Data Driven Non-Verbal Behavior Generation for Humanoid Robots

Taras Kucherenko
Department of Robotics, Perception and Learning
KTH Royal Institute of Technology
Sweden
tarask@kth.se

ABSTRACT

Social robots need non-verbal behavior to make an interaction pleasant and efficient. Most of the models for generating non-verbal behavior are rule-based and hence can produce a limited set of motions and are tuned to a particular scenario. In contrast, data-driven systems are flexible and easily adjustable. Hence we aim to learn a data-driven model for generating non-verbal behavior (in a form of a 3D motion sequence) for humanoid robots.
Our approach is based on a popular and powerful deep generative model: Variation Autoencoder (VAE). Input for our model will be multi-modal and we will iteratively increase its complexity: first, it will only use the speech signal, then also the text transcription and finally - the non-verbal behavior of the conversation partner. We will evaluate our system on the virtual avatars as well as on two humanoid robots with different embodiments: NAO and Furhat. Our model will be easily adapted to a novel domain: this can be done by providing application specific training data.

CCS CONCEPTS

• **Human-centered computing** → *Systems and tools for interaction design*; Empirical studies in HCI; • **Computing methodologies** → *Machine learning*; Artificial intelligence;

KEYWORDS

Non-verbal behavior, data driven systems, machine learning, deep learning, humanoid robot

ACM Reference Format:
Taras Kucherenko. 2018. Data Driven Non-Verbal Behavior Generation, for Humanoid Robots. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3242969.3264970

1 INTRODUCTION

Robots are moving out of the industrial environment and becoming an integral part of our life. This puts robots in the contact with human and raises the importance of human-robot interaction.

Non-verbal behavior is a crucial part of human-human communication [16, 20]. We convey plenty of information using non-verbal behavior, such as intent, emotional state, and attitude [17]. Non-verbal behavior is also often used to disambiguate a message [9]. In order for robots to have human-like communication capabilities, they need to be able to produce and perceive non-verbal communication in a manner resembling humans. This can enable pleasant, natural and efficient interactions between humans and robots.

Most of the existing work on modeling non-verbal behavior for social humanoid robots is based on rule-based methods [2, 4, 23]. While those systems often perform well, they require extensive encoding of expert knowledge. Apart from that, they suffer from limited flexibility and variability.

Data-driven systems in contrary do not require expert knowledge, because they are learned from data. They also allow for more variability, since a non-verbal behavior generated by a data-driven system is not limited to a set of rules. These properties make data-driven systems an attractive research direction in robotic non-verbal behavior modeling.

The goal of this work is to build a data-driven non-verbal behavior generation system using machine learning tools. Deep Neural Networks (DNN) have become the state-of-the-art tool across many domains of human data, such as speech recognition [11], computer vision [12], and machine translation [5]. These kinds of methodology have been also widely applied to human skeleton modeling: for motion prediction [19] and classification [6]. For this reason, we believe this is a promising approach to the problem we aim to address.

Non-verbal behavior of humans is correlated with both their own verbal behavior, as well as the behavior (verbal and non-verbal) of the conversational partner [9]. We want to learn this correlation in order to create a generative model of non-verbal behavior for a humanoid robot. This task is highly non-trivial, as all these factors interact in a complex manner. So we are going to do it step by step.

We will begin by learning a mapping from a human speech signal (mainly its energy and prosody) to the corresponding upper body motion. Then we will add the text transcription of the speech signal as additional input to the system. The mapping will first be used to generate plausible upper-body motion for a virtual human, which will then be re-targeted to a humanoid robot. Later we will incorporate the non-verbal behavior of the conversational partner as an additional input to the system. Finally, we will learn a similar mapping for facial expressions and execute it on the robot Furhat. As a result, our system will be capable of producing non-verbal behavior in two different modalities: upper body motions and facial expressions.

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3264970

In summary, we want to find answers to the following questions:

(1) Can upper body gestures, generated based only on the speech signal, appear natural to a human?
(2) How can we measure if the robot motion appears natural?
(3) Which architecture of DNN is the best for generating non-verbal behavior (given limited amount of training data)?
(4) How can the non-verbal behavior of a conversational partner be incorporated?

2 RELATED WORK

2.1 Rule-based systems

Generation of non-verbal behavior has been traditionally done by a rule-based system [7, 10, 13, 21]. The BEAT toolkit [7] can be used to generate synchronized non-verbal behaviors and synthesized speech of the given text input by a virtual character. This system uses linguistic and contextual information of the text to choose an appropriate gesture. The mapping is contained in a set of rules from the non-verbal conversational research. Ng-Thow-Hing et al. [21] proposed a model for synchronized gesture motion generation with an arbitrary text input. They could generate many different gesture types: emblems, deictics, metaphoric and iconic gestures and beats. Fernández-Baena et al. [10] produced gestures based on the speech signal. Their rule-based system was derived for a particular corpus and for the Spanish language only, so it cannot generalize to other situations, which is a typical problem of rule-based systems.

