# 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

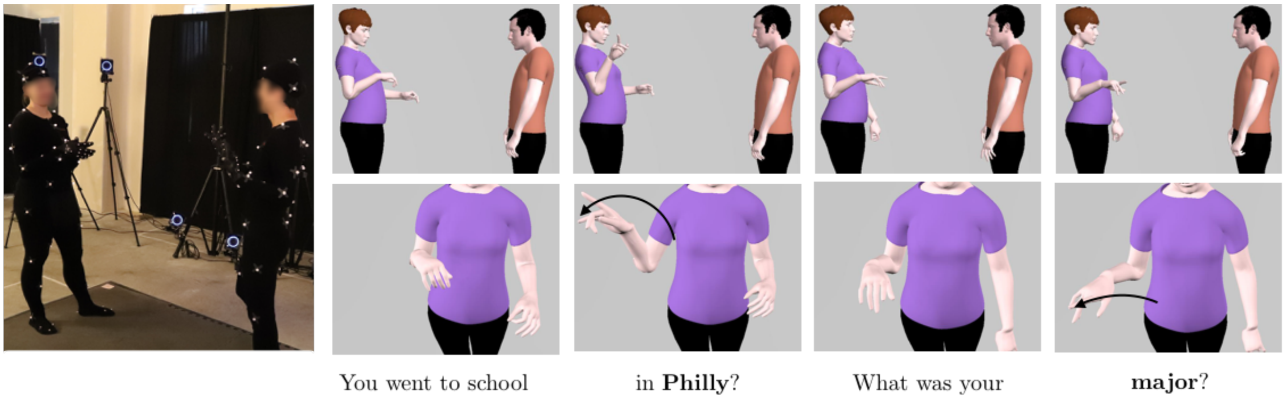


Figure 1: With our capture setup (left), we collected a large-scale, conversational motion and audio dataset of 50 people (right). Participants often make pointy gestures with high pitch emphasis when referring to a place or asking a question.

## Abstract

*We present a 16.2-million-frame (50-hour) multimodal dataset of two-person face-to-face spontaneous conversations. Our dataset features synchronized body and finger motion as well as audio data. To the best of our knowledge, it represents the largest motion capture and audio dataset of natural conversations to date. The statistical analysis verifies strong intraperson and interperson covariance of arm, hand, and speech features, potentially enabling new directions on data-driven social behavior analysis, prediction, and synthesis. As an illustration, we propose a novel real-time finger motion synthesis method: a temporal neural network innovatively trained with an inverse kinematics (IK) loss, which adds skeletal structural information to the generative model. Our qualitative user study shows that the finger motion generated by our method is perceived as natural and conversation enhancing, while the quantitative ablation study demonstrates the effectiveness of IK loss.*

## 1. Introduction

Real-time motion synthesis in conversational settings is becoming more important with the increased demand for telepresence through virtual/augmented reality and applications in 3D animation and social games. Motion synthesis is commonly done via physics-based trajectory optimization [17], data-driven generative models [20, 16], or a combination of the two [24]. In the latter two approaches, the availability of high-quality data is crucial for the quality of synthesized motion. Recent advancement in deep learning provides many advanced temporal generative models [11, 5, 23]. In leveraging such powerful models, the scale of the training dataset is important. However, when it comes to conversational settings, *high-quality, large-scale* motion capture dataset is not available due to the challenges involved in capturing face-to-face conversations in a holistic manner: the dataset must capture the complex *multimodal* interaction including spontaneous verbal and nonverbal gestures and voice, and must be captured across many subjects for diversity.

We introduce a large-scale, multimodal (body, finger, and audio) dataset of two-person conversations. Our dataset consists of 50-hour motion capture of two-person conversational data, which amounts to 16.2 million frames. To the best of our knowledge, our dataset is the largest dataset of conversational motion and voice, and has unique content: 1) nonverbal gestures associated with casual conversations

| | Ours | CMU Mocap [1] | Panoptic [14] | MMAC [2] | BigHand [33] | FM [15] | Human3.6M [13] |
|---|---|---|---|---|---|---|---|
| # of subjects | 50 | 108 | - | 25 | 10 | 2 | 3 |
| # of hours | **50** | - | 5.5 | 15 | 20 | 0.5 | 20 |
| # of frames | **16.2M** | - | 1.5M | 6.3M | 2.2M | 54K | 3.6M |
| # of sequences | 200 | 2605 | 65 | 125 | - | 56 | |
| Seq. Length (min) | 7-20 | 0.1-1 | 0.5-25 | 3-10 | - | 2-4 | - |
| Audio | **Yes** | No | Yes | - | No | No | No |
| Captured parts | Body, **Hand** | Body | Body, Hand | Body | Hand | Body, Hand | Body |
| Content | **Conversation** | Jog, walk | Dance, game | Cook | Schemed, random | Conversation | Talk, eat |
| Multi-subjects | Yes | Few | Yes | No | No | No | No |

Table 1: Comparisons of our dataset with publicly available motion capture datasets. Ours is unique in its scale, content, and multimodality. For other databases, we approximated the number of hours and frames based on the number of sequences and lengths when it was not directly available. Note that the content row shows only a few example motion types in each dataset.

without any scripted motions, 2) full body as well as finger motions, and 3) synchronized audio data, separately captured for each participant with directional microphone (*i.e.*, no audio bleeding). Our dataset will enable many future researches in multimodality analysis, modeling, and synthesis of social communication behaviors, especially in utilizing advanced deep learning models.

