf{p}_{t-1}), ...]$;
  $R_{combined} = R_{audio} + R_{pose}$ (elem. wise);
  sort $R_{combined}$, sort its indices into $I_{combined}$;
  $i = I_{combined}[k]$;
  $append(\mathbf{G}, \hat{\mathbf{g}}^i_{(0:(N-1))})$;
  $append(\mathbf{P}, \hat{\mathbf{p}}^i_{(0:(N-1))})$;
  $t = t + N$;
end

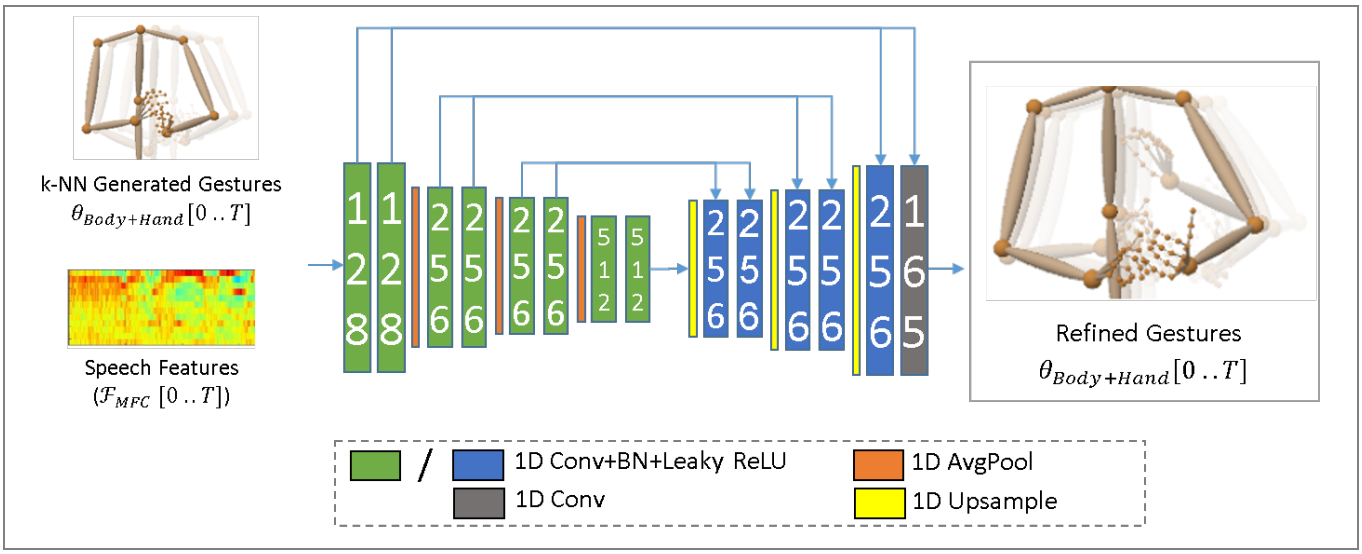
2 NETWORK ARCHITECTURE

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 parameters (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.

