A Framework for Integrating Gesture Generation Models into Interactive Conversational Agents

Demonstration Track

Rajmund Nagy*
KTH, Stockholm, Sweden

Taras Kucherenko*
KTH, Stockholm, Sweden

Birger Moell
KTH, Stockholm, Sweden

André Pereira
KTH, Stockholm, Sweden

Hedvig Kjellström
KTH, Stockholm, Sweden

Ulysses Bernardet
Aston University, Birmingham, UK

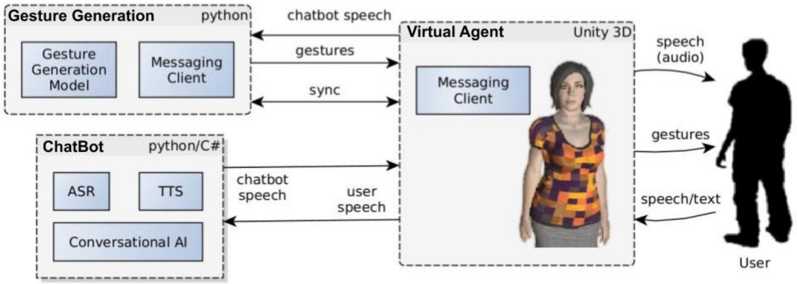

Figure 1: Architecture of the framework for integrating gesture generation models.

ABSTRACT

Embodied conversational agents (ECAs) benefit from non-verbal behavior for natural and efficient interaction with users. Gesticulation – hand and arm movements accompanying speech – is an essential part of non-verbal behavior. Gesture generation models have been developed for several decades: starting with rule-based and ending with mainly data-driven methods. To date, recent end-to-end gesture generation methods have not been evaluated in a real-time interaction with users. We present a proof-of-concept framework, which is intended to facilitate evaluation of modern gesture generation models in interaction.

We demonstrate an extensible open-source framework that contains three components: 1) a 3D interactive agent; 2) a chatbot backend; 3) a gesticulating system. Each component can be replaced, making the proposed framework applicable for investigating the effect of different gesturing models in real-time interactions with different communication modalities, chatbot backends, or different agent appearances. The code and video are available at the project page https://nagyrajmund.github.io/project/gesturebot.

KEYWORDS

conversational embodied agents; non-verbal behavior synthesis

ACM Reference Format:
Rajmund Nagy, Taras Kucherenko, Birger Moell, André Pereira, Hedvig Kjellström, and Ulysses Bernardet. 2021. A Framework for Integrating Gesture Generation Models into Interactive Conversational Agents: Demonstration Track. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 3 pages.

1 INTRODUCTION

Humans use non-verbal behavior to signal their intent, emotions and attitudes [9, 16]. Similarly, embodied conversational agents (ECAs) can be more engaging when having appropriate nonverbal behavior [21]. It is therefore desirable to enable conversational agents to communicate nonverbally.

Currently existing implementations of ECAs rely primarily on pre-recorded animations or require handcrafted specification of the motion [11, 15, 17], e.g. in XML-based formats [10]. However, recent developments in the field of gesture generation make it possible to produce realistic gestures in a purely data-driven fashion [1, 12, 23]. To date, many of these recent gesture generation methods have not been evaluated in a real-time interaction with users. A potential reason for the lack of evaluation in interaction for the recent models is the difficulty of setting up an interactive conversational agent.

In this work, we outline a framework for embedding data-driven gesture generation models into conversational agents. We envision that this framework with accelerate development of interactive embodied agents with end-to-end gesticulation models.

Our framework is modular, which enables it to be used for a wide range of scientific investigations about intelligent virtual agents, such as experimenting with their voices, gestures, breathing, conversational complexity or gender. For our demonstration, we integrate the speech- and text-driven model developed by Kucherenko et

*Both authors contributed equally to the paper.

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

al. [13] into an ECA built with Unity, and we show the flexibility of our approach by demonstrating our framework with two different chatbot backends.

2 SYSTEM DESCRIPTION

Our open source system is composed of a 3D virtual agent in Unity, a chatbot backend with text-to-speech capabilities and a neural network that generates gesturing motion from speech (Figure 1). The communication between the components consists of sending audio, text or motion file in a message; in our implementation, this is facilitated by the open-source Apache ActiveMQ message broker[1] and the STOMP protocol[2].

Each component is replaceable and is described in the corresponding section below.

2.1 Virtual agent in Unity

We provide the virtual environment and the user interface as a Unity scene. The end user interacts with the conversational agent through voice input or a text field.

By using Unity, we ensure that the system can be easily extended with new modules by other researchers in the future. Furthermore, it makes it possible to tailor the environment and the character model according to the requirements of the application (e.g., explore virtual/mixed reality applications).

2.2 Chatbot backend

The user's input message is sent to the chatbot backend that produces the agent's response as text and audio. The chatbot backend is comprised of a speech recognition system, a neural conversational model and a text-to-speech synthesizer. For our demonstration, we present two implementations of this component.

In the first configuration, Google's popular DialogFlow platform [19] is used with its automatic speech recognition module to enable voice-based interaction with the agent. The interfacing to DialogFlow is implemented in a C# module inside Unity.

In the second configuration, we adapt the open-domain chatbot BlenderBot [18] and a text-to-speech model called Glow-TTS [8] to build a virtual agent with free-form conversation capabilities. We leverage open-source implementations (provided by HuggingFace [22] and Mozilla TTS[3]) to seamlessly integrate the two models into the Python backend.

