Learning Speech-driven 3D Conversational Gestures from Video

Ikhsanul Habibie
Max Planck Institute for Informatics

Weipeng Xu
Facebook Reality Labs

Dushyant Mehta
Max Planck Institute for Informatics

Lingjie Liu
Max Planck Institute for Informatics

Hans-Peter Seidel
Max Planck Institute for Informatics

Gerard Pons-Moll
University of Tübingen

Mohamed Elgharib
Max Planck Institute for Informatics

Christian Theobalt
Max Planck Institute for Informatics

ABSTRACT

We propose the first approach to synthesize the synchronous 3D conversational body and hand gestures, as well as 3D face and head animations, of a virtual character from speech input. Our algorithm uses a CNN architecture that leverages the inherent correlation between facial expression and hand gestures. Synthesis of conversational body gestures is a multi-modal problem since many similar gestures can plausibly accompany the same input speech. To synthesize plausible body gestures in this setting, we train a Generative Adversarial Network (GAN) based model that measures the plausibility of the generated sequences of 3D body motion when paired with the input audio features. We also contribute a new corpus that contains more than 33 hours of annotated data from in-the-wild videos of talking people. To this end, we apply state-of-the-art monocular approaches for 3D body and hand pose estimation as well as 3D face performance capture to the video corpus. In this way, we can train on orders of magnitude more data than previous algorithms that resort to complex in-studio motion capture solutions, and thereby train more expressive synthesis algorithms. Our experiments and user study show the state-of-the-art quality of our speech-synthesized full 3D character animations.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

gesture synthesis, character control, audio-driven pose estimation

ACM Reference Format:
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 *21th ACM International Conference on Intelligent Virtual Agents (IVA '21), September 14–17, 2021, Virtual Event, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3472306.3478335

This work was supported by the ERC Consolidator Grant 4DRepLy (770784). Gerard Pons-Moll is funded by the Deutsche Forschungsgemeinschaft (DFG. German Research Foundation) - 409792180.

IVA '21, September 14–17, 2021, Virtual Event, Japan

ACM ISBN 978-1-4503-8619-7/21/09.
https://doi.org/10.1145/3472306.3478335

1 INTRODUCTION

Virtual human characters are a crucial component in many computer graphics applications, such as games or shared virtual environments. Traditionally, their generation requires a combination of complex motion capture recordings and tedious work by animation experts to generate plausible appearance and movement. The particular challenges include the animation of the conversational body gestures of a talking avatar, as well as the facial expressions that accompany the audio in conveying the emotion and mannerisms of the speaker. Both are traditionally achieved by manually specified key-frame animation. Automated tools for animating facial expressions and body gestures directly from speech would drastically ease the effort required, and allow non-experts to author higher quality character animations. Further on, such tools would enable users to drive real-time embodied conversational virtual avatars of themselves populating shared virtual spaces, and animate them with on-the-fly facial expressions and body gestures in tune with speech. In psycho-linguistics studies, it has been shown that user interfaces showing avatars with plausible facial expressions, body gestures, and speech are perceived as more believable and trustworthy [52]. It was also shown that non-verbal behavior is important for conveying information [16], providing a view into the speaker's internal state, and both speech and body gestures are tightly correlated, arising from the same internal process [24, 35].

Prior work on speech-driven virtual characters has been limited either to the generation of co-verbal body gestures through heuristic rule-based [34] or learning-based [11, 29, 30] approaches, or the generation of facial expressions [23] and head movements [42] in tune with speech. Many learning-based approaches use motion and gesture training data captured in a studio with complex motion capture systems [2, 11, 12, 28–30, 48]. In this way, it is hard to record large corpora of data reflecting gesture variation across subjects, or subject-specific idiosyncrasies revealing only in long term observation.

We propose the first approach to jointly generate synchronized conversational 3D gestures of the arms, torso, and hands, as well as a simple but expressive 3D face and head movement of an animated character from speech. It is based on the following contributions: **(1)** We contribute a new set of 3D training data annotations[1] from

[1]The dataset can be found in our project webpage: http://gvv.mpi-inf.mpg.de/projects/3d_speech_driven_gesture/

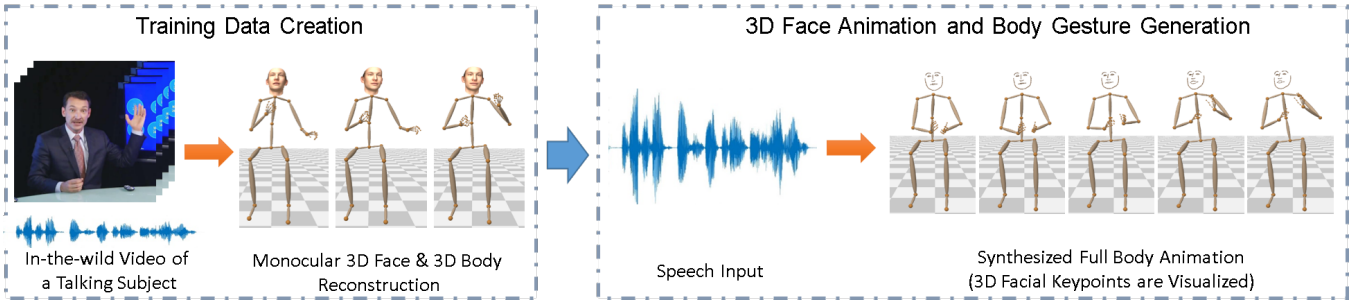

Figure 1: We propose the first approach to jointly synthesize the synchronous 3D conversational body gestures and 3D face animations of a virtual character from speech input. It is trained using our new contributed 3D facial expression, body, and hand pose annotation for a large corpus of in-the-wild video of talking people. Images courtesy of Dr. Mark Kubinec.