Comparisons of our dataset with some widely used motion capture datasets are given in Table 1. Our dataset excels in multiple dimensions: in scale, it is 2.5 times larger than [13] and many more times larger than others; in completeness, it simutaneously captures two people's body and finger motion instead of capturing just one person [2, 8, 25, 13], capturing just body [1, 2, 8, 25, 13] or not capturing audio [1, 8, 25, 13]. Our dataset is in 90 fps, which is 1.8 times higher frame rate than [13] (50 fps). The high frame rate is especially beneficial for fine-grained motion prediction. Moreover, the length of each captured sequence is 7-20 minutes, much longer than most compared datasets. Such lengths are closer to real world conversations and therefore more diverse and spontaneous human behaviors may naturally emerge. Furthermore, to facilitate future study of both person-specific model and generic model, we intentionally include both *deep* capture – two participants participate in many capture sessions resulting in a large amount of data for each, and *wide* capture – in total we have approximately 50 participants mostly recruited through public advertisements.

As highlighted in Figure 1, our dataset captures expressive gestures and voice variations that arise during conversations. Based on a statistical analysis, we found that paired participants covary their voices or motions to agree with their counterparts' statements or answer questions. Further, our data reveals strong covariance between a participant's own two hands as well as strong covariance of paired participants' arm joints and voice features (Section 3). These findings suggest that we can utilize various intrapersonal and interpersonal features to generate rich and realistic gestures in conversational settings.

To showcase the usage of our large-scale dataset, we train deep temporal neural network models to synthesize natural finger motion from upperbody joint angles and audio in real time. We choose finger motion as the synthesizing target because it contains rich social signals and is critical in social interaction setting. However, it has been rarely studied with deep learning models, mainly due to the lack of large-scale multimodal data. We tackle this problem in this paper as proof of the benefits of our dataset.

To ensure the synthesis of expressive gestures (*e.g.*, pointing index finger), we propose to use inverse kinematics (IK) loss, which incorporates kinematic constraints into the training of a temporal neural network. IK loss can be used with any generative models that output joint angles. It is formed using forward kinematics operations, so gradients taken on it penalize each joint angle according to how much it affects the target pose, *e.g.*, fingertip pose. Although IK loss has been commonly used in motion synthesis and reconstruction [27, 28], its usage in the end-to-end training of a temporal neural network is unseen.

We apply IK loss to temporal neural networks such as Long Short-Term Memory (LSTM) [11], Variational Recurrent Neural Network (VRNN) [5], and Temporal Convolutional Network (TCN) [23]. All our models meet the real-time constraint: they generate every frame in less than 0.002 seconds. Our qualitative user study shows that our model can generate natural looking and informative finger motions during conversations. Our quantitative ablation study indicates that training with IK loss causes smaller fingertip position error with negligible effect on joint angle error.

The key contributions of this paper are summarized as:

- the largest currently available face-to-face multimodal conversation dataset that contains body and finger motions as well as audio;
- a statistical analysis of combined body, hand, and acoustic features during face-to-face converations that verifies previously used heuristics in gesture synthesis;
- the innovative application of IK loss to train temporal neural network to synthesize realistic finger motion.

# 2. Related Work

## 2.1. Human Motion Capture Datasets

There exist many 3D human motion datasets using marker-based, markerless, and depth sensor-based tracking systems. Our dataset contains 50-hour, two-person, face-to-face social conversational interactions, capturing body and finger motions as well as speech. This unique focus and large scale clearly differentiate our work from existing datasets.

**Single Person Motion Datasets** The CMU motion capture dataset [1] is one of the most widely used motion capture datasets in the research community. It contains both single-person and two-person interactions, and its diverse motions range from locomotion to sport activities. Although large scale, its number of sequences per motion type is relatively small. This contrasts with our dataset, which provides a large number of sequences of one kind: two-person conversational motion. The CMU Multi-Modal Activity Database (MMAC) [2], closer in spirit to ours, captures a large amount of multimodal human motion sequences, all related to kitchen activities.

In the HumanEVA dataset [25], 3D motion capture data is accompanied by synchronized video clips, which may be useful for human pose estimation from videos. In the Human3.6M dataset [13], human daily activities such as eating were captured. Multimodal human actions were captured in [22] using an optical motion capture system, multiple cameras, depth sensors, IMUs, and microphones. The University of Edinburgh Interaction Database [3] includes both body and finger motion, but it focuses only on human-object interactions. Some datasets are dedicated to 3D hand pose capture [7, 33]. The Finger Motion (FM) [15] dataset is closest to ours in terms of its content, as it contains full body and hand motions in conversational settings. However, the dataset contains only scripted motions by actors and has only single-person data without audio.

**Multi-Person Interaction Datasets** Similar to our approach, Lu *et al*. [18] also used a marker-based motion capture system to capture two-person interactions. Their action categories include object handovers and conversations. However, the activities are scripted. Ye *et al*. [32] captured human interactions using three hand-held Kinects. Their capture setting is much simpler than ours, allowing interactions in a natural environment, but they do not capture finger motion. The Utrecht Multi-Person Motion dataset [29] provides synchronized video and 3D motion capture data of multi-person interactions. Recently, Joo *et al*. [14] captured human social interactions in a sophisticated dome called Panoptic Studio, which mounted many RGB cameras, depth sensors, and microphones. While the available data types are similar, ours is much larger than these datasets.

## 2.2. Finger Motion Synthesis

Physics-based approaches combine various kinematic, task-specific, and style constraints. Liu [17] optimizes trajectories using contact and kinematic constraints for object-hand interaction. Pollard *et al*. [24] and Zhao *et al*. [34] combine a small dataset with physical simulation to generate grasping motions. These methods are effective when task-specific constraints or external contact constraints are clear, but cannot be immediately applied to verbal and nonverbal gesture synthesis.

