# On the Importance of Representations for Speech-Driven Gesture Generation

Extended Abstract

Taras Kucherenko
KTH Royal Institute of Technology
Stockholm, Sweden
tarask@kth.se

Dai Hasegawa
Hokkai Gakuen University
Sapporo, Japan
dhasegawa@hgu.jp

Naoshi Kaneko
Aoyama Gakuin University
Sagamihara, Japan
kaneko@it.aoyama.ac.jp

Gustav Eje Henter
KTH Royal Institute of Technology
Stockholm, Sweden
ghe@kth.se

Hedvig Kjellström
KTH Royal Institute of Technology
Stockholm, Sweden
hedvig@kth.se

## ABSTRACT

This paper presents a novel framework for automatic speech-driven gesture generation applicable to human-agent interaction, including both virtual agents and robots. Specifically, we extend recent deep-learning-based, data-driven methods for speech-driven gesture generation by incorporating representation learning. Our model takes speech features as input and produces gestures in the form of sequences of 3D joint coordinates representing motion as output. The results of objective and subjective evaluations confirm the benefits of the representation learning.

## KEYWORDS

Gesture generation; social robotics; representation learning; neural network; deep learning; virtual agents

**ACM Reference Format:**
Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. 2019. On the Importance of Representations for Speech-Driven Gesture Generation. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

## 1 INTRODUCTION

Conversational agents are rapidly becoming commonplace and many of us will soon interact with them regularly in our day-to-day lives. Humans use non-verbal behaviors to signal their intent, emotions, and attitudes in human-human interactions [7, 8]. Therefore, conversational agents also need the ability to perceive and produce non-verbal communication in a human-like way.

An important part of non-verbal communication is gesticulation: gestures communicate a large share of non-verbal content [9]. To achieve natural human-agent interaction, it is hence important to enable conversational agents to accompany their speech with gestures in the way people do.

Most existing work on generating hand gestures relies on rule-based methods [1, 6, 10]. These methods are rather rigid as they can only generate gestures that are incorporated in the rules. Consequently, it is difficult to fully capture the richness of human gesticulation in rule-based systems. In this paper, we follow a line of data-driven methods [2, 3, 5, 11], which learn to generate human gestures from a dataset of human actions. More specifically, we use speech data, as it is highly correlated with hand gestures [9] and has the same temporal character (speech and gestures are parallel, aligned time-sequences). These properties make speech-driven gesture generation an interesting direction to explore.

Our contribution is a novel speech-driven gesture generation method that extends the deep-learning-based method in [5] to incorporate representation learning. Numerical evaluation results show that our system learns a mapping from human speech signals to corresponding upper body motions better than a baseline model based on [5]. A subsequent user study indicates that representation learning indeed improved on the baseline model in terms of the perceived naturalness of generated gestures.

## 2 REPRESENTATION LEARNING FOR SPEECH-MOTION MAPPING

### 2.1 Problem formulation

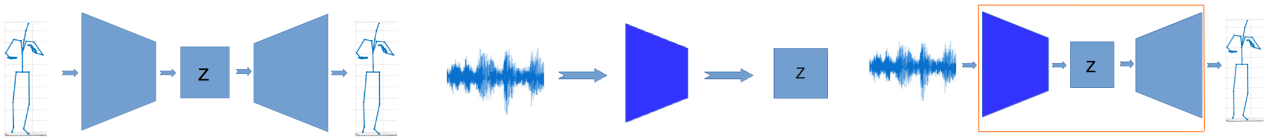
We frame the problem of speech-driven gesture generation as follows: given a sequence of speech features $\mathbf{s} = [\mathbf{s}_t]_{t=1:T}$—here Mel Frequency Cepstral Coefficients (MFCCs)—the task is to generate corresponding, natural-looking gestures $\hat{\mathbf{g}} = [\hat{\mathbf{g}}_t]_{t=1:T}$.

The ground truth gestures $\mathbf{g}_t$ and predicted gestures $\hat{\mathbf{g}}_t$ are typically represented as sequences of poses (3D joint coordinates): $\mathbf{g}_t = [x_{i,t}, y_{i,t}, z_{i,t}]_{i=1:n}$, where $n$ is the number of keypoints of the human body.

### 2.2 Baseline speech-to-motion mapping

Our model builds on the work of Hasegawa et al. [5], with their approach functioning as our baseline system. This baseline speech-gesture Deep Neural Network (DNN) [5] takes MFCC features of a speech recording as input and generates a sequence of gestures frame by frame. The network contains three fully connected layers followed by a recurrent network layer with so-called Gated Recurrent Units (GRUs) [4] and finally a linear output layer.

*Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.* 




(a) Step1: Motion representation learning (b) Step2: Mapping speech to motion repr. (c) Step3: Combining the learned components

**Figure 1: The proposed encoder-decoder DNN for speech-to-motion mapping.**

## 2.3 Proposed approach

Our intent with this paper is to extend the baseline model by leveraging the power of representation learning. Our proposed approach contains three steps:

(1) We use a Denoising Autoencoder (DAE) [13] to learn a compact representation **z** of the motion (see Fig. 1a).
(2) We learn a mapping from the chosen speech features **s** to the learned motion representation **z**, using the same network architecture as in the baseline model (see Fig. 1b).
(3) The two learned mappings are chained together to turn speech input **s** into motion output **g** (see Fig. 1c).