more than 33 hours of in-the-wild videos of talking subjects, which was used for learning a purely 2D gesturing model, without face expression synthesis, before [15]. To create ground truth, we apply monocular in-the-wild 3D body pose reconstruction [36], 3D hand pose reconstruction [55], and monocular dense 3D face reconstruction [14] on these videos. **(2)** We propose a CNN architecture that synthesizes face, body, and hand gestures from speech input. It has a common encoder for facial expression, body, and hands gesture which learns the inherent correlation between them and three decoder heads to jointly generate realistic motion sequences for face, body, and hands. In addition to facial expressions and head poses in tune with audio, it synthesizes plausible conversational gestures, such as beat gestures that humans use to emphasize spoken words, and gestures that reflect mood and personal conversational style. Note that, the goal is not to generate gestures relating to semantic speech content, or carrying specific language meaning, like in sign language. **(3)** Synthesis of body gestures is a multi-modal problem; several gestures could accompany the same utterance. To prevent convergence to the mean pose in training and ensure expressive gesture synthesis, the prior 2D work of Ginosar et al. [15] used adversarial training [17]. We improve upon this work by not only designing a discriminator that can measure whether the synthesized body and hand gestures look natural, but also the plausibility of the synthesized gestures when paired with the ground truth audio features. Figure 1 summarizes of our contribution. We evaluate our approach through extensive user study, where the participants rate our results as more natural and in-tune than the baseline methods. Please refer to our supplementary video for qualitative results.

2 RELATED WORK

Prior work looked at the problems of body gesture and face animation synthesis from audio input largely in separation. Problem settings differ, as conversational gesture synthesis is a much more multi-modal setting than speech-driven face animation where the viseme to phoneme mapping is much more unique. In this paper, we look at them in combination.

Speech-Driven Body Gestures and Head Motion. The prior art can be grouped into rule-based and data-driven methods. The seminal work by Cassell et al. [5] and Cassell [4] show that automatic body gesture and facial expression generation of a virtual character can be synchronized with the audio by using a set of manually defined rules. Other work incorporate linguistic analysis [6] into an extendable rule-based framework. Marsella et al.[34] develop a rule-based system to generate body gesture (and facial expression) by analyzing the content of the text input and audio. However, such methods heavily rely on the study of language-specific rules and cannot easily handle non-phoneme sounds.

To overcome these problems, data-driven approaches, which do not rely on expert knowledge in the linguistic domain, have attracted increasing attention. Neff et al. [37] propose a method to create a person-specific gesture script using manually annotated video corpora, given the spoken text and performer's gesture profile. The gesture script is then used to animate a virtual avatar. Levine et al. [30] use a complex motion capture setup to capture 45 minutes of training data and trained a Hidden Markov Model to select the most probable body gesture clip based on the speech prosody in real time. Levine et al. [29] mapped the audio signal into a latent kinematic feature space using a variant of Hidden Conditional Random Fields (CRF). The learned model is then used to select a gesture sequence via reinforcement learning approach. Marioooryad and Busso [33] use a combination of Dynamic Bayesian Networks (DBN) to synthesize head pose and eyebrow motion from speech. Sadoughi et al. [41] extended this approach by modeling discourse functions as additional constraints of the DBNs. Sadoughi et al. [42] use a learning-based approach that can leverage text-to-speech (TOS) system to synthesize head motion and propose a method that can solve the mismatch between real and synthetic speech during training. Chiu and Marsella [8] train a Conditional Restricted Boltzmann Machine (CRBM) to directly synthesize sequences of body poses from speech. Chiu and Marsella proposed using Gaussian Process Latent Variable Models (GPLVM) to learn a low dimensional embedding to select the most probable body gestures from a given speech input [9].

In recent years, deep learning has demonstrated its superiority in automatically learning discriminative features from big data. Bidirectional LSTM was used by Takeuchi et al.[47], Hasegawa et al.[20], and Ferstl and McDonnell [11] to synthesize body gestures from speech. Similarly, Haag and Shimodaira [18] used LSTM to synthesize head motion from speech. Kucherenko et al.[25] propose a denoising autoencoder to learn lower dimensional representation of body motion and then combines it with an audio encoder

to perform audio-to-gesture synthesis at test time. Lee et al. [28] contribute a large scale motion capture dataset of synchronized body-finger motion and audio, and propose a method to predict finger motion based on both audio and arm position as input. Ferstl et al. [12] use a multi-objective adversarial model and make use of a classifier that is trained to predict the gesture phase of the motion to improve gesture synthesis quality.

Recent work also try to incorporate text-based semantic information to improve generation quality of body gestures from speech [26, 53]. Alexanderson et al. [2] propose a normalizing flow-based generative model that can synthesize multiple plausible 3D body gesture from the same speech input and also allows some degrees of control to the synthesis. Ahuja et al.[1] show that a single learning-based mixture model can be trained to perform gesture style transfer between multiple speakers. In contrast, our focus is to find the best solution of predicting all relevant body modalities from audio using a single framework, which is a challenging problem even when trained in a person-specific manner.

Deep learning approaches typically require a large scale training corpus of audio and 3D motion pairs, which is usually captured with complex and expensive in-studio motion capture systems. To tackle this, Ginosar et al. [15] propose a learning-based speech-driven generation of 2D upper body and hand gesture model from a large scale in-the-wild video collection. With this solution, they are able to build an order of magnitude larger corpus from community video. Similarly, the method of Yoon et al. [54] was trained using ground truth 2D poses extracted from TED Talk videos via OpenPose [3]. Their model employs a Bidirectional LSTM to map audio input into a sequence of 2D human body pose. In our work, we contribute additional 3D face, hand, and body annotation for the dataset of Ginosar et al. [15]. Furthermore, in contrast to existing methods, we synthesize not only the 3D upper body and hand gestures, but also head rotation and facial expression of the speaker.