2.3 Gesture generation model

Based on the output of the chatbot backend, the gesture generation model synthesizes the corresponding motion sequence. We adapt a recent gesture generation model called Gesticulator [13], which is an autoregressive neural network that takes acoustic features combined with semantic information as its input, and generates the corresponding gesticulation as a sequence of upper-body joint angles, which is a widely used representation in computer animation and robotics.

In the original paper [13], the network was trained on the Trinity Speech-Gesture dataset [6], consisting of 244 minutes of speech and motion capture recordings of spontaneous monologues acted out by a male actor. The input features – log-power spectrograms for audio and BERT [4] word embeddings for text – are extracted and concatenated at the frame level, and a 1.5 s (30 frames at 20 FPS) sliding window of input features is used for predicting every motion frame, motivated by gesture-speech alignment research.

We tailor the base Gesticulator model to our interactive agent with the following adjustments:

1. We replace the audio features from spectrograms to the extended Geneva Minimalistic Acoustic Parameter Set [5], normalized to zero mean and unit variance. We qualitatively found that it results in better motion for synthesized voice.
2. The text transcriptions that are used for training the model contain precise word timing information, which is usually not available in real-time settings. Therefore, when interacting with a user, we approximate it with speech utterance lengths that are proportional to the syllable count.
3. Finally, we replace the BERT word embedding with FastText [3] (which has significantly lower dimensionality) in order to reduce the feature extraction time.

3 LIMITATIONS

At the current stage of development, each of the components has some important limitations:

1. Both available chatbot backends introduce several seconds of processing time before the agent responds to the user, which might currently affect immersion in the interaction.
2. The synthesized voice of the agent yields out-of-distribution audio samples which significantly degrade the quality of the generated motion. This could be improved by replacing audios in the dataset with synthetic audios [20] or by training a TTS model on the audio from a speech-gesture dataset [2].
3. Finger motion is not modelled by the gesture generation model due to poor data quality in the dataset.

However, the system's modular design of replaceable components allows addressing these limitations in the future. For instance, it is straightforward to replace Gesticulator with a model that generates full-body motion or to change the chatbot backend as shown in our two distinct examples with DialogFlow and Blenderbot.

4 CONCLUSIONS AND FUTURE WORK

We have presented a framework for integrating state-of-the-art data-driven gesture generation models with embodied conversational agents. As highlighted by the GENEA gesture generation challenge [14], the gesture generation field needs a reliable benchmark. The proposed framework provides such a possibility; it can be used to compare gesture generation models in real-time interactions.

There are many directions in which this work can be extended in the future. More diverse gesticulation can be achieved by choosing a probabilistic gesture generation model instead of a deterministic one like Gesticulator. Incorporating stylistic control [1, 7] in ECA to allow expression of different emotions is a promising direction. Moreover, the proposed framework can be used in user studies to investigate the human perception of different gesticulation.

[1] https://activemq.apache.org/
[2] https://stomp.github.io/
[3] https://github.com/mozilla/TTS

REFERENCES

[1] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum* 39, 2 (2020), 487–496.

[2] Simon Alexanderson, Éva Székely, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Generating coherent spontaneous speech and gesture from text. In *20th ACM International Conference on Intelligent Virtual Agents*. 1–3.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[4] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019)*.

[5] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[6] Ylva Ferstl and Rachel McDonnell. 2018. Investigating the Use of Recurrent Motion Modelling for Speech Gesture Generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA '18)*. Association for Computing Machinery, New York, NY, USA.

[7] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2020. Understanding the predictability of gesture parameters from speech and their perceptual importance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[8] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*.

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

[10] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*. Springer.

[11] Stefan Kopp and Ipke Wachsmuth. 2004. Synthesizing multimodal utterances for conversational agents. *Computer animation and virtual worlds* 15, 1 (2004).

[12] Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. 2021. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *International Journal of Human–Computer Interaction* (2021). https://doi.org/10.1080/10447318.2021.1883883

[13] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction.*

[14] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020. In *Proceedings of the International Conference on Intelligent User Interfaces*.

[15] Margot Lhommet, Yuyu Xu, and Stacy Marsella. 2015. Cerebella: automatic generation of nonverbal behavior for virtual humans. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[16] Michael Nixon, Steve DiPaola, and Ulysses Bernardet. 2018. An Eye Gaze Model for Controlling the Display of Social Status in Believable Virtual Humans. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 1–8. https://doi.org/10.1109/CIG.2018.8490373

[17] Igor Rodriguez, Aitzol Astigarraga, Txelo Ruiz, and Elena Lazkano. 2016. Singing minstrel robots, a means for improving social behaviors. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2902–2907.

[18] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

[19] Navin Sabharwal and Amit Agrawal. 2020. Introduction to Google Dialogflow. In *Cognitive Virtual Assistants Using Google Dialogflow*. Springer, 13–54.

[20] Najmeh Sadoughi and Carlos Busso. 2016. Head Motion Generation with Synthetic Speech: A Data Driven Approach.. In *INTERSPEECH*. 52–56.

[21] Maha Salem, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2011. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *Proceedings of the International Symposium on Robot and Human Interactive Communication*. IEEE.

[22] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[23] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4303–4309.