# Real-time Gesture Animation Generation from Speech for Virtual Human Interaction

Manuel Rebol
mrebol@american.edu
American University, Graz University Of Technology
Graz, Austria

Christian Gütl
c.guetl@tugraz.at
Graz University Of Technology
Graz, Austria

Krzysztof Pietroszek
pietrosz@american.edu
American University
Washington, USA

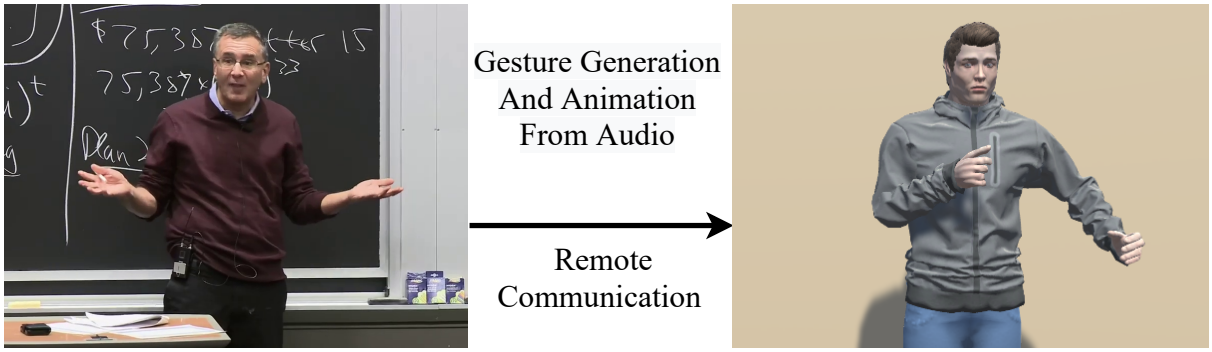


**Figure 1: Real-time gesture generation for professor Jonathan G. On the left, we show the original pose in a video. Our system captures the audio and predicts gestures as shown in the right image. Our system can be used in remote communication scenarios where no video capture is possible.**

## ABSTRACT

We propose a real-time system for synthesizing gestures directly from speech. Our data-driven approach is based on Generative Adversarial Neural Networks to model the speech-gesture relationship. We utilize the large amount of speaker video data available online to train our 3D gesture model. Our model generates speaker-specific gestures by taking consecutive audio input chunks of two seconds in length. We animate the predicted gestures on a virtual avatar. We achieve a delay below three seconds between the time of audio input and gesture animation. Code and videos are available at https://github.com/mrebol/Gestures-From-Speech

## CCS CONCEPTS

• **Human-centered computing** → *Interaction design*; • **Applied computing** → *Media arts*; • **Computing methodologies** → **Virtual reality**.

## KEYWORDS

Gestures, Animation, NUI


*CHI '21, May 08–13, 2021, Yokohama, Japan*

ACM ISBN 978-1-4503-8095-9/21/05.
https://doi.org/10.1145/3411763.3451554

**ACM Reference Format:**
Manuel Rebol, Christian Gütl, and Krzysztof Pietroszek. 2021. Real-time Gesture Animation Generation from Speech for Virtual Human Interaction. In *CHI '21: The ACM CHI Conference on Human Factors in Computing Systems, May 08–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3411763.3451554

## 1 INTRODUCTION

Equipping digital representations of humans with non-verbal communication skills finds many applications in human-computer systems, from telehealth to entertainment to education, because non-verbal communication is an essential component of communication process.

In our previous work, we tackled the problem of unrealistic human-agent communication by introducing a generative model that synthesizes natural-looking gestures [5]. In this work, we expand the proposed model to generate gestures from speech in real time. Our approach utilizes the large amount of video data available to train a model to learn the relationship between speech and gesture. In contrast to our previous work work [5] and recent related work [2], we extract and predict 3D gestures in real time, with relatively small latency.

Our models are speaker-specific because body language is specific to every individual. We train a model individually for every speaker for that we generated a dataset. Our model also captures the individual gesturing style of speakers from different genres. We animated the individual gesture predictions on a virtual avatar to provide a natural human-like representation to the observer.

Our real-time gesture generation system can be used in remote communication applications. Whenever there exists no possibility to video capture the speaker, our system can be used to animate a virtual human instead of capturing the speaker. By having both audio and a visual representation, a more vivid communication experience is achieved. We support to move the communication into Virtual Reality because we animate the generated gestures in a 3D environment. Our system can also be used in low-bandwidth remote communication where sending audio is possible, but video is not supported. In this scenario, the remote communication partner would be able to watch at a virtual representation of the speaker.

