Should Beat Gestures Be Learned Or Designed? A Benchmarking User Study

Pieter Wolfert
Ghent University - imec
Ghent, Belgium
pieter.wolfert@ugent.be

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

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

Tony Belpaeme
Ghent University - imec
Ghent, Belgium
tony.belpaeme@ugent.be

***Abstract*—In this paper, we present a user study on generated beat gestures for humanoid agents. It has been shown that Human-Robot Interaction can be improved by including communicative non-verbal behavior, such as arm gestures. Beat gestures are one of the four types of arm gestures, and are known to be used for emphasizing parts of speech. In our user study, we compare beat gestures learned from training data with hand-crafted beat gestures. The first kind of gestures are generated by a machine learning model trained on speech audio and human upper body poses. We compared this approach with three hand-coded beat gestures methods: designed beat gestures, timed beat gestures, and noisy gestures. Forty-one subjects participated in our user study, and a ranking was derived from paired comparisons using the Bradley Terry Luce model. We found that for beat gestures, the gestures from the machine learning model are preferred, followed by algorithmically generated gestures. This emphasizes the promise of machine learning for generating communicative actions.**

I. Introduction

Social robots are often depicted in popular media as robots that are already fully capable of interacting naturally with humans, on a verbal and nonverbal level. This, however, is not the case. In fact, most robots rely on manual or rule-based methods for gesture generation, which affects the duration of the interaction, and makes the interaction less appealing to the interacting human [1]. Up to 55% of human communication is non-verbal [2], and a large part of that is gesticulation, the use of gestures and body language to convey a message. Humans are also reading nonverbal behavior in robots in the same way as they do in other humans [3], [4]. From psycholinguistic research, we know that in human-human interaction, nonverbal behavior is very important. For example, Holler et al. [5], [6] showed that when humans questioned each other, the amount of time it took to answer a question was lower when the questioner used more speech gestures. Another example is that of Hömke, Holler, and Levinson [7]. In their study, eye blinking and head nodding in a virtual avatar were manipulated. They found that the answers of participants were shorter when the question raised by the avatar was followed with longer eye blinks and head nodding. These studies confirm the importance of nonverbal behavior in human communication.

An important aspect of non-verbal behavior is gesticulation, more precisely co-speech gestures. McNeill [8] categorized co-speech gestures into four categories: deictic, iconic, metaphoric, and beat gestures. Deictic gestures are better known as pointing gestures, whereas iconic gestures have a close relationship with the semantic context of speech. Metaphoric gestures are related to iconic gestures, but metaphoric gestures present a more abstract concept. Beat gestures are gestures that do not present meaning in some sort but are used to emphasize parts of the uttered speech. Recent work by Youngwoo et al. [9] showed that the relation between (written) speech and gesture use could be learned from videos. This was done by extracting human poses and subtitles from TEDx videos and learning this relation with a sequence to sequence neural network [10]. However, this approach lacks proper alignment in time between gesture and speech audio and is based on public speakers giving a well-prepared talk. Another problem is that this model is not able to learn very precise iconic or metaphoric gestures. Alignment of speech audio with gesticulation is very important, because these gestures are there to support verbal communication. Proper alignment can be reached when the relationship of gestures and speech is modeled based on audio features, of which the work by Kucherenko et al. [11] is an example. In this work, an encoder-decoder neural network is trained on gestures and speech of Japanese speakers. The model generates mostly beat gestures for all the speech features, indicating that it is not trivial for a neural network to capture more complex gestures.

In this paper we present the results of a user study where we compared generated beat gestures with gestures that were created manually, to find out whether gestures generated with machine learning are preferred over manually generated beat gestures. Currently, researchers in this field compare their outcome with their previous outcomes, but not with a well-established baseline. In this paper, we aim to compare beat gestures generated with machine learning, with three baseline conditions: designed beat gestures, timed beat gestures and noisy gestures (not specifically beat gestures). For the machine learning part, we used the model by Kucherenko [11], which is trained with new data from an English-speaking Irish actor [12], as this makes it possible to run a comparison study with English speakers. Our contribution is a step forward to benchmarking in the gesture generation field for Human-Robot Interaction (HRI).