The key of data-driven hand motion synthesis is interpolation among nearest neighbor trajectories from motion libraries. The quality and computational tractability of these approaches largely depend on the quality of the motion library, the motion query algorithm, and the objective function. The collected trajectories are segmented [19, 15, 21, 20, 26] by some predefined gesture phases. Combination of several cost terms are used for interpolation such as proximity of the pose to the data [15], smoothness of the motion [19] or transition likelihood between segments [26].

Another common approach is to learn probabilistic generative models for hand motions. Generative models are useful in real-time motion synthesis, as it can generate motions given history of observations without requiring the whole trajectory. Levine *et al*. [16] use a Hidden Markov Model (HMM) to generate arm motions given speech, though finger motinos are not considered in this work. Mousas *et al*. [20] use an HMM to generate finger motions given wrist poses of a performer. Analogous to interpolation-based methods, these methods require clustering the database to similar states in order to train discrete HMMs.

We encourage interested readers to refer to a survey paper [31] for more comprehensive overview.

## 2.3. Temporal Generative Models

Recent advancement of deep learning resulted in many models with high capacity. Oord *et al*. [23] proposed a dilated temporal convolutional network for generating raw audio. Holden *et al*. [12] use an autoencoder to find low dimensional manifold for human motions. Walker *et al*. [30] use Variational AutoEncoder (VAE) to encode past poses and decode future poses, while using a Generative-Adversarial Network to generate fully rendered images. Habibi *et al*. [9] combines VAE with LSTM [11] to generate human motions. With our large-scale dataset, many of these methods could be readily applied to finger motion synthesis.

# 3. Dataset Construction

Our dataset consists of 50 sessions of two-people conversations. Each session is approximately one hour and has 4-6 subsessions, either free-talking or video retelling (Section 3.2). We provide body and hand poses, raw audio data,

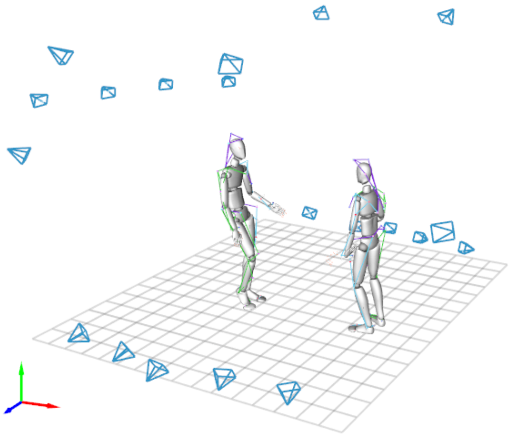


Figure 2: Location of the 24 cameras: 14 cameras were placed on each side of the participants to best capture fingers.

and acoustic features. All data is processed in 90 fps except for raw audio.

### 3.1. Capture System

Our capture system included 24 OptiTrack Prime 17W cameras surrounding a capture area of approximately 3 $m \times 3\ m$ (Figure 2) and two directional microphone headsets. Ten of the 24 cameras were placed at 1.6 $m$ height. The remaining 14 cameras were placed at each side of the two participants, ten for bottom-to-top views and four for top-to-bottom views, to get the best capture of the finger motions. Participants wore marker-attached suits and gloves. We followed the standard marker arrangement provided by OptiTrack for body capture and glove marker arrangement suggested in Han *et al*. [10].

Before the actual data capture each participant was asked to follow a recorded video of gestures and large body movements, which helped the participant to familiarize with the motion capture setup and the suit. The recorded body and finger motions were used to calibrate the bone lengths and translations per participant.

All motion capture data was converted to a joint angle representation. All joints are represented by local transforms with respect to parent joints; a global pose of the pelvis is provided with respect to a fixed frame, which can be used to track the global movement of the subject as well as to measure the distance between the two subjects. We processed finger motion capture data using Han *et al*.'s method [10], which automatically labels finger markers and computes joint angles via optimization based on inverse kinematics.

We provide raw audio of each person, which was synchronized with motion data by BrightEye 56 Pulse Generator and recorded by OctaMic XTC. In addition, we provide the Geneva Minimalistic Acoustic Parameter set (GeMAPS) [6], a comprehensive set of acoustic features that captures various aspects of vocal expressions. The features include frequency (*e.g*., pitch, jitter), energy, amplitude (*e.g*., shimmer, loudness), and spectral features.

### 3.2. Conversational Tasks

To inspire spontaneous conversations, we experimented with various conversational tasks, from which we learned two lessons. First, the conversational tasks should provide sufficient context to engage the participants. Second, to accommodate a wide variety of participants, the task should not require too much background knowledge. Therefore, we chose two main tasks for our participants: *free conversation* around a given topic and *video retelling*.

Free conversation topics were chosen from a comprehensive set originally designed for casual conversations in English classes [4]. Some example topics are:

- Where are you planning to go for your next vacation?
- What good restaurants do you know of around here?

In each capture session, the pair of participants engaged in 2-3 such conversations. We told participants to freely drift from the topic, similar to how people segue topics in casual conversation.