Speech-Driven Facial Expressions. Current techniques can be classified into: 1) face model-based [7, 10, 32, 38, 49, 50] and 2) non-model based. Model-based approaches parameterize expressions in terms of blendshapes, and estimate these parameters from the audio input. Non-model based approaches, however, directly map the audio into 3D vertices of a face mesh [23] or 2D point positions of the mouth [46]. In Karras et al. [23], an LSTM is used to learn this mapping, and in Suwajanakorn et al. [46], final photorelistic results are generated. Cudeiro et al. [10] use DeepSpeech voice recognition [19] to produce an intermediate representation of the audio signal. This is then regressed into the parameters of the FLAME face model [31]. Taylor et al. [49] use an off-the-shelf speech recognition method to map the audio into phoneme transcripts. A network is trained to translate the phonemes into the parameters of a reference face model. Tzikrakis et al. [50] use a Deep Canonical Attentional Warping (DCAW) to translate the audio into expression blendshapes. Pham et al. [38] directly maps the audio to the blendshape parameters even though their results suffer from strong jitter. While current audio-driven facial expression techniques produce interesting results, most of them show results on voice data recorded in controlled studios with minimal background noise [7, 10, 23, 32, 38, 49]. Cudeiro et al. [10] showed interesting results in handling different noise levels. Nevertheless, there is currently no audio-driven technique that estimates high quality facial expressions in-the-wild, as well as estimating the head motion and body conversational gestures. Our method uses the face model as the first category. In contrast to other methods, we apply a simple but effective approach to jointly learn the 3D head and face animation with body gestures, by directly regressing the facial parameters captured from a large corpus of community video.

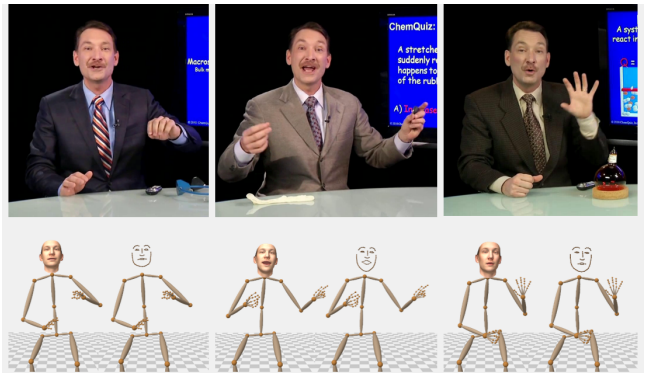

Figure 2: We annotate the in-the-wild conversational video corpus of Ginosar et al. [15] with 3D parameters of dense facial reconstruction, hand poses, and body poses. Images courtesy of Dr. Mark Kubinec.

3 DATASET CREATION

3.1 Creating 3D Annotations from Video

A major bottleneck for previous speech-driven animation synthesis work is the generation of sufficient training data. Many methods resort to complex in-studio capture of face and full-body motion with multi-camera motion capture systems. We therefore propose the first approach to extract automatic annotations of 3D face animation parameters, 3D head pose, 3D hands, and 3D upper body gestures from a large corpus of community video with audio. In this way, much larger training corpora spanning over long temporal windows and diverse subjects can be created more easily.

In particular, we use the dataset of Ginosar et al.[15] which features 144 hours of in-the-wild video of 10 subjects (e.g. talk show hosts) talking into the camera in both standing and sitting poses. From these videos, Ginosar et al.extracted 2D keypoints of the arms and hands, as well as 2D sparse face landmarks. They used a subset of these annotations to train a network synthesizing only 2D arm and finger motion from speech. While showing the potential of speech-driven animation, their approach does not synthesize 3D body motion; does not synthesize 3D motions of the torso, such as leaning, which is an element of personal speaker style; and does not predict 3D head pose and detailed face animation parameters. To train a method jointly synthesizing the latter more complete 3D animation parameters in tune with input speech, we annotate the dataset with state-of-the-art 3D face performance capture and monocular 3D body and hand pose estimation algorithms, see Fig. 2.

For monocular dense 3D face performance capture, we use the optimization-based tracker [14] that predicts parameters of a parametric face model, specifically: 64 expression blend shape coefficients, 80 PCA coefficients of identity geometry, 80 PCA coefficients of face albedo, 27 incident illumination parameters, and 6 coefficients for 3D head rotation and position. The face tracker expects tightly cropped face bounding boxes as input. We use the face tracker from Saragih et al.[44] for bounding box extraction and temporally filter the bouding box locations; we experimentally found this to be more stable than using the default 2D face landmarks in Ginosar's dataset for bounding box tracking. For training our algorithm, we only use the face expression coefficients $\theta_{Face} \in \mathbb{R}^{64}$ and head rotation coefficients $\mathbf{R} \in SO(3)$ (we use the 3D head position found by the body pose tracker).

For 3D body capture, we need an approach robust to body self-occlusions, occlusions by other people, occlusions of the body by a desk (sitting poses by talk show hosts) or occlusions by camera framing not showing the full body, even in standing poses. We therefore use the XNect [36] monocular 3D pose estimation approach designed to handle these cases. Specifically, in each video frame we extract 3D body keypoint predictions from *Stage II* of XNect for the 13 upper body joints (2 for head, 3 for each arm, 1 for neck, 1 for spine, 3 for hip/pelvis). This results in a 39-dimensional representation $\mathcal{K} \in \mathbb{R}^{39}$ for the body pose. We group the head rotation $\mathcal{R}$ predicted by the face tracker together with 3D the body keypoints $\mathcal{K}$ in a 42-dimensional vector $\theta_{Body} \in \mathbb{R}^{42}$.