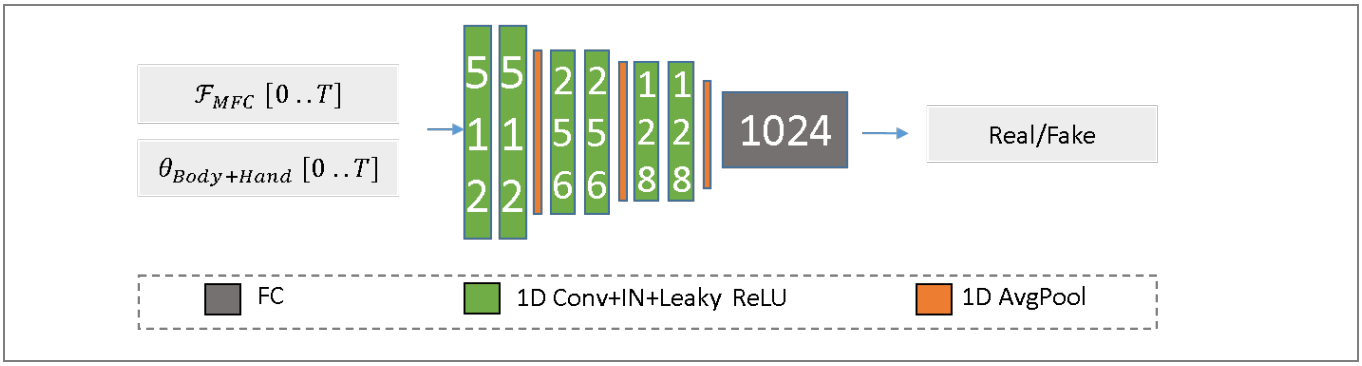
We employ a standard WGAN-GP formulation to train the method. To this end, we also remove the last sigmoid layer of the discriminator. The generator is updated after every 5 iterations to ensure that the average of the combined real and fake critic training curve fluctuates around 0.

3 ADDITIONAL EVALUATION OF CONTROL QUALITY

Figure 2 show 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



(a) Generator architecture.



(b) Discriminator architecture.

Fig. 1. The proposed network for the cGAN gesture resynchronization. The generator takes as input the MFCC audio feature and the 3D gesture generated by the k-NN and produces a refined 3D gesture. The numbers in the blocks represent the number of feature channels output by the block. Since we employ Wasserstein GAN formulation with Gradient Penalty, the last layer of the discriminator or critic network does not include a sigmoid activation.

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.

Total duration	11.4 hours
Total unique videos	105 videos
Total unique clips	9624 clips
Duration per clip	64 frames @ 15 fps

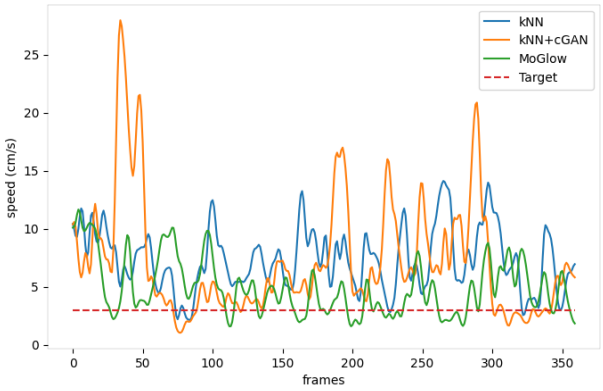
4 SEARCH DATABASE

We use the non-overlapping "Oliver" sequences of the in-the-wild speech-to-gesture data originally prepared by Ginosar et al. [2019] and 3D tracked by Habibie et al. [2021] as our search database. This data consists of more than 11 hours of audio-gesture pairs of 3D face, body, and hand poses tracked using state-of-the-art monocular trackers, and contain various range of conversational 3D upper body and hand gestures. Table 1 describes the detail of the search database we use in our experiments.