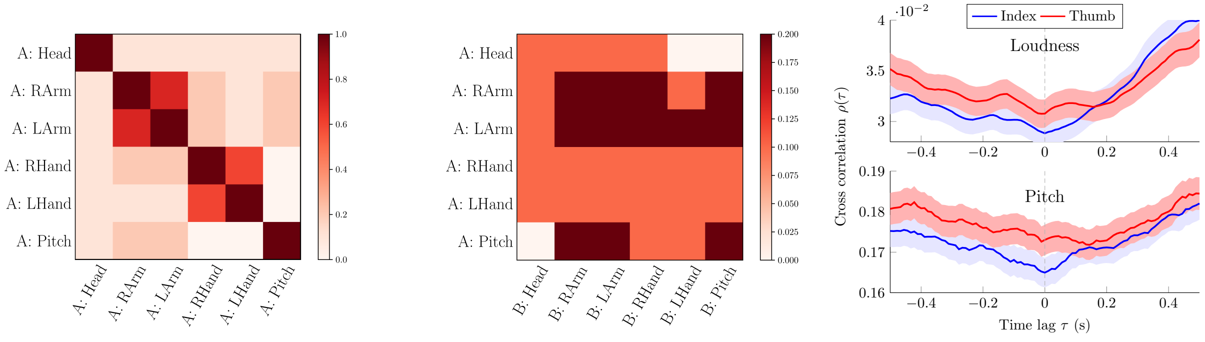
In addition, participants engaged in two subsessions of video retelling. To start, one participant watched a 5-minute video while the other waited outside the room. Then the participant who watched the clip *told* the story to the other participant, during which the other person could interrupt for clarification questions. After the telling, the participant who heard the story *retold* the story to the first participant, during which the first participant could interrupt to correct the retelling. This design was to engage participants in spontaneous conversational behaviors such as explaining, active listening, interruptions, and questions.

If participants were highly engaged, we let the conversation continue until it ended naturally. As a result, the conversations ranged in length from 7 to 20 minutes.

### 3.3. Captured Data Analysis

We first investigate what correlates between how people speak and how they use their hands by evaluating the covariance of the upper body, finger joints, and acoustic features. Figure 3a and 3b show intraperson and interperson covariance. We take the average joint angle per finger. For visual simplicity, we grouped the fingers into left and right hands and grouped wrist, elbow, and shoulder joints into left and right arms. Among the possible covariance pairs, we use the maximum values to represent each cell: *e.g*., the covariance of left index finger and right wrist may correspond to the cell (LHand, RArm) in Figure 3a.

Our statistical findings are coherent with heuristics used in the previous gesture synthesis work. Jörg *et al*. [15] empirically noted that wrist joints were strongly correlated with finger motions. Indeed, strong covariance exists between



(a) Intraperson covariance

(b) Interperson covariance

(c) Temporal correlation

Figure 3: (a) Single-person covariance among joints and pitch. Strong off-diagonal covariances among the two arm and hand joints indicate that the two hand motions are correlated. (b) Strong covariance is found between the pitch of one participant and the pitch and arm motion of the other. (c) Acoustic features and proximal joints are temporally correlated.

the arm and finger joints of the same side. Our analysis further shows that the two hands are strongly correlated, implying that people often move both hands together. Levine *et al.* [16] noted pitch, loudness, and duration as key features contributing to gestures. Likewise, we observe that pitch covaries with both left and right arm joints (Figure 3a). Our analysis further indicates that the audio features covary with the other person's arms (Figure 3b). Qualitatively, we observed that while one person talks and makes gestures, the other person often verbally responds, such as answering questions or agreeing to the other's comments, which accounts for the covariance.

Figure 3c shows that the velocities of the proximal finger joints are temporally correlated with two audio features: loudness and pitch. We measure temporal correlation using the Pearson cross-correlation coefficients[1].

## 4. Real-Time Finger Motion Synthesis

To illustrate the use of our dataset on data-driven motion synthesis, we train temporal generative models using our dataset to synthesize finger motion. We also propose to use inverse kinematics loss, described below.

Formally, generative models learn the probability function $p(\mathbf{r}_t|\mathbf{r}_{<t}, \mathbf{y}_{\le t})$, where $\mathbf{r}_t$ is the joint angles of the fingers at time $t$, and $\mathbf{y}_{\le t}$ is the history of observations. We assume that the joint angles and angular velocities of upper body joints and acoustic features are observed. Partial observation of body joints is commonly assumed in finger motion synthesis [15, 19, 21] where the goal is to fill in the finger pose to match the context of the body pose. For acoustic features, we use loudness and alpha-ratio, which is the ratio of the summed energy from 50–1000 Hz and 1–5 kHz. For body features, we use upper body joints as well as relative transforms of wrist joints with respect to a root joint.

### 4.1. Inverse Kinematics (IK) Loss

Many temporal generative models [11, 12, 23] can produce complex motions. However, the structures in data are implicitly learned by the model and cannot be directly regulated, even when prior knowledge is available. The human skeleton, in fact, has a rich set of kinematic constraints that place *different weights* to different joints. For example, moving a shoulder results in a larger wrist movement than moving an elbow by the same amount because of the additional transform between the shoulder and elbow. When generating joint motions for a human skeleton, the learner must take into account the cascaded effects across joints.

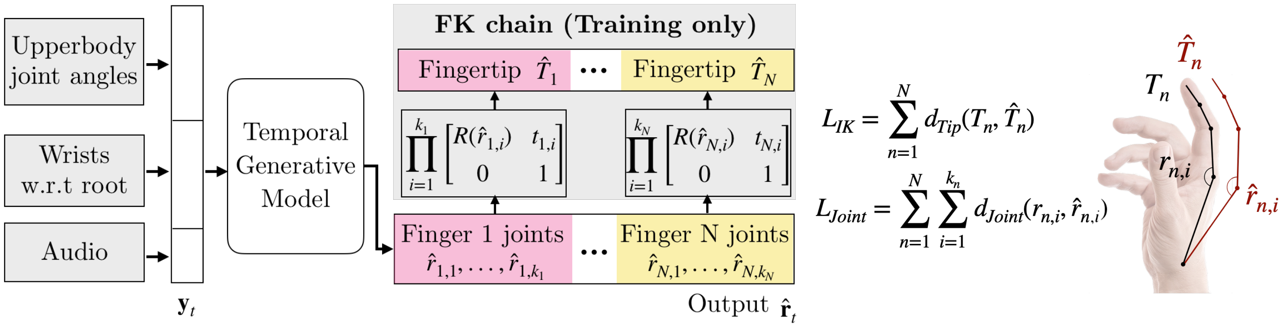
We use IK loss (Figure 4) to incorporate this kinematic structural knowledge. To compute it, we first form a kinematic chain by combining the output joint angles of the generative model with the bone translations. Let $\hat{\mathbf{r}}$ be the joint angles generated by the temporal model a particular timestep, and let $\hat{r}_{n,i}$ be the angles of the joints for finger $n$. We augment these joint angles with bone translations $t_{n,i}$ and form homogenous transformation matrices. Finally, we multiply them in the order given by the skeletal structure, starting from the base joint to the distal joint. The forward kinematics of finger $n$ can be represented as:

