Learning Non-verbal Behavior for a Social Robot from YouTube Videos

Patrik Jonell
KTH*
Stockholm, Sweden
pjjonell@kth.se

Taras Kucherenko
KTH*
Stockholm, Sweden
tarask@kth.se

Erik Ekstedt
KTH*
Stockholm, Sweden
erikekst@kth.se

Jonas Beskow
KTH*
Stockholm, Sweden
beskow@kth.se

***Abstract*—Non-verbal behavior is crucial for positive perception of humanoid robots. If modeled well it can improve the interaction and leave the user with a positive experience, on the other hand, if it is modelled poorly it may impede the interaction and become a source of distraction. Most of the existing work on modeling non-verbal behavior show limited variability due to the fact that the models employed are deterministic and the generated motion can be perceived as repetitive and predictable. In this paper, we present a novel method for generation of a limited set of facial expressions and head movements, based on a probabilistic generative deep learning architecture called Glow. We have implemented a workflow which takes videos directly from YouTube, extracts relevant features, and trains a model that generates gestures that can be realized in a robot without any post processing. A user study was conducted and illustrated the importance of having any kind of non-verbal behavior while most differences between the ground truth, the proposed method, and a random control were not significant (however, the differences that were significant were in favor of the proposed method).**

***Index Terms*—Facial expressions, non-verbal behavior, generative models, neural network, head movement, social robotics**

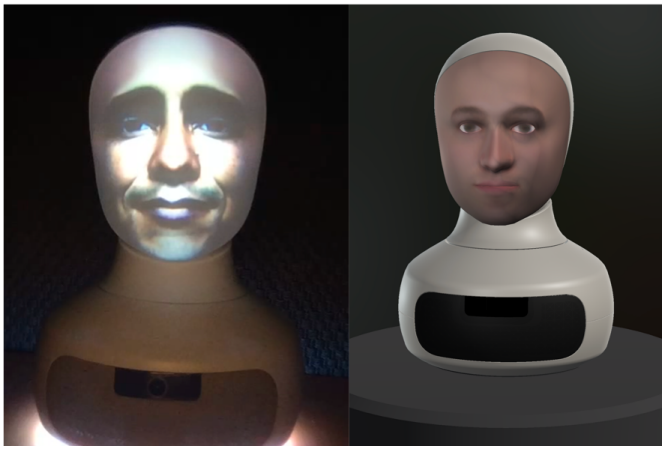

Fig. 1: To the left: a Furhat robot with former American president Barack Obama's face. To the right: A Furhat robot in the Fruhat simulator environment.

I. Introduction

Non-verbal cues play a crucial role in human communication, e.g. to convey information and express emotions. People often read and interpret these non-verbal cues from robots as they would from another person [1]. Hence, realistic and human-like non-verbal communication is crucial for social robots to achieve effective and enjoyable human-robot interactions [2].

Human non-verbal behaviors contain random variation, which lead to variable, alternating behaviors as a person speaks. For robots to be equipped with a similar level of variety in their non-verbal behaviors, it is possible to use probabilistic generative models trained on human examples. Unlike deterministic methods, these models can produce variations in the output for the same input. Video platforms, such as YouTube, are a perfect source of human-human interactions and can be used to extract data for training such models.

In this paper we present how we automatically collected video clips from YouTube and trained a generative model, using the Glow architecture [3], for generating head motion and facial expressions for a robot based on speech input. In our specific showcase we have concentrated on head movements based on videos from former American President Barack Obama's weekly addresses to the nation. We evaluated our method using Amazon Mechanical Turk.

*KTH Royal Institute of Technology

II. Related work

Non-verbal behavior generation has gained a lot of attention in the last few years [4]–[8]. We review the latest advances related to the two main components of our work: probabilistic models for motion generation, and facial expression and head motion generation.

