Supplementary Material:
Talking With Hands 16.2M: A Large-Scale Dataset of Synchronized Body-Finger Motion and Audio for Conversational Motion Analysis and Synthesis

Gilwoo Lee[†] Zhiwei Deng[*] Shugao Ma[‡] Takaaki Shiratori[‡]
Siddhartha S. Srinivasa[†] Yaser Sheikh[‡]
[†]University of Washington [*]Simon Fraser University [‡]Facebook Reality Labs
{gilwoo, siddh}@cs.uw.edu zhiweid@sfu.ca {tshiratori,shugao,yasers}@fb.com

1 Detail of the motion synthesis model

The conditional inputs to the generative model are upper body joints, global wrist positions, and acoustic features. The upper body joints are local joints with respect to the parents, starting from the pelvis joint. The joint rotations are represented as quaternions. We also provide joint velocities as additional inputs, approximated as $q[t] - q[t-1]$ where q is the quaternion representation. For the global wrist positions, we compute the relative transforms from head to left wrist and to right wrist. Although this information is redundant given the upper body joints, we empirically found it to be a strong signal for the generative model to utilize. For the acoustic features, we use two features among the features extracted from GeMAPS [1]: loudness and alpha-ratio, which is the ratio of the summed energy from 50–1000 Hz and 1–5 kHz. The output of the model is the finger position and velocity. The proximal joints of the fingers are represented as quaternions, and the distal joints are represented as axis-angle. The proximal joints and the angle part of the distal joints are generated. During training we train the model to generate joint velocities as well, but only the positions are used in visualization.

The inputs and outputs of the models are in 10 fps, resampled from the original data (90 fps) and normalized. The models are implemented with PyTorch 0.4.0 [4]. The ADAM [3] optimizer with learning rate of 1e-3 is used. The final models contain 30M parameters. We believe such model capacity is important for leveraging our large scale dataset for modeling complex and subtle hand motion.

1.1 TCN

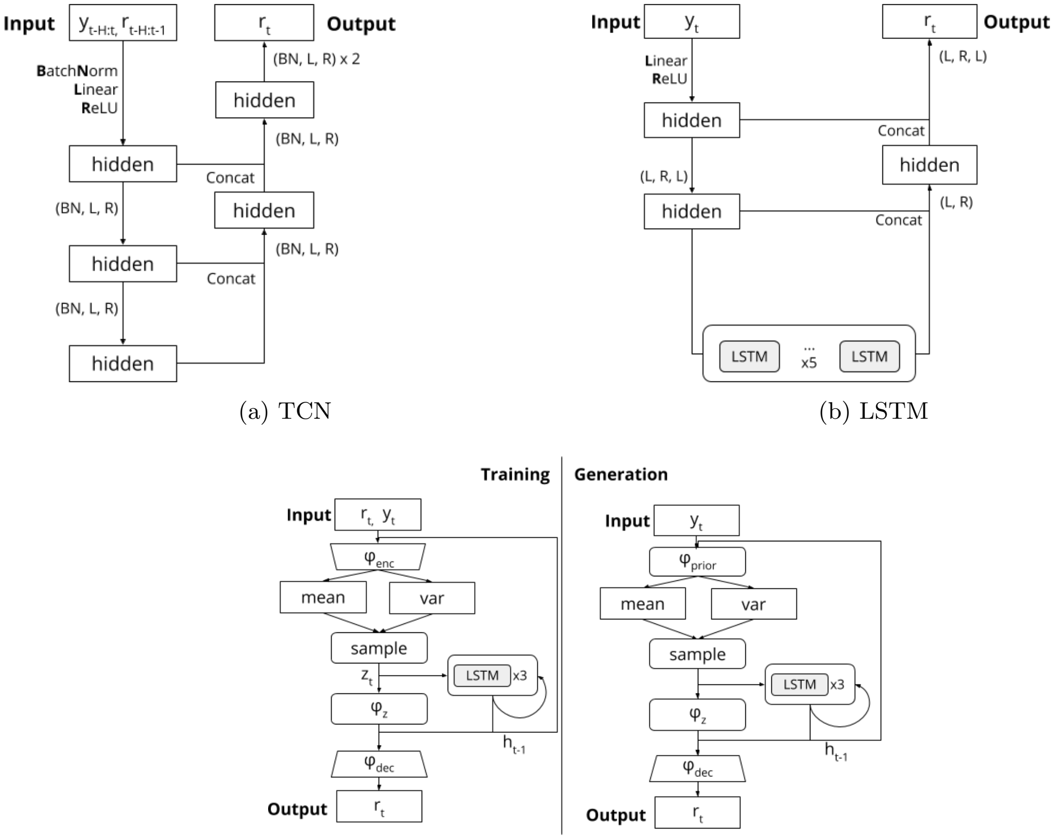
Figure 1(a) is the model diagram for TCN. For all hidden layers, 2000 units are used. Previous 10 second (100 frames) is used for generating 0.1 second (1 frame), i.e. $H = 100$ in the figure.

1.2 LSTM

Figure 1(b) is the model diagram for LSTM. The input y_t goes through two encoders before the LSTM step, and the output goes through a decoder. The hidden layers inside of LSTMs have 600 units. All other hidden layers have 800 units.

1.3 VRNN

Figure 1(c) is the model diagram for VRNN. The conditional VRNN has stacks of 3 LSTM [2] layers of hidden layers of 1000 dimensions. The prior, encoder, and decoder networks are constructed with two layers of linear and ReLU units, with hidden layers of 1000 dimensions. Both the prior and encoder outputs map to a vector of 2000 dimensions via linear operation, which is used to form the latent Gaussian distribution of 1000 dimensions.



(a) TCN

(b) LSTM

(c) VRNN

Figure 1: Generative models used for finger motion synthesis. For VRNN, ϕ_{enc}, ϕ_{prior} are jointly trained via KL divergence loss.

References

[1] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

[2] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[3] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[4] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.