$$\hat{T}_n = fk(\hat{r}_{n,1}, \cdots \hat{r}_{n,k_n}) = \prod_{i=1}^{k_n} \begin{bmatrix} R(\hat{r}_{n,i}) & t_{n,i} \\ 0 & 1 \end{bmatrix}$$

where $k_n$ is the number of joints on finger $n$ and $R$ is a function that maps a rotation vector $\hat{r}$ to a $3 \times 3$ rotation matrix. In our implementation, $\hat{r}$ is a quaternion.

The pose difference between $\hat{T}_n$ and the true $T_n$ corresponds to the pose difference of fingertip $n$ (Figure 4b),

[1] For two stochastic processes $(X_t, Y_t)$ that are jointly wide-sense stationary, the cross correlation with timelag $\tau$ is given by $\rho_{XY}(\tau) = \frac{1}{\sigma_X \sigma_Y}\mathbb{E}[(X_t - \mu_X)(Y_{t+\tau} - \mu_Y)]$, where $\sigma_X, \sigma_Y$ are standard deviations, and $\mu_X, \mu_Y$ are the means of $X$ and $Y$.


Upperbody joint angles
Wrists w.r.t root
Audio
$\mathbf{y}_t$
Temporal Generative Model
FK chain (Training only)
Fingertip $\hat{T}_1$
Fingertip $\hat{T}_N$
$\prod_{i=1}^{k_1} \begin{bmatrix} R(\hat{r}_{1,i}) & t_{1,i} \\ 0 & 1 \end{bmatrix}$
$\prod_{i=1}^{k_N} \begin{bmatrix} R(\hat{r}_{N,i}) & t_{N,i} \\ 0 & 1 \end{bmatrix}$
Finger 1 joints $\hat{r}_{1,1}, \ldots, \hat{r}_{1,k_1}$
Finger N joints $\hat{r}_{N,1}, \ldots, \hat{r}_{N,k_N}$
Output $\hat{\mathbf{r}}_t$
$L_{IK} = \sum_{n=1}^{N} d_{Tip}(T_n, \hat{T}_n)$
$L_{Joint} = \sum_{n=1}^{N} \sum_{i=1}^{k_n} d_{Joint}(r_{n,i}, \hat{r}_{n,i})$
$\hat{T}_n$
$T_n$
$r_{n,i}$
$\hat{r}_{n,i}$

(a) Training scheme with forward kinematics (FK) chain

(b) IK loss

Figure 4: (a) Given observations of body pose and audio features $\mathbf{y}_t$, a temporal generative model predicts joint angles $\hat{\mathbf{r}}_\mathbf{t}$, which undergoes forward kinematics operations only during training. (b) The output of the FK chain forms IK loss.

which we call the IK loss:

$$L_{IK}(\mathbf{r}, \hat{\mathbf{r}}) = \sum_{n=1}^{N} d_{Tip}(T_n, \hat{T_n})$$

where $d_{Tip}$ is a distance metric of our choice on the SE3 space. Intuitively, having $L_{IK}$ helps the network learn what matters more, *i.e.*, fingertip poses. We use only the positional difference in our implementation. This loss is used in addition to the joint angle loss:

$$L_{Joint}(\mathbf{r}, \hat{\mathbf{r}}) = \sum_{n=1}^{N} \sum_{i=1}^{k_n} d_{Joint}(r_{n,i}, \hat{r}_{n,i})$$

where $d_{Joint}$ is a distance metric of our choice on the SO3 space. Here, we take mean squared error on quaternions. Finally, the overall loss is a weighted combination of these terms, with a manually tuned weight $\lambda$:

$$L_{data}(\mathbf{r}, \hat{\mathbf{r}}) = L_{IK}(\mathbf{r}, \hat{\mathbf{r}}) + \lambda L_{Joint}(\mathbf{r}, \hat{\mathbf{r}})$$

$L_{data}$ can be combined with other loss terms needed by the generative models. For example, when using a variational model, we combine it with a distributional loss.

The forward kinematics chain operation is performed only during training. Note that the loss for joint angles and IK loss are complementary: if $L_{Joint}$ is zero, then $L_{IK}$ is zero.

## 4.2. Temporal Generative Models

IK loss can be applied to any generative models that output joint angles. We implemented three models: LSTM, VRNN, TCN. Here we discuss the high-level implementation and leave the detail to the supplementary material.

**Long Short-Term Memory** LSTM [11] maintains a hidden state $\mathbf{h}_t$ that encodes the history of previously generated states $\mathbf{r}_{<t}$ and conditions $\mathbf{y}_{<t}$. At each timestep, condition $\mathbf{y}_t$ is encoded through two encoders before being passed to a stacked LSTM of five layers. Each encoder consists of layers of linear and Rectified Linear Units (ReLUs). The output of the LSTM goes through two decoders, composed similarly to the encoders. We connect encoder outputs to decoders: the first decoder inputs the second encoder's and the LSTM's output. To generate $\mathbf{r}_t$, the second decoder inputs the first decoder's and the first encoder's output.