2.2 Data-driven systems

Recently, limitations of the rule-based system motivated rising interest in the data-driven systems. Liu et al. [18] used a data-driven approach for a social robot, serving as a shopkeeper in a fully autonomous way. However, it had no non-verbal behavior, only moving from A to B and speech production. Admoni and Scassellati [1] built a model for generating non-verbal behavior, where both gaze and gestures were discretized. For decision making, they used k Nearest Neighbor (kNN) with majority votes. This system was not evaluated on robots, but while predicting human behavior in their dataset.
Several authors have explored data-driven approaches for the task of **speech to motion mapping**. For example, Chiu et al. [8] predicted the discrete set of co-verbal gestures, using a machine learning tool based on DNN and Conditional Random Fields (CRF). They worked with a virtual character. Takeuchi et al. [24] used DNN to produce upper body gestures based on the speech signal. However, user studies could not confirm any improvement over the baseline, probably due to quiver in the generated motion. Sadoughi et al. [22] bridged data-driven and rule-based approaches. They used Probabilistic Graphical Models (PGM) with an additional hidden node, which contained contextual information. They evaluated only 3 hand gestures and 2 head motions.

The novelty of our approach, compared to earlier work, is that we aim to generate general, continuous smooth non-verbal behavior, not only a discrete set of motions.

3 PROBLEM FORMULATION

The main focus of our work is generating kinesics motion, mainly upper body gestures and facial expressions. In their influential paper [9], Ekman and Friesen identify five major categories of kinesics:

- *emblems*: which has a direct verbal analogy (such as the "ok" gesture). Emblems are usually used to convey a message, which could be otherwise expressed through words.
- *illustrators*: which create a visual image and support the spoken message (such as holding your hands apart to indicate size). These gestures are typically less conscious and intentional as emblematic gestures.
- *adaptors*: body adjustments and other movements made with little awareness (such as shifting body and/or feet position when seated). The interpretation of adaptors is very difficult, often speculative and uncertain [9], especially for a robot, which does not need to adjust its pose.
- *affect displays*: expressing emotion, primarily through facial expressions. Our emotions and hence affect displays depend on the context of the verbal message as well as on the emotional state of the conversational partner.
- *regulators*: body movements that control, adjust, and sustain the flow of a conversation (for example nods and gaze behaviors). Regulators are most affected by the non-verbal behavior of the conversational partner.

In summary, most of a human's non-verbal behavior is correlated with the verbal communication of the person speaking or/and with the behavior of the conversational partner. Learning this correlation is the core of our approach to the non-verbal behavior generation.

The application scenario would be a robot having a conversation with a human while producing verbal and non-verbal signals. The input to our generative model would be a text of the utterance to be produced, its speech signal, as well as the verbal and non-verbal behavior of the conversational partner. Output would be the speech utterance and the corresponding non-verbal behavior for the robot. For the humanoid robot NAO, which has a physical body, we will generate mostly emblems and illustrators. While for the robotic head Furhat we will mostly generate affect displays.

4 OUR APPROACH

We decompose the challenging task of conditioning a generative model (for non-verbal behavior) on the text of the utterance, its speech signal and the verbal and non-verbal behavior of the conversational partner into smaller steps, and will add one modality at a time:

(1) Speech only,
(2) Speech and text,
(3) Speech, text and non-verbal behavior of a dialogue partner.

4.1 General Framework

We will start by learning a mapping from a human speech signal to the corresponding upper body motion sequence of this human:

$$\mathbf{m} \sim F(\mathbf{s}) \tag{1}$$

where $\mathbf{s} = (s_1, s_2, ...s_t)$ is a sequence of the relevant features (for example F0 contour, energy of the speech signal, as well as their

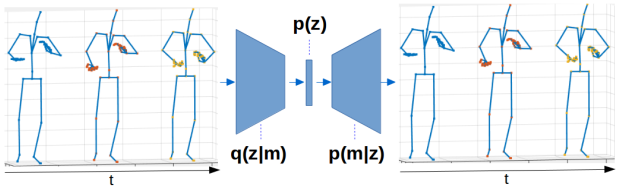

Figure 1: Illustration of the motion representation learning.