To perform hand tracking, we employ the state-of-the-art monocular 3D hand pose estimation method of Zhou et al.[55]. To ensure good prediction results, we first tightly crop the hand images using the 2D hand keypoint annotations provided by Ginosar et al.[15] before feeding it to the 3D hand pose predictor. Since the hands can be occluded or out of view, we also employ an off-the-shelf cubic interpolation method to fill-in potentially missing 3D hand pose information. This results in 21 joints prediction for each hands, which we group into a 126-dimensional vector $\theta_{Hand} \in \mathbb{R}^{126}$.

To improve the robustness of our data, we exclude the data if the prediction confidence of the face landmarks or hand keypoints within a certain number of frames falls below a given threshold. This is obtained by reinterpreting the maximum value of the 2D joint heatmap prediction of the body parts produced by the tracker as a confidence measure. We also remove 4 out of 10 subjects provided by Ginosar et al.[15] due to the low resolution of the videos which lead to poor quality 3D dense face reconstruction results. Our final 3D dataset consists of more than 33 hours of videos from 6 subjects. We use the same training, validation, and test split as the original 2D dataset, which make up to around 80%, 10%, and 10% of the total data respectively, even after accounting for the excluded data.

We temporally smooth our 3D body and hand pose prediction as well as the head rotation results using a Gaussian filter with a standard deviation of $\sigma = 1.5$ to improve visual quality of our output. The same filter is also applied to the ground truth sequences in our video results.

3.2 Audio features pre-preprocessing

Similar to Suwajanakorn et al.[46], we compute the MFC coefficients of each input video frame after normalizing the audio using FFMPEG [13, 39]. We make use of CMU Sphinx [27] for computing the coefficients, and use 13 MFC coefficients and an additional feature to account for the log mean energy of the input. These, together with their temporal first derivatives, yield a 28-dimensional vector $\mathcal{F}_{MFC} \in \mathbb{R}^{28}$ representing the speech input at each time step. MFCC encodes the characteristics of how human speech is perceived, which make it useful for a wide range of applications such as speech recognition. Encoding the characteristics of speech perception make MFC coefficients a good representation for predicting facial expressions because modulation of face shapes is a part of the speech production process. For predicting body gestures, the change of MFCC features over the sequence carries the rhythm information needed to produce beat gestures.

4 METHOD

Our approach produces a temporal sequence of 3D facial expression parameters, head orientation, 3D body, and 3D hand pose keypoints given a speech signal as input. Temporal variations in these aforementioned parameters contain the gestural information. As described in section 3.2, the speech input is pre-processed to yield MFC based feature frames $\mathcal{F}_{MFC}[t] \in \mathbb{R}^{28}$ for each discrete time step t. We indicate the facial expression parameters at each time step as $\theta_{Face}[t] \in \mathbb{R}^{64}$, 3D keypoints for both hands as $\theta_{Hand}[t] \in \mathbb{R}^{126}$, and the head orientation and 3D body keypoints are represented together as $\theta_{Body}[t] \in \mathbb{R}^{42}$. The temporal sequences are sampled at $15Hz$.

4.1 Network Architecture

Similar to other adversarial learning-based approaches, our model consists of 2 main neural networks that we refer as the generator network G and discriminator network D. We follow the design of prior human motion and gesture synthesis approaches [15, 21, 22] by using 1D convolutional networks to model the temporal relationship of the audio and body features across different frames.

We employ a 1D convolutional Encoder-Decoder architecture for the generator network G to map the input audio feature sequence $\mathcal{F}_{MFC}[0:T]$ to 3D face expression parameter sequence $\theta_{Face}[0:T]$, 3D body pose parameter sequence $\theta_{Body}[0:T]$, and 3D hand pose parameter sequence $\theta_{Hand}[0:T]$ which is also trained in a supervised manner.

Our 1D convolutional architecture for the generator G is adapted from a reference implementation [51] of the U-Net [40] architecture originally proposed for 2D image segmentation. Our architecture utilizes a single encoder, comprised of 8 1D [Conv-BN-ReLU] blocks with a kernel size of 3, and is interleaved with MaxPool after every second block except the last. The last block is followed by an upsampling layer (nearest neighbour). Each of face, body, and hand sequences utilize a separate decoder to learn body-part specific motion characteristics. The decoders are symmetric with the encoder, and comprised of 7 1D [Conv-BN-ReLU] blocks and a final 1D convolution layer, interleaved with upsampling layers after every second block. The decoders, being symmetric with the encoder, utilize skip connectivity from the corresponding layers in the encoder. The discriminator network is designed to predict whether its input audio and pose features are real or not. This network comprised of 6 1D [Conv-BN-ReLU] blocks with a kernel size of 3, and

