On the Importance of Representations for Speech-Driven Gesture Generation

Taras Kucherenko[1], Dai Hasegawa[2], Naoshi Kaneko[3],
Gustav Eje Henter[1], Hedvig Kjellström[1]

1 - KTH Royal Institute of Technology
Stockholm, Sweden

2 - Hokkai Gakuen University
Sapporo, Japan

3 - Aoyama Gakuin University
Sagamihara, Japan

Read the full paper

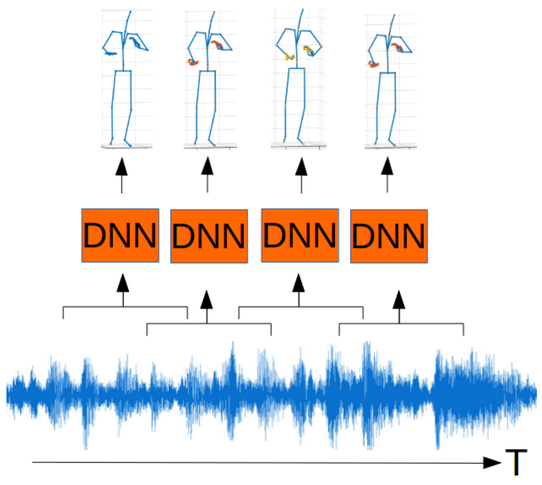
Deep-learning based speech-driven gesture generation becomes more natural using representation learning

MOTIVATION

Gestures transmit a large share of non-verbal content in communication. To achieve natural human-agent interaction it is important that conversational agents accompany their speech with gestures in the way people do.

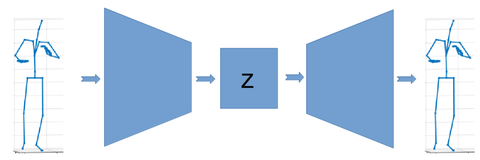
GENERAL FRAMEWORK

We follow a line of data-driven methods [2,3], which learn to generate human gestures from a dataset of human actions. We learn a mapping from the speech sequence to the 3D motion on a dataset of recorded motion sequences [1]. Our framework is illustrated below.

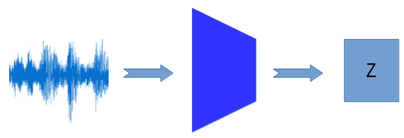


PROPOSED METHOD

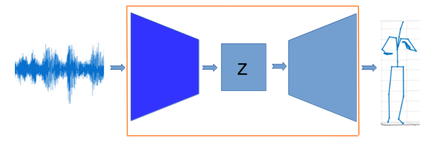
We extend recent deep-learning-based method [2] for speech-driven gesture generation by incorporating representation learning using the Denoising Autoencoder (DAE). Our system has three stages:



1. Apply representation learning to learn a motion representation by DAE



2. Learn a mapping from the speech signal to the learned motion representation by another NN

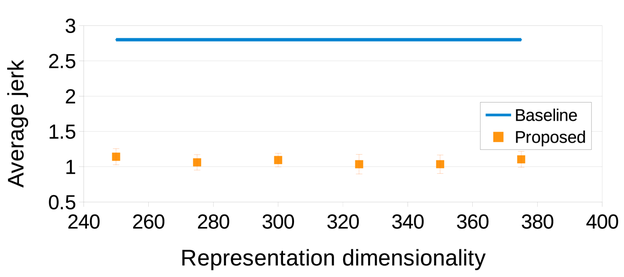


3. The two learned mappings are chained together to turn speech input into motion output

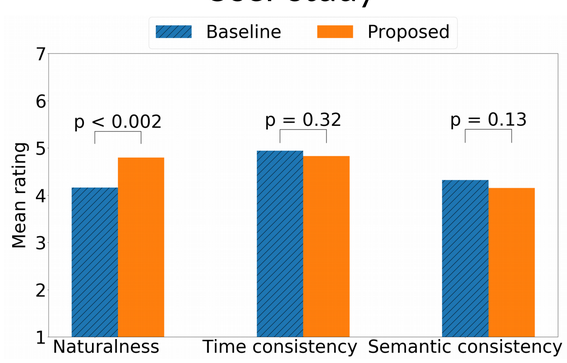
MAIN RESULTS

We evaluated different representation sizes and conducted a user study.

Evaluating jerkiness

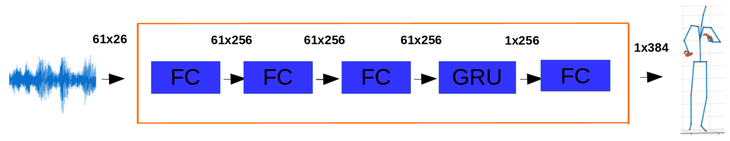


User study



BASELINE MODEL

Our baseline model and architecture in step 2 are closely based on the work of Hasegawa [2]. This model is illustrated below.



FC = Fully Connected layer, GRU = Gated Recurrent Units

DISCUSSION

Our experiments show that representation learning improves the performance of the speech-to-gesture neural network both objectively and subjectively.

Follow-up work:
Video (with a link to the paper and the code)
youtu.be/Iv7UBe92zrw

ACKNOWLEDGEMENT

The authors would like to thank Iolanda Leite, Simon Alexanderson and Sanne van Waveren for useful discussions and comments. This work is funded by Stiftelsen för Strategisk Forskning.

REFERENCES

[1] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. 2017. Creating a gesture-speech dataset for speech-based automatic gesture generation. International Conference on Human-Computer Interaction. Springer, Cham,
[2] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018 Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network. International Conference on Intelligent Virtual Agents. ACM.
[3] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predictingco-verbal gestures: a deep and temporal modeling approach. International Conference on Intelligent Virtual Agents. ACM.