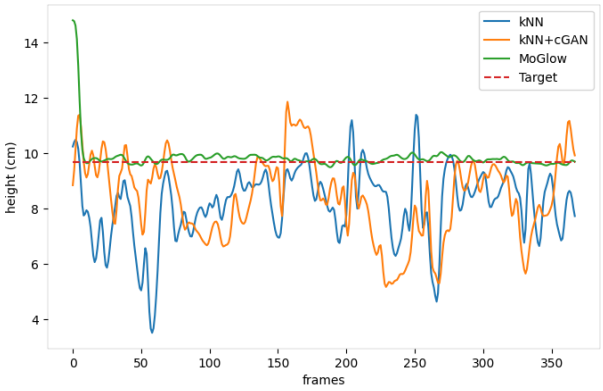
5 USER STUDY

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

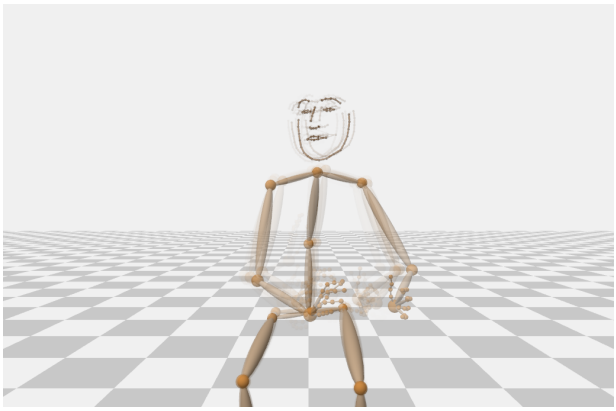


(a) Wrist position of the synthesis using "low speed" control

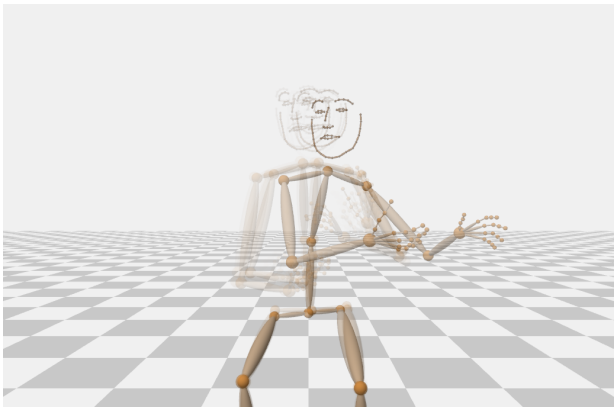


(b) Wrist position of the synthesis using "low hand" control

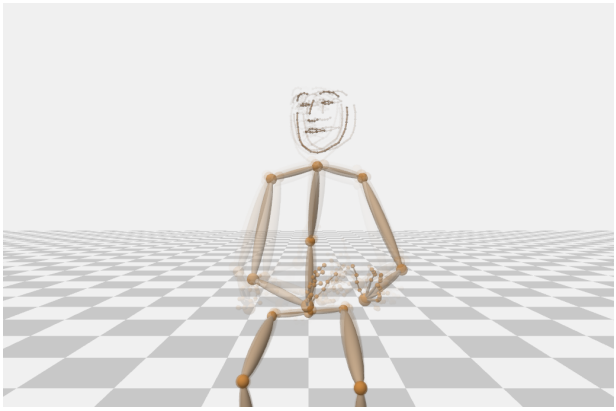
Fig. 2. Controlled synthesis result comparison of slow left hand speed (a) and low hand height (b) between the k-NN (ours, blue), k-NN+cGAN (ours, orange), and MoGlow ([Alexanderson et al. 2020], green) over a test sequence.



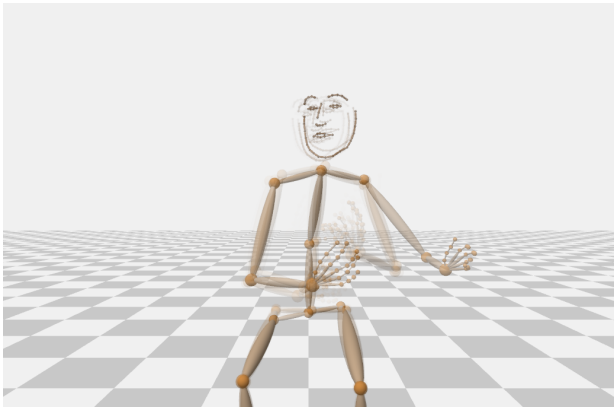
(a) Low left hand control result of k-NN



(b) High left hand control result of k-NN



(c) Low left hand control result of k-NN + cGAN



(d) High left hand control result of k-NN + cGAN

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

Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum* (2020).

S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. 2019. Learning Individual Styles of Conversational Gesture. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning Speech-driven 3D Conversational Gestures from Video. In *Proceedings of the International Conference on Intelligent Virtual Agents*.