A. Probabilistic models for motion generation

Habibie et al. presented a probabilistic motion generation model using Variational Autoencoders. They encode a sequence of control signals into a latent representation using a CNN-based encoder, then use an LSTM decoder to synthesize the motion with a low frame rate, and finally up-sample the motion to a higher frame rate using a neural network [4].

Vougioukas et al. used a probabilistic neural network, namely a Generative adversarial network (GAN), for speech-driven facial animation. The model takes speech segments and a still face as an input and produced a sequence of facial expressions as output [6].

Recently, a probabilistic generative deep learning architecture called Glow [3] was applied for generating locomotion

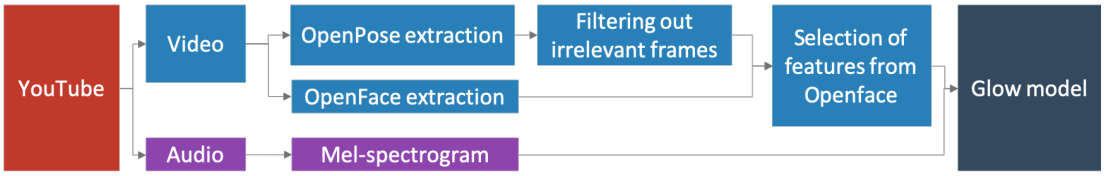

Fig. 2: Pre-processing steps.

from a provided trajectory [8]. Following the success of the Glow model architecture in locomotion generation, we apply it to facial expressions and head motion generation.

B. Facial expression and head motion generation

Several models have previously been successfully applied for head motion generation [5], [9] and facial expression generation [10], [11]. Karras et al. [10] trained a CNN-based neural network using speech and an emotion vector as input to generate facial behaviors for a virtual character with very little training data. Most of these models are however deterministic, and produce the same output for a given input, while we aim to build a probabilistic model instead. The most relevant work regarding head motion generation is the one of Greenwood et al. [9], where they predicted head pose from speech using a conditional variational autoencoder. This work is similar to ours, since it is also probabilistic. The main difference is that their model relies on approximate inference, while our model is exact and does not need any approximations.

III. Method

A. Model

The model was based on Glow, but modified to produce sequences of facial parameter vectors conditioned on audio features instead of unconditional RGB images. Our implementation used the PyTorch [12] GitHub repository github.com/chaiyujin/glow-pytorch as a base. The model was trained to generate a fixed-length output of 160 frames conditioned on the speaker's audio features (speech spectra). Like in [8] the conditioning information was concatenated with the other inputs to the network in the affine coupling layers. Our code is publicly available at github.com/jonepatr/glow-non-verbal-robot-behavior[1]. For more details on the Glow model and its application to motion generation please refer to [3], [8].

B. Data

A workflow was implemented to extract relevant features from both audio and video from specified YouTube IDs. In our experiments we used videos from the White House "Weekly address to the Nation" produced by the Obama administration. The videos were downloaded with the highest resolution available in 30fps as mp4 files. The audio was downloaded as raw wav files with a sample rate of 16kHz. The dataset consisted of approximately 25h of video material primarily of Barack Obama but occasionally featuring other people as well.

C. Pre-processing of data

The data processing is illustrated in Fig. 2.

The audio was normalized over each track and processed using the Librosa [13] Python package. The features consisted of 80-channel mel-spectrograms extracted using overlapping window frames of 2048 samples (0.128s), and with a hop length of 0.033s (in order to match the video frame rate).

Videos from YouTube may include scenes which are not of interest, e.g. segments showing multiple people or scenery, etc. The OpenPose [14] toolkit was used to classify frames of interest: frames containing only one person with a prediction certainty over 0.7 (max is 1) and where all facial landmarks were contained within the frame. After filtering out relevant frames, approximately 23h of video material remained. Once the relevant parts of the videos were defined we used the OpenFace [15] toolkit to extract the facial action units [16]. In our experiment we used head rotation (pitch, yaw and roll) and four action units corresponding to eye-brow movement (AU01, AU02, and AU04) and blinking (AU45), since these were considered accurate and reliable after manual inspection of the output from OpenFace.