**Variational Recurrent Neural Network** VRNN [5] follows a Bayesian perspective and assumes that a latent space $Z$ controls hand skeleton motions. We assume that the latent distribution is unit Gaussian. To encode $\mathbf{r}_{<t}$, $\mathbf{z}_{<t}$, and $\mathbf{y}_{\leq t}$, a recurrent neural network is used: $\mathbf{h}_t = RNN(\mathbf{h}_{t-1}, \mathbf{r}_{t-1}, \mathbf{z}_{t-1}, \mathbf{y}_t)$. Our implementation uses a stacked LSTM of 5 layers. During training, we follow the standard variational inference method by defining a proposal distribution (*i.e.*, the encoder), and we maximize the evidence lowerbound (ELBO). During inference, given conditional data $y_{\leq t}$ and an initial prior distribution $p(\mathbf{z}_0) = \mathcal{N}(0, I)$, our model first generates $\mathbf{r}_0$ by sampling from $p(\mathbf{z}_0)$ and $p(\mathbf{r}_0|\mathbf{z}_0, \mathbf{h}_0)$. Subsequently, $\mathbf{r}_t$ is generated from the decoder, which defines $p(\mathbf{r}_t|\mathbf{z}_t, \mathbf{h}_t)$.

**Temporal Convolutional Network** Our TCN [23] takes one second of the observation history and previously generated finger motions as input. It encodes the history through three layers of batch normalization, fully connected linear and ReLUs, and it decodes through three similar layers. Skip connections connect the first and second layers of encoder outputs to decoder layers.

# 5. Experiments

For the training of the three generative models, we used the data from one of the *deep capture* participants, who participated in 11 sessions. We used only the portions where at least one of the wrists were above pelvis – this was a good indication of active gestures. In total, we used approximately 120 minutes of data.

All our implementation was made on an Intel CPU E5-

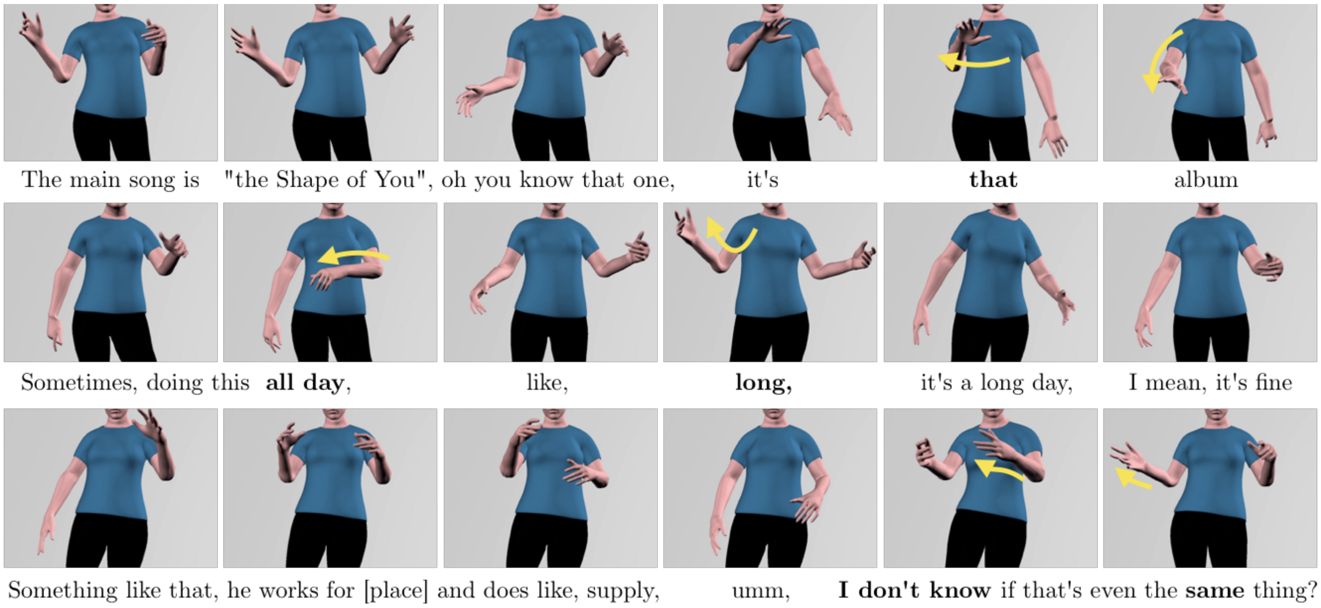

The main song is
"the Shape of You", oh you know that one,
it's
that
album
Sometimes, doing this
all day,
like,
long,
it's a long day,
I mean, it's fine
Something like that, he works for [place] and does like, supply,
umm,
I don't know if that's even the same thing?

Figure 5: Results of finger motion synthesis from body poses and audio by our model.

2630 with an NVDIA Titan Xp Graphics card with 12GB memory. On this machine, RNN, VRNN, and TCN processed one frame in $1.27 \times 10^{-3}s$, $1.53 \times 10^{-3}s$, and $1.16 \times 10^{-3}s$ on average, respectively. This allows at most 500 fps.

Two important criteria for evaluating the quality of the synthesized finger motion are 1) whether it looks like natural human motion and 2) whether it reflects personal characteristics. As these criteria are perceptual, we performed a qualitative user study to evaluate them. We asked 18 participants to evaluate the richness of motion, naturalness, and personal motion characteristics.

For quantitative evaluation, we computed the mean squared error between the generated motion and motion capture on a left-out test set, for fingertip positions and joint angles. Results demonstrate the effectiveness of IK loss.

Figure 5 shows some of the finger motions generated by our model that were used for the user study.