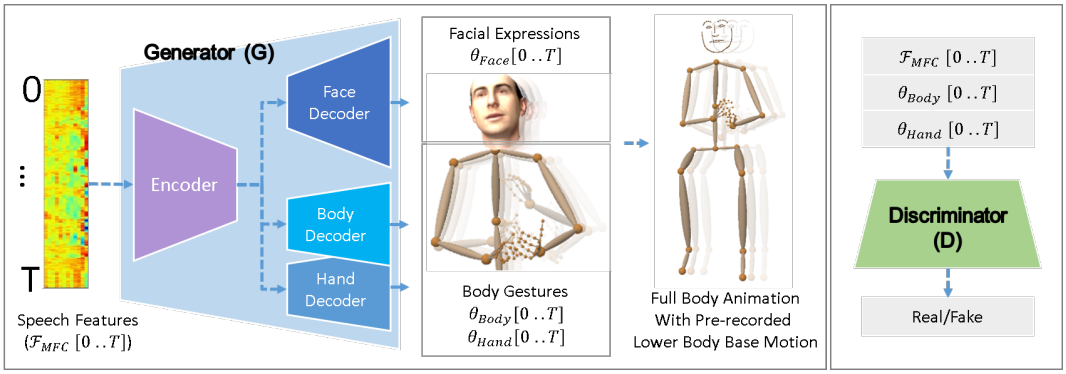

Figure 3: Our approach produces a temporal sequence of 3D facial expression parameters, head orientation, and 3D keypoints of the upper body and hands given a speech signal as input. We employ an adversarial loss in which the discriminator network tries to distinguish whether the input audio and body pose features are real or generated by the generator network.

is interleaved with MaxPool after every second block. Afterwards, it is followed a linear and sigmoid activation layers.

A schema of the architecture is shown in Figure 3.

4.2 Training Details

For each sequence of sampled speech features $\mathcal{F}_{MFC}[0 .. T-1]$, and annotated 3D face expression parameter sequence $\theta_{Face}[0 .. T-1]$, 3D body pose parameter sequence $\theta_{Body}[0 .. T-1]$, and 3D hand pose parameter sequence $\theta_{Hand}[0 .. T-1]$, we extract 64 frame ($\approx$ 4sec) sub-sequences in a sliding window manner, with a 1 to 5 frame overlap between consecutive sub-sequences depending on the number of data points of the subject. Each mini-batch for training comprises of a random sampling of such 64-frame sub-sequences extracted from all training sequences. We use Adam for training, with a learning rate of $5e-4$, a mini-batch size of 25, and trained until 300,000 iterations per subject. Since the generator network is fully convolutional, at deployment our network can handle input speech features of arbitrary duration.

We supervise our generator network G with the following loss terms:

$$\mathcal{L}_{Reg} = w_1 * \mathcal{L}_{Face} + w_2 * \mathcal{L}_{Body} + w_3 * \mathcal{L}_{Hand}. \quad (1)$$

$\mathcal{L}_{Face}$ is the L2 error of facial expression parameters

$$\mathcal{L}_{Face} = \sum_{t=0}^{T-1} \|\theta_{Face}[t] - \hat{\theta}_{Face}[t]\|_2.$$

$\mathcal{L}_{Body}$ is the L1 error of 3D body keypoint locations and head orientation, and $\mathcal{L}_{Hand}$ is the L1 error of 3D hand keypoint locations

$$\mathcal{L}_{Body} = \sum_{t=0}^{T-1} \|\theta_{Body}[t] - \hat{\theta}_{Body}[t]\|_1,$$

$$\mathcal{L}_{Hand} = \sum_{t=0}^{T-1} \|\theta_{Hand}[t] - \hat{\theta}_{Hand}[t]\|_1.$$

We define $w_1 = 0.37$, $w_2 = 600$, and $w_3 = 840$ to ensure that each term is equally weighted during training.

In practice, we observe that only employing L1 or L2 error for body keypoints results in less expressive gestures, as has also been pointed out in prior work on 2D body gesture synthesis of Ginosar et al. [15]. Inspired by the adversarial training approach of Ginosar et al.[15], we show that incorporating an adversarial loss using a discriminator network D which is trained to judge whether an input pose is real or fakely generated by the generator G, can lead to more expressive gestures that are also in-sync with the speech input. When trained together with the generator network in a minimax game scenario, it will push the generator to produce a higher quality 3D body and hand pose synthesis in order to fool the discriminator. We follow similar approach to the work of Ferstl et al.[12] by using not only the pose, but also the audio features as input to the discriminator. This way, the discriminator is not only tasked to measure if the input gesture looks real, but it also needs to determine if the gesture is in-sync with the input audio features or not. Since the multi-modality of the body gestures mainly occurs for the body and hands, we exclude the facial expression parameters from the adversarial loss formulation:

$$\mathcal{L}_{Adv}(G, D) = \mathbb{E}_{\mathcal{F}_{MFC}}[\log(1 - D(\mathcal{F}_{MFC}, G^*(\mathcal{F}_{MFC})))] + \mathbb{E}_{\mathcal{F}_{MFC},\theta_{Body},\theta_{Hand}}[\log D(\mathcal{F}_{MFC}, \theta_{Body}, \theta_{Hand})] \quad (2)$$

where G^* indicates that we only use the predicted θ_{Body} and θ_{Hand} outputs of the original generator network G.

Combined with the direct supervision loss, our overall loss is:

$$\mathcal{L} = \mathcal{L}_{Reg} + w \cdot \min_D \max_G \mathcal{L}_{Adv}(G, D) \quad (3)$$

where w is set to be 5.

Our networks are trained on subject specific training sets in order to capture the particular gesture characteristics of the subject.

5 RESULTS

Our proposed approach addresses essential aspects of animating virtual humans: synthesizing facial expressions, body, and hand gestures in tune with speech. For visualization of our results, as well as for the user study, to allow observers to focus on the face and body motion, we render an abstract 3D character that showcases all the important skeletal and facial elements without the risk of falling in the uncanny valley, following similar approaches in prior

work [15, 30]. Since our approach only predicts upper body motion, we fuse it with a pre-recorded base motion of the lower body in both sitting and standing scenarios.