For implementation details we refer readers to our code on GitHub.[1]

# 3 EXPERIMENTS

## 3.1 Data and setup

For our experiments, we used a gesture-speech dataset collected by Takeuchi et al. [12]. The dataset contains 1,047 utterances, of which we used 904 for training, 53 for validation, 45 for development and 45 for testing. The models were thus learned from 171 minutes of training data at 20 frames per second, equating to 206,000 frames.

## 3.2 Objective comparison

We first evaluated how different dimensionalities for the learned motion representation affected the prediction accuracy of the full system. The evaluation of the motion jerkiness (see Fig. 2) demonstrates that our system produces smoother motion, since jerkiness decreased by two thirds.

## 3.3 User study

We conducted a 1×2 factorial-design user study with representation learning as the within-subjects factor (baseline vs. proposed). We used 10 utterances randomly selected from the 45 test utterances. Example videos from the study are provided at vimeo.com/album/5667276. Participants responded to statements (the same as in [5]) relating to the *naturalness*, *time consistency*, and the *semantic consistency* of the motion. Each aspect was assessed through agreement/disagreement (on a seven-point Likert scale) with three distinct statements. The order of the speech utterances was fixed for every participant, but the gesture conditions were counter-balanced.

19 native speakers of Japanese (17 male, 2 female) with an average age of 26 participated in our user study. Fig. 3 illustrates the results. A paired-sample $t$-test was conducted to evaluate the impact of the motion encoding on the rating of the produced gestures. We found a significant difference in naturalness (but not the other aspects) between the baseline (M=4.16, SD=0.93) and the proposed model (M=4.79, SD=0.89), with $p$<0.002.

[1]Our code can be found at github.com/GestureGeneration/Speech_driven_gesture_generation_with_autoencoder

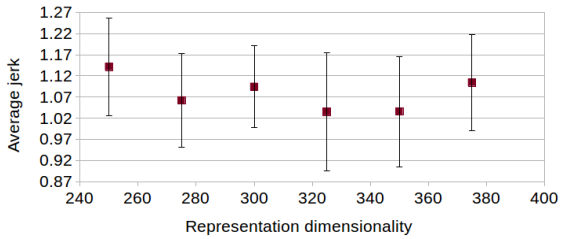



**Figure 2: Average jerk of our model. Our model is significantly closer to the ground truth jerk of 0.54 than the baseline, which at 2.8±0.3 is far off the top of the vertical scale.**

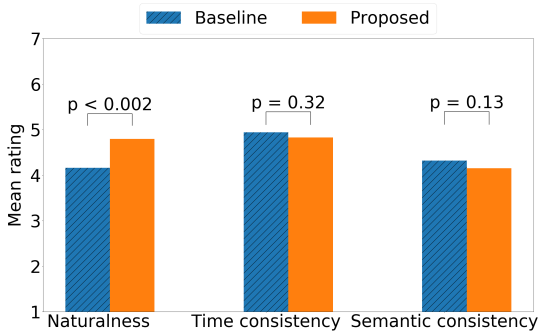



**Figure 3: User study results. Higher is better. We note a significant difference in naturalness, but not the other aspects.**

# 4 CONCLUSIONS

We have presented a new model for speech-driven gesture generation that extends prior work using deep learning for gesture generation by incorporating representation learning. The motion representation is learned first, after which a network is trained to predict such representations from speech, instead of directly mapping speech to raw joint coordinates as in prior work.

Our experiments show that representation learning improved the performance of the speech-to-gesture neural network both objectively and subjectively. In particular, the proposed method generated smoother motion and was rated as significantly more natural than the baseline method in a user study.

## REFERENCES


[1] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. Beat: the behavior expression animation toolkit. In *Annual Conference on Computer Graphics and Interactive Techniques*.

[2] Chung-Cheng Chiu and Stacy Marsella. 2011. How to train your avatar: A data driven approach to gesture generation. In *Proc. International Workshop on Intelligent Virtual Agents*. 127–140.

[3] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: a deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents*.

[4] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation* (2014), 103.

[5] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. ACM, 79–86.

[6] Chien-Ming Huang and Bilge Mutlu. 2012. Robot behavior toolkit: generating effective social behaviors for robots. In *ACM/IEEE International Conference on Human Robot Interaction*.

[7] Mark L. Knapp, Judith A. Hall, and Terrence G. Horgan. 2013. *Nonverbal Communication in Human Interaction*. Wadsworth, Cengage Learning.

[8] David Matsumoto, Mark G. Frank, and Hyi Sung Hwang. 2013. *Nonverbal Communication: Science and Applications*. Sage.

[9] David McNeill. 1992. *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago press.

[10] Victor Ng-Thow-Hing, Pengcheng Luo, and Sandra Okita. 2010. Synchronized gesture and speech production for humanoid robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

[11] Najmeh Sadoughi and Carlos Busso. 2017. Speech-driven animation with meaningful behaviors. *arXiv preprint arXiv:1708.01640* (2017).

[12] Kenta Takeuchi, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa, and Hiroshi Sakuta. 2017. Creating a gesture-speech dataset for speech-based automatic gesture generation. In *International Conference on Human-Computer Interaction*. Springer, 198–202.

[13] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* 11, Dec (2010), 3371–3408.