## 2 OUR APPROACH

Our approach is divided into three core components. Our input component provides the GAN with the gestures. We estimate the 2D human pose in each frame of input video sequences. Once the 2D human pose is estimated, it is projected into 3D and handed over to the GAN. The main component contains the gesture generation GAN. Our GAN takes two inputs during training, the raw audio of the speech and the gestures in human pose format. We extract the raw audio from our video dataset. The third component animates the generated gestures on a character in virtual reality.

### 2.1 Extracting Gestures from Video

We extract the training data for our GAN model from a readily-available large source of data: 2D videos of people performing lectures and speeches. Recent advancements in 3D human pose estimation from 2D video allow us to process many online videos of lectures and speeches and extract 3D body poses and gestures from them.

We encode the gestures from video in the 3D human pose format. We take precautions to ensure high quality of data, both before and after extracting the body language from the video. We eliminate those fragments of the videos where the hands and upper body of the speaker is not entirely visible, *e.g.* because it is being partially occluded by an object.

We approach the gesture extraction in two steps. In the first step, we extract the 2D Human Pose from the raw video using the OpenPose framework [1, 7]. In the second step, we project the 2D pose into 3D space for the large body parts and hands separately. For the 3D body pose estimation we use the model implemented by Pavllo *et al.* [4]. For the 3D hand pose estimation we implement a model similar to Zimmerman *et al.* [9]. As a result of the process, we created a large dataset of motion captures that correspond to the gestures used by humans when speaking.

### 2.2 Generating Gestures from Speech

We implement the Generative Adversarial Neural Network (GAN) [3] framework which allows us to model the multimodal task of predicting gestures from speech. The GAN framework consists of the gesture generator $G$ and the motion discriminator $D$.

*Gesture Generator.* The objective of the gesture generator is twofold. First, the generator is trained to predict gestures close to ground truth gestures extracted from video input. Second, the main objective of the generator inside the GAN framework is to fool the discriminator. In our case, the discriminator is fooled by predicted gestures with realistic motion. We implement a UNet [6] architecture to generate gestures encoded in the human pose format from speech. Inside the UNet, the skip connections forward low-level prosodic features extracted from the input audio. These features are necessary to predict smaller beat gestures. The bottleneck extracts high-level features that contain information about long input sequences. This is useful to predict the posture of the speaker.

The objective of the generator is enforced by a regression loss on the prediction given pseudo ground truth gestures. The loss function for the generator is defined as

$$\mathcal{L}_{\text{Gen}}(G) = \mathbb{E}_{(\mathbf{s},\mathbf{p})}\left[||\mathbf{p} - G(\mathbf{s})||_1\right] + \tag{1}$$

$$\lambda_{\text{bone}}\, \mathbb{E}_{(\mathbf{s})}\left[||B(G(\mathbf{s}_t)) - B(G(\mathbf{s}_{t-1}))||_1\right], \tag{2}$$

where vector $\mathbf{s}$ refers to the input speech and vector $\mathbf{p}$ refers to the pseudo ground truth body keypoints. The function $B$ computes the bone length which is computed by the euclidean distance between pairs of keypoints at consecutive time steps $t$ and $t-1$. Therefore, the second term in Equation 1 ensures that the bone length stays constant over time. The first term ensures that the predicted output matches the ground truth gestures extracted from the video. The hyperparameter $\lambda_{\text{bone}} \in (0, 1)$ is used to weight the importance of constant bone length in the prediction.

*Motion Discriminator.* Our discriminator ensures that the motion of the generated gestures is similar to the motion extracted from video to avoid regression towards the mean gesture. We compute the motion between consecutive time steps by subtracting the keypoint positions:

$$M(\mathbf{v}) = \mathbf{v}_t - \mathbf{v}_{t-1}\ . \tag{3}$$

The discriminator receives either a real or a fake gesture sequences as input. The fake gesture sequence is predicted by the generator, whereas the real gesture sequence is directly obtained from the input video. The discriminator is trained to distinguish real and generated gestures. The complete GAN loss function including the discriminator $D$ is defined as

$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{(\mathbf{p})}\left[\log D(M(\mathbf{p}))\right] + \mathbb{E}_{(\mathbf{s})}\left[\log(1 - D(M(G(\mathbf{s}))))\right]\ . \tag{4}$$

The objective of the discriminator is to maximize this function. Consequently, the term loss is only true with respect to the generator. The discriminator learns to output $D(\cdot) \rightarrow 1$ if input motion is real and $D(\cdot) \rightarrow 0$ if the input motion is generated.