D. Training Parameters

Glow hyper-parameters L=2, K=32 were used. A fixed width of 160 frames was used for both audio features and facial parameters. For the affine coupling layer we used two convolutional layers of size 512 with ReLU activation functions in between. The Adam optimizer [17] was used with learning rate 0.001 and the Noam learning rate decay scheme.

IV. Evaluation

We evaluated the model by playing back the generated motion using a robot. We used the *Furhat* robot from Furhat Robotics[2] since it provides detailed control of both facial expression and head movements. Two experiments were set up on Amazon Mechanical Turk for evaluating the robots/virtual agent's non-verbal behavior. Motion generated using the proposed method (using only audio as input), was tested against the ground truth (motion and audio are from the same sequence), a misaligned sequence (motion and audio are from different sequences) referred to as "Random Alignment" and a model with no head movement referred to as "Still".

In the first experiment we used videos from a physical Furhat robot replaying the facial behaviors. Audio files of Barack Obama's speech were used and the default lipsync

[1]GitHub commit 7753fa1b41bd4691baedc5c225ffe36010f50652

[2]https://www.furhatrobotics.com

from the Furhat Robot Development Kit (RDK) was used. Since the robot applied smoothing on the signal, a second experiment was also conducted but using the simulator included in the RDK which did not smooth out the signal.

A. Evaluators

All recruited participants were required to have an acceptance rate of 97% or more.

1) Study 1: We recruited 45 participants who had completed 5000 previous tasks. Only evaluators classified as masters[3] by Amazon were used for the study. Out of these, 17 were discarded since they did not pass the control questions.

2) Study 2: We recruited 50 participants who had completed 10,000 previous tasks. Out of these, 24 were discarded since they did not pass the control questions.

B. Stimuli

Ten stimuli for each condition were generated, which resulted in a total of 40 stimuli. The generated motion was 5.3s long (160 frames at 30fps) and the generated videos were approximately 6s long. Additionally stimuli with intentional distortion were added to control for "cheating" evaluators who submit completely random answers.

1) Study 1: The gestures were realized in a Furhat robot which was recorded performing the gestures using a web camera as shown to the left in Fig 1. The physical limitations imposed by the servo motors in the robot introduced a noticeable amount of smoothing of the head movements. Five control stimuli were created where the audio was removed and video was distorted.

2) Study 2: For the second evaluation the Furhat simulator included in the RDK was used as shown to the right in Fig 1. Three control stimuli each were generated containing intentional distortion of audio and video. For the video the same method was used as described in Study 1. The reason to separate the audio and video distortion was to catch participants that had the audio turned off.

C. Experiment setup

The videos were shown one by one, in random order, provided with three questions; "How coherent is the facial behavior and head movements with the voice?" (rated on a scale from 1 meaning "very incoherent" to 5 meaning "very coherent"), "How appropriate is the facial behavior for a (social) robot?" (rated on a scale from 1 to 5 where 1 was labelled "not at all" and 5 "extremely well") and finally whether or not there were any issues with the video.

D. Results

R [18] and lmerTest [19] were used to perform a linear mixed effects analysis of the two metrics. In total four models were fitted with the condition as a fixed effect and the evaluator as a random effect. Fitting the models with random slope and intercept for evaluator resulted in singular models, which were not used. A detailed result summary of the linear mixed effect model can be seen in Table I. All responses from "cheating" evaluators (43%) and all of the samples which were marked as having an issue (5%) were excluded from the analysis. In the table, "Coherent" refers to the question "How coherent is the facial behavior and head movements with the voice?" while "Appropriate" refers to "How appropriate is the facial behavior for a (social) robot?". The results for Study 1 and Study 2 are shown in Figure 3 and Figure 4 respectively.