### 5.1. Qualitative User study

Each participant first watched a reference motion capture of the character and then watched a set of clips that were generated by three methods: motion capture, static finger motion, and our method. The joint angles in the static case were fixed to be the mean angles of the corresponding joints in each clip. Across all clips, the motion of body joints other than fingers was taken from motion capture data.

Since our user study aimed to verify whether our method produced motions that were *natural*, *matched the audio*, and *matched the person's character*, we asked participants to score their agreement level via a five-point scale for the following statements for a set of 24 clips:

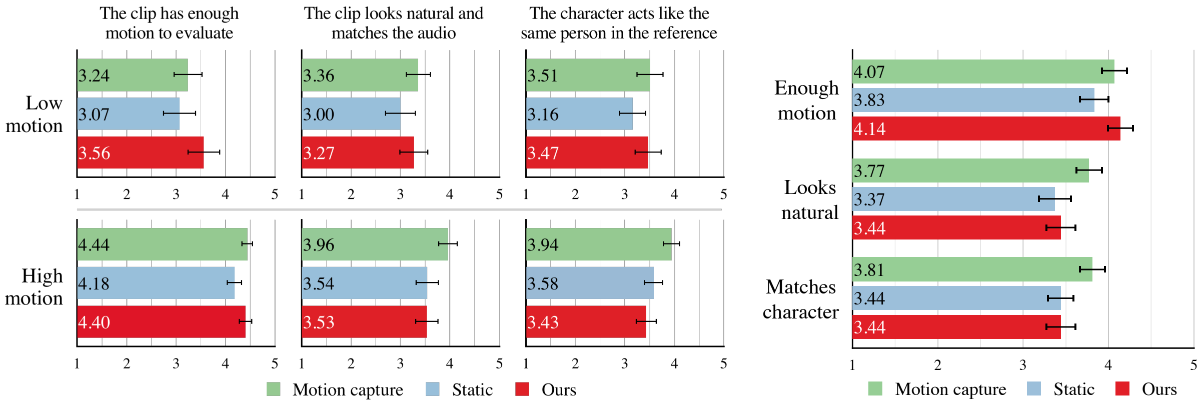
1. The clip has **enough motion** to evaluate
2. The clip looks **natural and matches the audio**
3. The character **acts like the same person** in the reference

The clips shown were randomly ordered. Participants were asked to evaluate only the finger motions.

We observed that the motion scores, *i.e.*, the responses to question 1, varied significantly across the clips. This variation implied that participants' judgements may be heavily influenced by how much motion they perceived in each clip. Thus, for an in-depth analysis, we split the clips into two groups according to the motion scores on the motion capture clips. For the clips of *low motion* scores (below 3.7), our method was comparable to the motion capture, while for the clips of *high motion* scores (above 3.7), our proposed method performed comparably to the static clips.

This result was initially counter-intuitive. However, as we assessed the clips, we noticed that the clips with high motion scores often contained large arm motions, while the clips with low motion scores contained smaller arm motions but distinctive finger motions, such as pointy gestures. That is, when there were large arm motions, participants ended up focusing on the arm motions instead of finger motions. This implies that the participants' responses on low motion clips are more accurate evaluations of the finger motions. Since the finger motion generated by our method were perceived to be significantly better than the static clips for the low motion clips, this evaluation shows that our method indeed produces natural motions that match the personal characteristics.

Figure 6b shows the overall participant responses in this



(a) User study responses per motion score group

(b) Overall responses

Figure 6: (a) Analysis on participants' responses, split into low and high motion groups. In the low motion group, our method is comparable to the motion capture. (b) Our method produces clips with enough motions that look comparably to static clips.

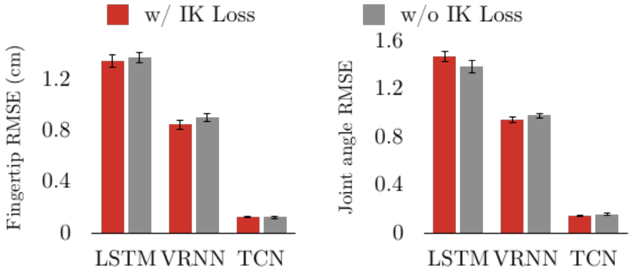


Figure 7: Fingertip position RMSE (left) and normalized joint angle RMSE (right) for models trained with IK loss and without. Models trained with IK loss results in smaller fingertip position errors.

user study. Due to the scores in the high motions clips, overall, participants perceived the motion generated by our model to be comparable to the static clips.

### 5.2. Effect of IK loss

We performed an ablation study to evaluate the benefit of using IK loss in training, comparing models trained with IK loss and joint angle loss to those trained only with the latter. Figure 7 shows the normalized mean squared error of the two models on fingertip positions and joint angles.

The model trained with IK loss reduced fingertip position error to a greater extent than the one without it, and it did not compromise joint angle error significantly. As discussed in Section 4, IK loss and joint angle loss are complementary: zero joint angle loss leads to zero IK loss. Our ablation study indeed verifies this complementarity.

## 6. Conclusion

This paper described our new multimodal dataset of human face-to-face social conversations. Our dataset is in large-scale: it consists of approximately one hour recordings for each of 50 two-person conversation sessions. Our dataset captures the multimodality of conversations: both body and finger motions, together with their individual audio data, were captured. As with existing motion capture datasets, the dataset currently lacks facial motion or gaze, which were technically challenging given an optical motion capture system or even with cameras at distance. Nonetheless, we believe that it contains sufficient and interesting social interaction data to be of benefit, as we illustrated with finger motion synthesis models trained with IK loss. Qualitative evaluation by a user study suggests that our method can generate natural looking and conversation enhancing finger motions. We also showed the advantage of using IK loss to train a generative motion sequence model in a quantitative ablation study. By publicly releasing this dataset, we hope to promote future research on analyzing, predicting, and synthesizing human social behaviors.