*GAN Objective.* We train the parameters of our model by combining the loss functions shown in Equation 1 and Equation 4. The final objective function is defined as

$$\min_G \max_D \mathcal{L}_{\text{GAN}}(G, D) + \mathcal{L}_{\text{Gen}}(G). \tag{5}$$

The generator $G$ has the objective to minimize this function whereas the discriminator $D$ aims to maximize $\mathcal{L}_{\text{GAN}}$.

### 2.3 Animating Gestures in Virtual Reality

Our Gesture GAN model produces 3D Human Pose sequences which we animate on virtual humans using rotation angles between bones and inverse kinematic computation. By connecting the keypoints predicted by our Gesture GAN, we create a skeletal

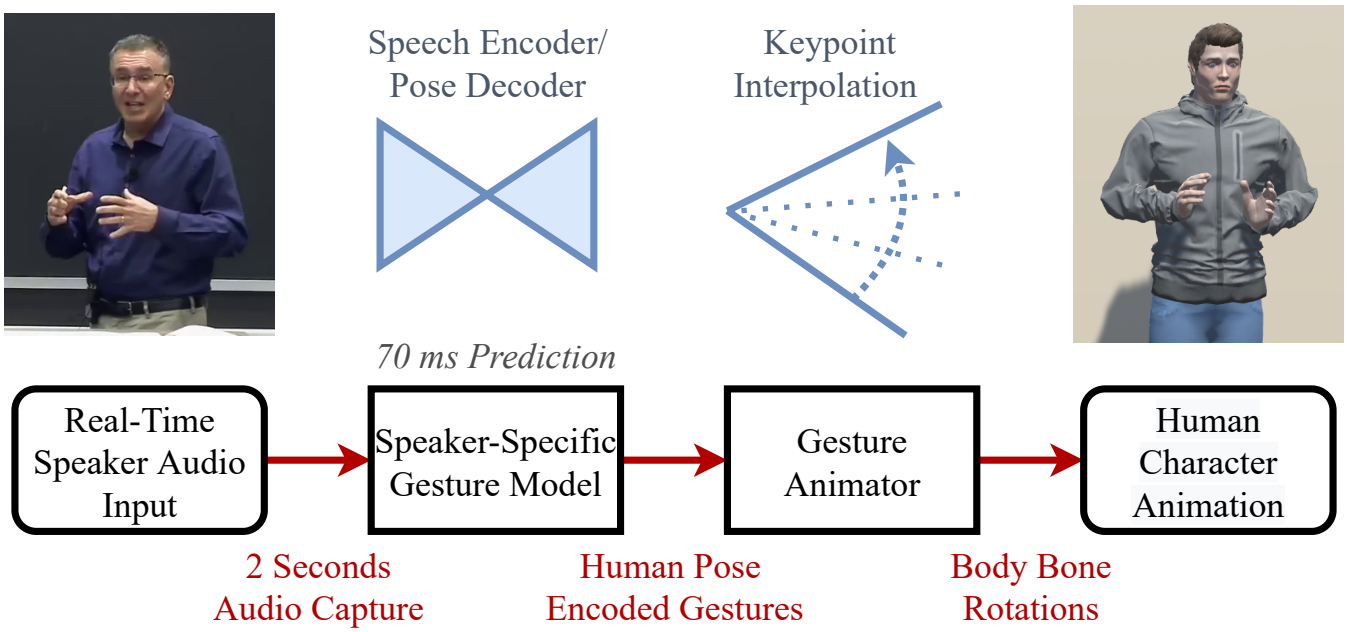


**Figure 2: System diagram. Our real-time gesture prediction system takes chunks of two seconds speaker audio as an input and animates the predicted gestures on a virtual human. In this process, a speaker-specific gesture model generates gestures in the human pose format. The gesture animator translates the encoded gestures into a 3D animation.**

representation of the human pose. We use this skeletal representation of the gestures to animate an avatar in virtual reality.

One challenge when transferring the predicted 3D Human Pose into a 3D animation is the missing information about anatomical details and ambiguities. Since we animate the gestures on a virtual human of different size and shape compared to the original speaker, we omit the information about the bone length of the generated skeleton. Instead, we only consider the rotation between the bones. To tackle the problem of missing rotation information, we use inverse kinematics on the human pose input.

Besides recovering missing rotation information, we also tackle implausible motion predictions. Although our Gesture GAN is trained to predict motion which is similar to real human motion, there exists no constraint which particularly enforces anatomically plausible motion. Hence, the predicted motion in some cases appears artificial, especially when animated on a virtual human avatar. This problem is very distracting for the viewer of the animation because anatomically implausible motion is quickly noticed. To overcome this issue, we introduce motion constraints on arms and fingers to filter out implausible motion. In addition to the motion constraints, we apply motion smoothing. Both optimizations combined produce an appealing animation.