[3]www.mturk.com/worker/help#what_is_master_worker

Results for Study 1

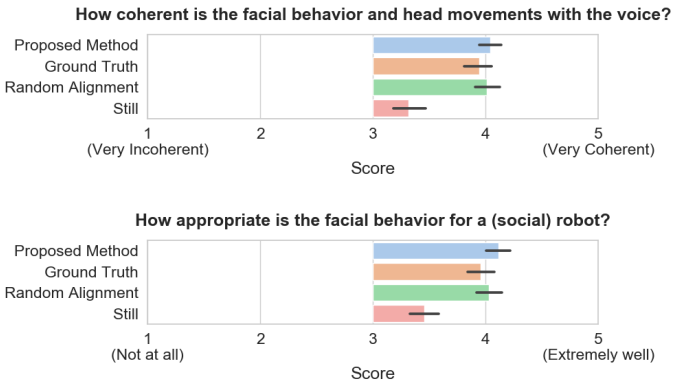

Fig. 3: *Study 1*. The mean score and 95% confidence interval for the question which is stated above each plot.

Results for Study 2

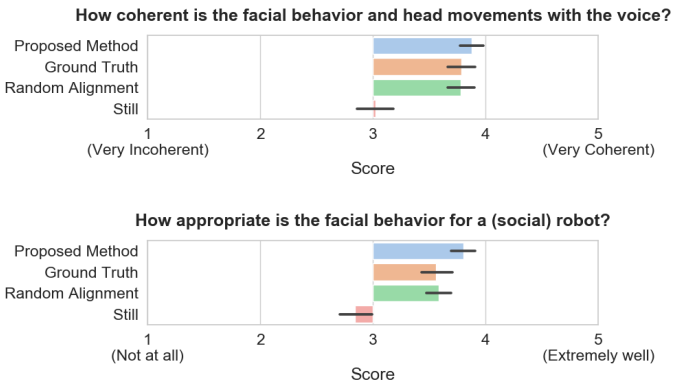

Fig. 4: *Study 2*. The mean score and 95% confidence interval for the question which is stated above each plot.

V. Discussion

There are three important findings in the user studies that will be discussed below.

1) Having no facial expressions or head movement is significantly worse than having some facial expression: This indicates that facial expressions are important for the perception of both the social robot and virtual agent, as it is clearly seen in both studies across both of the conditions.

TABLE I: Detailed summary of results from linear mixed effect analysis. The proposed method is used as a reference value, and the other conditions' estimates are relative to it

		Condition	Estimate	Std. Error	P-value
Study 1	**Coherent**	Proposed Method (Intercept)	4.04078	0.11215	<2e-16 ***
		Ground Truth	-0.10496	0.07637	0.170
		Random Alignment	-0.03183	0.07583	0.675
		Still	-0.72437	0.07584	<2e-16 ***
	Appropriate	Proposed Method (Intercept)	4.11309	0.11344	<2e-16 ***
		Ground Truth	-0.16658	0.07338	0.0234 *
		Random Alignment	-0.08606	0.07286	0.2378
		Still	-0.65525	0.07287	<2e-16 ***
Study 2	**Coherent**	Proposed Method (Intercept)	3.86822	0.12853	<2e-16 ***
		Ground Truth	-0.10376	0.07728	0.180
		Random Alignment	-0.08798	0.07614	0.248
		Still	-0.85576	0.07667	<2e-16 ***
	Appropriate	Proposed Method (Intercept)	3.79954	0.09645	<2e-16 ***
		Ground Truth	-0.24816	0.08591	0.00395 **
		Random Alignment	-0.21888	0.08465	0.00987 **
		Still	-0.96472	0.08525	<2e-16 ***

2) The proposed method is perceived as good as (for Study 1 and Study 2 in coherence) or even better (for Study 2 in appropriateness) compared to "Ground Truth" and "Random Alignment": One reason for that could be that our model exhibits more variability than the original motions, since it was trained on many different videos. Another reason could perhaps be that noise and artifacts in the output from the model, a result from not fully fitting the model to the data, are perceived as positive.