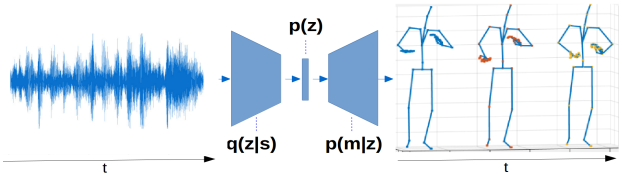

Figure 2: Illustration of the Speech-to-Motion mapping.

first derivatives) from the speech signal and $\mathbf{m} = (m_1, m_2, ...m_t)$ is a sequence of 3D positions of the joints of a human skeleton.

We express F as a random distribution conditioned on $\mathbf{s}$ from which $\mathbf{m}$ is sampled. We aim to learn this distribution from a dataset of recorded motion sequences [25]. A popular and powerful model for learning a probabilistic generative model is the Variational Autoencoder (VAE) [15].

4.2 Motion Encoding

In the first stage, we will learn a compact motion sequence representation using a VAE. A VAE consists of an encoding and a decoding network and a latent state variable (see Figure 1). In our model, the encoding network takes T time-frames of the motion as an input $\mathbf{m_t} = (m_{t+1}, m_{t+2}, ...m_{t+T})$ and produce the mean and variance for the Gaussian distribution for the lower-dimensional representation z , which is being decoded back to an original sequence.

The first model assumption is that the data $\mathbf{m}$ [1] is generated based on the latent variable z and has a Gaussian noise:

$$\mathbf{m} \sim \mathcal{N}(M_d(\mathbf{z}), \sigma_c^2 I) \tag{2}$$

where M_d is a DNN, σ_c is a constant variance.

The latent variable has the prior distribution:

$$z \sim \mathcal{N}(0, I) \tag{3}$$

The posterior over the latent variable $p(z|\mathbf{m})$ is intractable, so we approximate it with another Gaussian:

$$\mu(m), \sigma(m) = M_e(\mathbf{m}) \tag{4}$$

$$q(\mathbf{z}|\mathbf{m}) = \mathcal{N}(\mu(\mathbf{m}), \sigma(m)^2 I) \tag{5}$$

where M_e is another DNN.

The model is trained using variational inference on a set of recorded upper-body motions $\mathbf{m}_i$: both Neural Networks M_e and M_d are learned from data.

4.3 Speech To Motion Mapping

In order to obtain a probabilistic mapping from speech to motion, we propose to replace the encoder M_e with a mapping from speech to the latent state-space parameters μ and σ:

$$\mu(\mathbf{s}), \mu(\mathbf{s}) = S_e(\mathbf{s}) \tag{6}$$

where S_e is a DNN. The decoder M_d, learned as described in Section 4.2, is kept fixed during training of speech to motion mapping. The network is trained with a set of recorded upper-body motions $\mathbf{m}_i$ and the corresponding speech utterances $\mathbf{s}_i$. We optimize an Euclidean distances $\|\mu(\mathbf{s_i}) - \mu(\mathbf{m}_i)\|$ and $\|\sigma(\mathbf{s_i}) - \sigma(\mathbf{m}_i)\|$.

[1] We omit subscript and write m instead of m_t (to simplify the notations)

Testing During the testing the speech sequence is cut in the overlapping chunks and each of them is processed by the network. In order to achieve smoothness of the motion, we propose to change the computation of z to sample close from the previous sample. We do it by the following model assumptions:

$$\mathbf{z}_\tau \sim p(z_\tau | z_{\tau-1}, s_\tau) \tag{7}$$

$$p(z_\tau | z_{\tau-1}, s_t) \propto \mathcal{N}(z_t | \mu_\tau, \sigma_\tau^2 I) * \mathcal{N}(z_t | \mathbf{z}_{\tau-1}, \sigma_z^2 I) \tag{8}$$

where σ_z is a standard deviation in the latent space calculated over the training set. Additional smoothing might be applied to the resulting motion sequence.

5 RESEARCH PLAN

The research plan and tentative schedule are outlined below.

2018

(1) Learn a mapping from the speech sequence to the 3D motion sequence. Input will be a speech signal of a human. The output will be the 3D motion sequence that the human could have used while saying the speech from the input. This mapping will be a DNN based on VAE (as described in Section 4) trained on an existing dataset [25].
(2) Collect a dataset for further experiments: professional mime doing public speaking on similar topics with a diverse body language (we will record sound and 3D motion and will have the transcription given). We aim to record at least 4 hours.[2]
(3) Extend the previous system to incorporate the meaning (transcription) of the utterance so that the body language depends on what is being said.
(4) Execute mapping from human motion to robot motion on the NAO robot platform: from a predefined human skeleton sequence to a set of commands for NAO, using re-targeting and inverse kinematics.