## 2.4 Training the GAN

For the purpose of training our GAN, we generate a gesture dataset of over one hundred hours extracted from videos of four speakers from two domains. Specifically, we generated 72 hours of motion capture data for television show hosts Oliver and Ellen as well as 64 hours of motion capture data for professor Jonathan G. and Shelly K. We encode the gestures in the efficient human pose format. By using this format, we ignore irrelevant information such as background and the shape of different body parts of the speaker.

We evaluate the predictions of our Gesture GAN on our validation set using the Percent of Correct Keypoints (PCK) [8] metric with proximity radius $\alpha = 0.2$. The quantitative results are 31.0 PCK for speaker Ellen, 60.3 PCK for speaker Oliver, 40.6 PCK for speaker Shelly K. and 23.6 PCK for speaker Jonathan G. We observe that Oliver and Shelly K. who are in sitting position and therefore show less upper body movement achieve a higher PCK. In contrast, the PCK is lower for speakers which are standing (Ellen) and walking around (Jonathan G.).

## 2.5 System Design

Our real-time system consists of two main components, the speaker-specific gesture prediction model and the gesture animator. We show an overview of the system in Figure 2. The gesture prediction model contains a trained UNet model with a speech encoder and a human pose decoder. The output of the predictor are gestures in the form human pose keypoints. The human pose keypoints are taken as an input by the other component, the gesture animator. The animator reads keypoints and transforms them into animations for a virtual character. To achieve that, the keypoint positions which are located on joints are transferred into skeletal bone rotations. Incomplete information about rotation angles are estimated by using inverse kinematic computations. The motion between keypoint positions is interpolated using linear interpolation.

We take speech sequences of two seconds as input for our real-time system. The gesture prediction model reads the two second

audio input to predict gestures. The two second input provides context for the model to prediction beat gestures. Our model needs about 70 milliseconds to generate the corresponding output gestures from two second input. In our experiments we used a mid-range desktop computer setup with an Intel 8th generation i7 processor and a Nvidia Geforce GTX 1070 GPU. The generated gestures are passed to the animator which reads a two seconds of human pose data at the rate 15 poses per second. Besides interpolation between poses from within the two second sequences, the animator also interpolates between the last pose of the previous sequence and the first pose of the current sequence. The interpolation between sequences is computed more smoothly because the pose deviation between different sequences is larger. The total delay of our system is about 2.2 seconds which results from the two second sequence length, the prediction time, and the animation interpolation.

### 2.6 Future Work

Our gesture synthesizer is able to generate gestures for a given speaker with perceptual quality that could not be determined to be significantly different than the ground truth. However, because the training depends on the sound wave of the speech, the synthesis is speaker-specific and cannot generate good quality gestures for an arbitrary speaker. To address a wider spectrum of the applications, the predictor should also be able to synthesize gesture animations given a speech stream of an arbitrary speaker. One way to address this issue is to automatically transcribe the speech from the video into text using speech recognition algorithms and preserving the resulting text temporal alignment with the motion data. The resulting *speech* vector in our $< speech, gesture >$ training pair will now consists of text, rather than an audio sequence. This approach would significantly reduce the dimensionality of the input vector thus reducing the optimal size of the training data required.

## 3 CONCLUSION

We proposed a method of synthesizing gestures on avatars in real-time given input speech. Our model that translates speech to gestures is speaker-specific and it learns by observing speakers utilizing the large amount of video data available. We achieve a delay of less than three seconds when applying the gesture prediction in a real-time communication setting. This delay mainly comes from the fact that our model needs temporal context and is not able to predict gestures instantly. As a result, our system can be used most efficiently in one-way communication applications.

## 4 ACKNOWLEDGMENTS

This work was supported in part by the Marshall Plan Foundation and the National Science Foundation.

## REFERENCES

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

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

[3] 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*. 2672–2680.

[4] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Manuel Rebol, Christian Guetl, and Krzysztof Pietroszek. 2021. Passing a Non-verbal Turing Test:Evaluating Gesture Animations Generated from Speech. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 1–8.

[6] O. Ronneberger, P. Fischer, and T. Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

[7] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Yi Yang and Deva Ramanan. 2013. Articulated Human Detection with Flexible Mixtures of Parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2013), 2878–90.

[9] Christian Zimmermann and Thomas Brox. 2017. Learning to Estimate 3D Hand Pose from Single RGB Images. In *IEEE International Conference on Computer Vision (ICCV)*.