### Acknowledgments

Gilwoo Lee is partially supported by Kwanjeong Educational Foundation. This work was partially funded by the National Institute of Health R01 (#R01EB019335), National Science Foundation CPS (#1544797), National Science Foundation NRI (#1637748), the Office of Naval Research, the RCTA, Amazon, and Honda. We would like to thank Jesse Stiles for the audio recording setup, and Breannan Smith and Colin Lea for their constructive feedback.

# References

[1] CMU Graphics Lab motion capture database. http://mocap.cs.cmu.edu/,. 2, 3

[2] CMU multi-modal activity database. http://kitchen.cs.cmu.edu,. 2, 3

[3] Interaction database of the University of Edinburgh. http://www.ipab.inf.ed.ac.uk/cgvu/InteractionDatabase/interactiondb.html,. 3

[4] Teflpedia: Category:conversation questions. http://teflpedia.com/Category:Conversation_questions. Accessed: 2017-08-10. 4

[5] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C. Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Proc. Advances in neural information processing systems*, pages 2980–2988, 2015. 1, 2, 6

[6] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016. 4

[7] Andrew Gardner, Jinko Kanno, Christian A Duncan, and Rastko Selmic. Measuring distance between unordered sets of different sizes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 3

[8] Ralph Gross and Jianbo Shi. The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18, Carnegie Mellon University, 2001. 2

[9] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. In *Proc. British Machine Vision Conference*, 2017. 3

[10] Shangchen Han, Beibei Liu, Robert Wang, Yuting Ye, Christopher D. Twigg, and Kenrick Kin. Online optical marker-based hand tracking with deep labels. *ACM Transactions on Graphics*, 37(4), 2018. 4

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1, 2, 3, 5, 6

[12] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics*, 36(4), 2017. 3, 5

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2, 3

[14] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic Studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2, 3

[15] Sophie Jörg, Jessica Hodgins, and Alla Safonova. Data-driven finger motion synthesis for gesturing characters. *ACM Transactions on Graphics*, 31(6), 2012. 2, 3, 4, 5

[16] Sergey Levine, Christian Theobalt, and Vladlen Koltun. Real-time prosody-driven synthesis of body language. *ACM Transactions on Graphics*, 28(5), 2009. 1, 3, 5

[17] C. Karen Liu. Synthesis of interactive hand manipulation. In *Proc. SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 163–171, 2008. 1, 3

[18] David V. Lu, Annamaria Pileggi, and William D. Smart. Multi-person motion capture dataset for analyzing human interaction. In *Proc. RSS Workshop on Human-Robot Interaction*, 2011. 3

[19] Anna Majkowska, Victor B. Zordan, and Petros Faloutsos. Automatic splicing for hand and body animations. In *Proc. SIGGRAPH/Eurographics Symposium on Computer animation*, pages 309–316, 2006. 3, 5

[20] Christos Mousas and Christos-Nikolaos Anagnostopoulos. Real-time performance-driven finger motion synthesis. *Computers & Graphics*, 65:1–11, 2017. 1, 3

[21] Christos Mousas, Christos-Nikolaos Anagnostopoulos, and Paul Newbury. Finger motion estimation and synthesis for gesturing characters. In *Proc. Spring Conference on Computer Graphics*, pages 97–104, 2015. 3, 5

[22] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. Berkeley MHAD: A comprehensive multimodal human action database. In *Proc. IEEE Workshop on Applications on Computer Vision*, 2013. 3

[23] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016. 1, 2, 3, 5, 6

[24] Nancy S. Pollard and Victor B. Zordan. Physically based grasping control from example. In *Proc. ACM SIGGRAPH/Eurographics Symposium on Computer animation*, pages 311–318, 2005. 1, 3

[25] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1):4, Aug 2009. 2, 3

[26] Matthew Stone, Doug DeCarlo, Insuk Oh, Christian Rodriguez, Adrian Stere, Alyssa Lees, and Chris Bregler. Speaking with hands: Creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics*, 23(3), 2004. 3

[27] Jochen Tautges, Arno Zinke, Björn Krüger, Jan Baumann, Andreas Weber, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bernd Eberhardt. Motion reconstruction using sparse accelerometer data. *ACM Transactions on Graphics*, 30(3), 2011. 2

[28] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, 33(5):169, 2014. 2

[29] Nico van der Aa, Xinghan Luo, Geert-Jan Giezeman, Robby T. Tan, and Remco C. Veltkamp. UMPM benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *Proc. IEEE International Conference on Computer Vision Workshops*, pages 1264–1269, 2011. 3

[30] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proc. IEEE International Conference on Computer Vision*, pages 3352–3361, 2017. 3

[31] Nkenge Wheatland, Yingying Wang, Huaguang Song, Michael Neff, Victor Zordan, and Sophie Jörg. State of the art in hand and finger modeling and animation. *Computer Graphics Forum*, 34(2):735–760. 3

[32] Genzhi Ye, Yebin Liu, Nils Hasler, Xiangyang Ji, Qionghai Dai, and Christian Theobalt. Performance Capture of Interacting Characters with Handheld Kinects. In *Proc. European Conference on Computer Vision*, pages 828–841, 2012. 3

[33] Shanxin Yuan, Qi Ye, Björn Stenger, Siddhant Jain, and Tae-Kyun Kim. BigHand2.2M benchmark: Hand pose dataset and state of the art analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2605–2613, 2017. 2, 3

[34] Wenping Zhao, Jianjie Zhang, Jianyuan Min, and Jinxiang Chai. Robust realtime physics-based motion control for human grasping. *ACM Transactions on Graphics*, 32(6), 2013. 3