Since the synthesis of conversational gestures is a multi-modal problem, direct comparison with the tracked annotations would not be meaningful for all aspects of the synthesized results, particularly for evaluating the realism of the synthesized gestures. We evaluate our method through extensive user studies to judge the quality and the plausibility of our results, and we compare it to various baselines. Further, we measure the prediction of the facial expressions by comparing the 3D lip keypoints extracted from selected vertices of the predicted dense 3D face model against the automatically generated ground truth lip keypoints that we obtained from the source image. A qualitative example of our synthesis result is shown in Figure 4. Please see the accompanying video for extensive audio-visual results.

Table 1: User study result measuring both the naturalness and synchronization between the synthesized face+body+hand gesture and speech.

Method	Naturalness	Synchrony.
Ground truth	**4.29 ± 0.86**	**4.39 ± 0.77**
Conv.net. direct regress.	3.54 ± 1.11	3.78 ± 1.08
LSTM (adapted from [45])	3.15 ± 1.03	3.21 ± 1.11
Adv. loss on veloc. (adapted from [15])	3.03 ± 0.98	3.38 ± 0.95
Adv. loss on audio+3D pose (ours)	**4.05 ± 0.85**	**4.00 ± 0.91**

5.1 Baseline Comparisons

We evaluate our approach against other methods that perform body gesture prediction which use audio features as input. Other baseline methods are trained using the same MFCC features described in 3.2. Our first baseline is the direct regression 1D CNN model of our proposed network architecture without using the adversarial loss. Next, we compare our method against a Recurrent Neural Network (RNN)-based Long Short-term Memory (LSTM) architecture by Shlizerman et al.[45] which was originally designed to temporally predict 2D hand and finger poses. Since the original method was not designed to handle multi-modal data, we train three LSTM models for face, body, and hand gesture separately on our 3D data.

We trained an adaptation of Ginosar et al.[15] using our proposed model and trained the adversarial loss to distinguish between the real and fake synthesis of the gesture in the velocity space similar to their proposed approach and use this version as our baseline comparison. We also compare our method against the work of Alexanderson et al.[2] by retraining their method on our in-the-wild 3D data. Their model was originally trained on clean mocap data of 3D body pose without face or hand annotations. We found that the model is sensitive to the hyperparameters used. Because of this, we decided to only train it on the body and hand data to simplify the problem. We manually searched for an optimal set of hyperparameters that can produce the best results in terms of naturalness and synchronization based on the recommendation of the authors. Following their instruction, we conducted multiple experiments by varying the number of units H between 512, 700, and 800 and the number of flow-steps K from 8 up to 16. We found that the MoGlow-based model produces the best results when using the number of units $H = 800$ and number of steps $K = 10$.

Please also refer to our supplementary video for the qualitative results of the baseline methods.

5.2 Gesture Synthesis User Study Evaluation

We conducted two separate user studies for the qualitative evaluation of our proposed method.

For the first user study, we compare methods that synthesize the 3D face, body, and hand gestures from audio. In this study, the participants were shown 3 out of 6 randomly selected video sequences (12 seconds/sequence) synthesized by our proposed method, baselines, and the ground truth (tracked) annotations. This study involved 67 participants. Each user was asked to judge the naturalness and the synchronization between the audio and the generated 3D face and body gestures on a scale of 1 to 5, with 5 being the most plausible and 1 being the least plausible. As shown in Table 1, the ground truth sequences are perceived as both the most natural and in-tune with the input speech compared to other synthesized gesture videos, which is rated at 4.29 ± 0.86 and 4.39 ± 0.77, respectively. Compared to other baseline methods, the participants agree that our results look more natural and in tune with the speech audio with the score of 4.05 ± 0.85 in term of naturalness and 4.00 ± 0.91 in terms of synchronization with the speech.

Table 2: User study result measuring both the naturalness and synchronization between the synthesized body+hand gesture and speech. The users were specifically asked to ignore the quality of the facial expression.

Method	Naturalness	Synchrony.
MoGlow [2])	2.88 ± 1.02	3.11 ± 1.13
Ours	**4.01 ± 0.82**	**3.93 ± 0.92**

We also conducted a second user study evaluating only the synthesis of the 3D body and hand gestures and compare our method with the MoGlow-based model of Alexanderson et al.[2]. For this study, the participants were specifically asked to ignore the quality of the face expressions in the video. To ensure a fair comparison, all videos presented in this study were synthesized by using the 3D facial expression predicted by our method. Similar to the first study, each of the 45 participants were asked to rate the quality of the gestures from 3 out of 6 possible videos for each method on a scale between 1 to 5. As shown in Table 2, our method is rated as both more natural and in-sync with the audio.

5.3 Facial Expression Evaluation

In Table 3, we compare the 3D lip keypoints of the generated face vertices corresponding to the facial expressions predicted by various approaches against the image-based face tracker's 3D lip keypoints in a neutral head pose. The comparison was performed on the whole test set which consists of 578 sequences (12 seconds/sequence) across all subjects. As a sanity check baseline, we also compute the

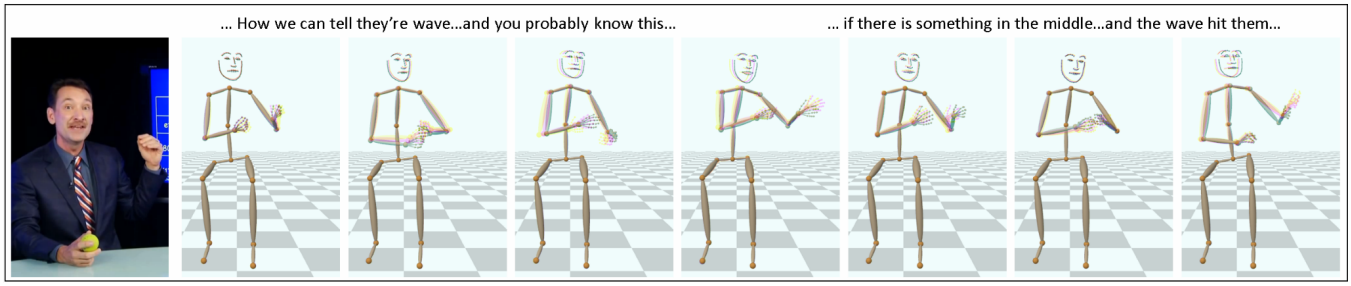

Figure 4: Qualitative results of our approach. Motion visualization is based on the Harris shutter effect. Image courtesy of Dr. Mark Kubinec.