3) There was no significant difference between "Random Alignment" and "Ground Truth": This seems to suggest that evaluators were not influenced by the timing of head motion and facial expression. Note that lipsync was always correctly aligned with the audio. This might also suggest that there were flaws in the realization of the gestures in the robot and virtual agent. The robot did for example smooth out the motions, probably leading to reducing the differences between the various conditions.

VI. Conclusion

In this paper, we presented a method for the automatic generation of facial expressions and head movements in a social robot. We trained a probabilistic generative model, called Glow, on a dataset of videos of Barack Obama collected from YouTube. Two user studies were conducted, one with a social robot and one with a virtual agent. The user studies concluded that the lack of head motion and facial expressions was significantly worse than having any kind of motion. For the remaining three conditions, i.e. the proposed method, ground truth, and randomly picked gesture sequences from the original dataset, most of the differences were not significant, however, those that were significant were in favor of the proposed method. The code is publicly available at github.com/jonepatr/glow-non-verbal-robot-behavior.

VII. Acknowledgments

We are grateful to Gustav Eje Henter for his help, discussions and feedback regarding the Glow model. We would also like to thank Sanne van Waveren for her helpful feedback. Finally the authors would like to acknowledge the support from the Swedish Foundation for Strategic Research, project EACare [20] under Grant No.: RIT15-0107.

References

[1] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *International Conference on Intelligent Robots and Systems, (IROS '05)*. IEEE, 2005, pp. 708–713.

[2] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 313–323, 2013.

[3] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 215–10 224.

[4] I. Habibie, D. Holden, J. Schwarz, J. Yearsley, and T. Komura, "A recurrent variational autoencoder for human motion synthesis," *IEEE Computer Graphics and Applications*, vol. 37, p. 4, 2017.

[5] N. Sadoughi and C. Busso, "Novel realizations of speech-driven head movements with generative adversarial networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '18)*. IEEE, 2018, pp. 6169–6173.

[6] K. Vougioukas, S. Petridis, and M. Pantic, "End-to-end speech-driven facial animation with temporal gans," *arXiv preprint arXiv:1805.09313*, 2018.

[7] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *International Conference on Intelligent Virtual Agents (IVA '19)*. ACM, 2019.

[8] G. E. Henter, S. Alexanderson, and J. Beskow, "Moglow: Probabilistic and controllable motion synthesis using normalising flows," *arXiv preprint arXiv:1905.06598*, 2019.

[9] D. Greenwood, S. Laycock, and I. Matthews, "Predicting head pose from speech with a conditional variational autoencoder," in *Conference of the International Speech Communication Association (Interspeech '17)*. ISCA, 2017, pp. 3991–3995.

[10] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 94:1–94:12, Jul. 2017. [Online]. Available: http://doi.acm.org/10.1145/3072959.3073658

[11] H. Chu, D. Li, and S. Fidler, "A face-to-face neural conversation model," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)*, 2018.

[12] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[13] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 01 2015, pp. 18–24.

[14] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," in *arXiv preprint arXiv:1812.08008*, 2018.

[15] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018, pp. 59–66.

[16] P. Ekman and W. V. Friesen, "Facial action coding system: A technique for the measurement of facial movement," *Consulting Psychologists Press Palo Alto*, vol. 12, 01 1978.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR '15)*, 2015.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/

[19] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[20] P. Jonell, J. Mendelson, T. Storskog, G. Hagman, P. Ostberg, I. Leite, T. Kucherenko, O. Mikheeva, U. Akenine, V. Jelic *et al.*, "Machine learning and social robotics for detecting early signs of dementia," *arXiv preprint arXiv:1709.01613*, 2017.