2019

(1) Execute the mapping from speech and text to motion on the NAO robot: make NAO accompany a speech stream with upper body motion. The speech stream will be generated from a given text.
(2) Evaluate the system from the step above by a user study, where users will evaluate naturalness of NAO saying an utterance 1) without moving at all, 2) with rule-based behavior generation and 3) with data-driven behavior generation (our system).

[2] Related work [24] used 2 hours of speech

(3) Design and execute an interaction scenario, where a human is talking to a NAO, which uses the system above to generate its non-verbal behavior.
(4) Extend this system so that it takes into account the human motion as well as text and speech for producing a robot motion. So the VAE will get yet another input: the MoCap data of the conversational partner.

2020

(1) Learn a mapping from the speech and text signal to a facial expression sequence. This mapping should be learning from data collected in a robot-patient interaction scenario (see the EACare project [14]).
(2) Implement the mapping described in step (1) on the Furhat [3] platform.
(3) Integrate the system from step (2) into the EACare [14] demonstrator.
(4) Evaluate the system from step (3) by a user study.

6 CONTRIBUTIONS

We expect the following contributions from this project:

(1) A data-driven generative model of human upper body motion
(2) A data-driven generative model of human facial expressions
(3) A data-driven generative model for the non-verbal behavior of the humanoid robots

All these mappings will be produced taking into account:

- speech signal
- speech transcription
- non-verbal behavior of the conversational partner

Since all of these systems are data-driven they would be adaptable to a novel and/or specialized domain: by collecting the application specific training data. For example, a model of the non-verbal behavior of a doctor can be built using recordings of doctors interacting with patients. This will be done in order to adapt the model for the EACare project [14].

ACKNOWLEDGEMENTS

I would like to thank my PhD advisor, Hedvig Kjellström, for the support and guidance on this work. This PhD project is supported by Swedish Foundation for Strategic Research project EACare under Grant No.: RIT15-0107.

REFERENCES

[1] Henny Admoni and Brian Scassellati. 2014. Data-driven model of nonverbal behavior for socially assistive human-robot interactions. In *International Conference on Multimodal Interaction*.
[2] Henny Admoni, Thomas Weng, Bradley Hayes, and Brian Scassellati. 2016. Robot nonverbal behavior improves task performance in difficult collaborations. In *ACM/IEEE International Conference on Human Robot Interaction*.
[3] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive Behavioural Systems*. 114–130.
[4] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *ACM/IEEE International Conference on Human Robot Interaction*.
[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
[6] Judith Bütepage, Michael Black, Danica Kragic, and Hedvig Kjellström. 2017. Deep representation learning for human motion prediction and classification. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[7] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. Beat: the behavior expression animation toolkit. In *Annual Conference on Computer Graphics and Interactive Techniques*.
[8] 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*.
[9] Paul Ekman and Wallace V Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica* 1, 1 (1969), 49–98.
[10] Adso Fernández-Baena, Raúl Montaño, Marc Antonijoan, Arturo Roversi, David Miralles, and Francesc Alías. 2014. Gesture synthesis adapted to speech emphasis. *Speech Communication* 57 (2014), 331–350.
[11] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.
[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[13] 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*.
[14] Patrik Jonell, Joseph Mendelson, Thomas Storskog, Goran Hagman, Per Ostberg, Iolanda Leite, Taras Kucherenko, Olga Mikheeva, Ulrika Akenine, Vesna Jelic, et al. 2017. Machine Learning and Social Robotics for Detecting Early Signs of Dementia. *arXiv preprint arXiv:1709.01613* (2017).
[15] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114* (2013).
[16] Mark L Knapp, Judith A Hall, and Terrence G Horgan. 2013. *Nonverbal communication in human interaction*. Wadsworth, Cengage Learning.
[17] Robert M Krauss, Yihsiu Chen, and Purnima Chawla. 1996. Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In *Advances in Experimental Social Psychology*. Vol. 28. 389–450.
[18] Phoebe Liu, Dylan F Glas, Takayuki Kanda, and Hiroshi Ishiguro. 2016. Data-driven HRI: Learning social behaviors by example from human–human interaction. *IEEE Transactions on Robotics* 32, 4 (2016), 988–1008.
[19] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.
[20] David Matsumoto, Mark G Frank, and Hyi Sung Hwang. 2013. *Nonverbal communication: Science and applications: Science and applications*. Sage.
[21] 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*.
[22] Najmeh Sadoughi and Carlos Busso. 2017. Speech-driven animation with meaningful behaviors. *arXiv preprint arXiv:1708.01640* (2017).
[23] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics* 4, 2 (2012), 201–217.
[24] 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 *International Conference on Human Agent Interaction*.
[25] 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.