Table 3: Quantitative comparison to baseline methods for lip motion prediction error against the ground truth (in mm).

Method	Oliver	Meyers	Ellen	Kubinec	Stewart	O'Brien
Conv. network direct regression	**0.29**	**0.35**	0.29	**0.28**	**0.39**	**0.37**
LSTM (adapted from [45])	0.3	0.36	0.30	0.32	0.41	0.39
Adv. loss on velocity (adapted from [15])	**0.29**	**0.35**	0.3	0.29	0.39	0.38
Random	0.49	0.57	0.47	0.43	0.57	0.52
Adv. loss on audio+3D pose (ours)	**0.29**	**0.35**	**0.28**	**0.28**	**0.39**	**0.37**

difference between the optimization-based tracked annotations of one sequence, to the optimization-based annotations on a *different* sequence *chosen randomly*. The evaluation shows that our proposed method achieves similar or slightly better performance against other proposed baselines. This result also demonstrate that our unified whole-body architecture is suitable for simultaneous face expression synthesis at decent quality, and better than simultaneous face synthesis with other body gesture synthesis architectures. Note here that we are not claiming that our design advances the state-of-the-art in face-only expression synthesis. This is outside the scope of our work and left for future work.

6 DISCUSSION

Although mouth expressions are strongly correlated with speech, the rest of the intended generation targets such as body gestures do not have a one-to-one mapping. Coupled with the noisy nature of our monocular data, as observed in our experiments, this multi-modal nature of the problem makes both designing and analyzing a stable expressive model challenging. We also observe that a lower value of $L1$ or $L2$ loss on the validation set does not always guarantee to produce a qualitatively better gesture synthesis, which further shows the importance of the adversarial loss. The data is also inherently noisy due to the use of 3D monocular trackers, which may to lead to jittery 3D motion that can affect the performance of our model and comparison baselines. However, we observe the effect of the noise to be minimal as it can be suppressed by applying temporal filters to the prediction output.

We also argue that the discriminator network can potentially be used as a plausibility metric to rate the quality of a gesture synthesis from speech, similar to how the inception score is used [43], if trained with enough gesture and noise variations. One way to validate this idea is to train the model to classify whether its audio-gesture pair input is in-sync or off-sync. The ground truth audio-gesture pairs can be used directly as in-sync (positive) samples, while off-sync (negative) samples can be prepared by pairing the audio sequence with a different gesture from a random pair.

When we train our discriminator network in this setup, it can reliably classify unseen test pairs of subject "Oliver" with a high accuracy of 87.4%. Unfortunately, since the classifier is trained only on ground-truth motion sequences, it is not yet possible to extend this model as a quantitative metric for gesture synthesis methods. When we tested this classifier on the baseline models, it produces inconsistent results. For example, it rates our proposed model to be more plausible than the ground truth sequences, which contradicts the result of the user study. A specific dataset containing different gesture noise characteristic may be required if we want to extend this classifier into a more general gesture plausibility metric.

7 CONCLUSION

We propose the first approach for full 3D face, body, and hand gesture prediction from speech to automatically drive a virtual character or an embodied conversational agent. We leverage monocular dense face reconstruction and body pose reconstruction approaches on in-the-wild footage of talking subjects to acquire training data for our learning-based approach, generating 3D face, body, and hand pose annotations for $\approx$ 33 hours of footage. Our key insight on incorporating an adversarial penalty not only on the 3D pose but also its combination with the audio input allows us to successfully generate expressive body gestures that are in-sync with the speech.

REFERENCES

[1] Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. 2020. Style Transfer for Co-speech Gesture Animation: A Multi-speaker Conditional-Mixture Approach. In *ECCV*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.).

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

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

[4] Justine Cassell. 2000. Embodied Conversational Interface Agents. *Commun. ACM* (2000).

[5] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated Conversation: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*.

[6] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. *BEAT: the Behavior Expression Animation Toolkit.*

[7] Y. Cha, T. Price, Z. Wei, X. Lu, N. Rewkowski, R. Chabra, Z. Qin, H. Kim, Z. Su, Y. Liu, A. Ilie, A. State, Z. Xu, J. Frahm, and H. Fuchs. 2018. Towards Fully Mobile 3D Face, Body, and Environment Capture Using Only Head-worn Cameras. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2018).

[8] Chung-Cheng Chiu and Stacy Marsella. 2011. How to Train Your Avatar: A Data Driven Approach to Gesture Generation. In *IVA*, Hannes Högni Vilhjálmsson, Stefan Kopp, Stacy Marsella, and Kristinn R. Thórisson (Eds.).

[9] Chung-Cheng Chiu and Stacy Marsella. 2014. Gesture Generation with Low-dimensional Embeddings. In *AAMAS*.

[10] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. *CVPR* (2019).

[11] Ylva Ferstl and Rachel McDonnell. 2018. Investigating the Use of Recurrent Motion Modelling for Speech Gesture Generation. In *IVA*.

[12] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2019. Multi-Objective Adversarial Gesture Generation. In *Motion, Interaction and Games*.

[13] FFmpeg Developers. 2016. FFMPEG. ffmpeg.org. ffmpeg.org

[14] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. [n.d.]. Reconstruction of Personalized 3D Face Rigs from Monocular Video. *ACM Trans. Graph.* ([n. d.]).

[15] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. 2019. Learning Individual Styles of Conversational Gesture. In *CVPR*.

[16] Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences* (1999).

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*.

[18] Kathrin Haag and Hiroshi Shimodaira. 2016. Bidirectional LSTM Networks Employing Stacked Bottleneck Features for Expressive Speech-Driven Head Motion Synthesis. In *IVA*.

[19] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep Speech: Scaling up end-to-end speech recognition. arXiv:1412.5567

[20] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network. In *IVA*.

[21] Daniel Holden, Jun Saito, and Taku Komura. 2016. A Deep Learning Framework for Character Motion Synthesis and Editing. *ACM Trans. Graph.* (2016).

[22] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. 2015. Learning Motion Manifolds with Convolutional Autoencoders. In *SIGGRAPH Asia 2015 Technical Briefs*.

[23] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven Facial Animation by Joint End-to-end Learning of Pose and Emotion. *ACM Trans. Graph.* (2017).

[24] Adam Kendon. 2004. *Gesture: Visible Action as Utterance.* Cambridge University Press.

[25] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing Input and Output Representations for Speech-Driven Gesture Generation. In *IVA*.

[26] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A Framework for Semantically-Aware Speech-Driven Gesture Generation. In *ICMI*.

[27] Paul Lamere, Philip Kwok, Evandro Gouvêa, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf. 2003. The CMU SPHINX-4 Speech Recognition System.

[28] G. Lee, Z. Deng, S. Ma, T. Shiratori, S.S. Srinivasa, and Y. Sheikh. 2019. Talking With Hands 16.2M: A Large-Scale Dataset of Synchronized Body-Finger Motion and Audio for Conversational Motion Analysis and Synthesis. In *ICCV*.

[29] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. 2010. Gesture controllers. In *ACM Trans. Graph.*

[30] Sergey Levine, Christian Theobalt, and Vladlen Koltun. 2009. Real-time Prosody-driven Synthesis of Body Language. *ACM Trans. Graph.* (2009).

[31] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics* (2017).

[32] Yilong Liu, Feng Xu, Jinxiang Chai, Xin Tong, Lijuan Wang, and Qiang Huo. 2015. Video-audio Driven Real-time Facial Animation. *ACM Trans. Graph.* (2015).

[33] S. Mariooryad and C. Busso. 2012. Generating Human-Like Behaviors Using Joint, Speech-Driven Models for Conversational Agents. *IEEE Transactions on Audio, Speech, and Language Processing* (2012).

[34] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. [n.d.]. Virtual Character Performance from Speech. In *SCA*.

[35] David McNeill. 2000. *Language and Gesture.* Cambridge University Press.

[36] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. 2020. XNect: Real-time Multi-Person 3D Motion Capture with a Single RGB Camera. *ACM Trans. Graph.* (2020).

[37] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture Modeling and Animation Based on a Probabilistic Re-creation of Speaker Style. *ACM Trans. Graph.* (2008).

[38] Hai Xuan Pham, Yuting Wang, and Vladimir Pavlovic. 2018. End-to-end Learning for 3D Facial Animation from Speech. In *ICMI*.

[39] Werner Robitza. 2019. ffmpeg-normalize. github.com/slhck/ffmpeg-normalize. https://github.com/slhck/ffmpeg-normalize

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*.

[41] Najmeh Sadoughi, Yang Liu, and Carlos Busso. 2014. Speech-Driven Animation Constrained by Appropriate Discourse Functions. In *ICMI*.

[42] Najmeh Sadoughi, Yang Liu, and Carlos Busso. 2017. Meaningful head movements driven by emotional synthetic speech. *Speech Communication* (2017).

[43] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. 2016. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*.

[44] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. 2011. Deformable Model Fitting by Regularized Landmark Mean-Shift. *IJCV* (2011).

[45] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. 2018. Audio to body dynamics. In *CVPR*.

[46] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: Learning Lip Sync from Audio. *ACM Trans. Graph.* (2017).

[47] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. 2017. Speech-to-Gesture Generation: A Challenge in Deep Learning Approach with Bi-Directional LSTM. In *Proceedings of the 5th International Conference on Human Agent Interaction*.

[48] Kenta Takeuchi, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa, and Hiroshi Sakuta. 2017. Creating a Gesture-Speech Dataset for Speech-Based Automatic Gesture Generation. In *HCI International – Posters' Extended Abstracts*, Constantine Stephanidis (Ed.).

[49] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A Deep Learning Approach for Generalized Speech Animation. *ACM Trans. Graph.* (2017).

[50] Panagiotis Tzirakis, Athanasios Papaioannou, Alexander Lattas, Michail Tarasiou, Björn W. Schuller, and Stefanos Zafeiriou. 2019. Synthesising 3D Facial Motion from "In-the-Wild" Speech. *CoRR* abs/1904.07002 (2019).

[51] Naoto Usuyama. 2018. github.com/usuyama/pytorch-unet. github.com/usuyama/pytorch-unet

[52] Susanne Van Mulken, Elisabeth André, and Jochen Müller. 1998. The persona effect: How substantial is it? In *People and computers XIII*. Springer, 53–66.

[53] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Trans. Graph.* (2020).

[54] Y. Yoon, W. Ko, M. Jang, J. Lee, J. Kim, and G. Lee. 2019. Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots. In *ICRA*.

[55] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. 2020. Monocular Real-time Hand Shape and Motion Capture using Multi-modal Data. In *CVPR*.