II. Model Description

A. Problem Formulation

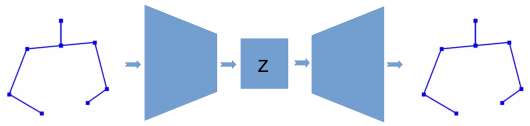
We consider the task of learning a mapping from a human speech signal to the corresponding upper body motion sequence: $\mathbf{m} = F(\mathbf{s})$, where $\mathbf{s} = (s_1, s_2, ...s_t)$ is a sequence of the prosodic features from the speech signal and $\mathbf{m} = (m_1, m_2, ...m_t)$ is a sequence of 3D positions of the joints of a human skeleton. We describe speech features and motion joints below.

1) Speech Features: We considered four prosodic features, extracted with a window length of 5.55 ms, resulting in 180 fps, which were subsequently sub-sampled by averaging to 60 fps. Those four features are: the energy of the speech signal, the logarithm of the F0 (pitch) contour and their numerical derivatives. The pitch and intensity value were extracted from audio using Praat [13] and normalized as in [14].

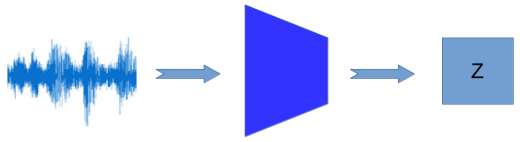
2) Human Skeleton: Since we are analyzing beat gestures, we consider only the upper body and ignore fingers, as they are not relevant for beat gestures. The resulting skeleton contains 8 joints: head, neck, left shoulder, right shoulder, left elbow, right elbow, left hand, and right hand.

B. Deep-Learning Based Solution

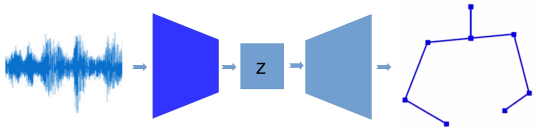
In our user study, we evaluate the state-of-the-art method for generating beat gestures based on speech [11].



(a) MotionED: Representation learning for the motion



(b) SpeechE: Mapping speech to motion representations



(c) Combining the learned components: SpeechE and MotionD

Fig. 1: How the encoder-decoder DNN for speech-to-motion mapping [11] is constructed. Every trapezium denotes a neural network, z denotes encoded representation of motion.

The machine learning model for speech-driven gesture generation used in this paper is depicted in Figure 1. First, a lower-dimensional representation of human motion is learned using a Denoising Autoencoder neural network. This neural network consists of a motion encoder *MotionE* and a motion decoder *MotionD*. Second, another neural network *SpeechE* is trained to map from speech to a corresponding motion representation. At test time, the speech encoder and the motion decoder networks are combined: *SpeechE* predicts motion representations based on a given speech signal and *MotionD* then decodes these representations to produce motion sequences. We refer the reader to the original paper by Kucherenko et al. [11] for more details on the network architecture. The code was taken from their GitHub[1] repository.

Since the human skeleton considered in our experiments is much simpler than the one in the original paper and the dataset is significantly smaller, the network was significantly reduced. For the Denoising Autoencoder (Figure 1a) the representation dimensionality was 20 instead of 325. The speech-to-representation neural network (Figure 1b) was also significantly simplified: the hidden layer size was reduced to 36; the amount of layers to 3; the batch size was reduced to 128; and the initial learning rate to 0.0005.

C. 3D Upper Body Modelling

For manual beat gesture generation, a 3D model of a human's upper body was modelled (with the joints as specified in section II-A2) using URDF (Unified Robot Description Format). URDF is a XML file format in which joints, dimensions and links are specified, and therefore this file can describe the kinematic information of the described agent. URDF can be used for running simulations with ROS (Robot Operating System) [15]. Having this URDF file, our 3D model of the human upper body, it becomes possible to calculate inverse and forward kinematics with the Python Module IKPY [2]. This serves as the basis for our modelled beat gestures.

III. Experimental Setup

A. Dataset

The model was trained on the Trinity College Conversational Dataset [12]. This dataset contains video recordings of one actor, who is allowed to speak freely about any topic he wishes. Together with the video and audio, the motion of the actor was captured using MoCap (Motion Capture) system. In total there are 23 takes of roughly 10 minutes. Due to issues with the synchronization of video and audio in parts of the dataset, only one recording, number 30, was used for training. Recording number 26 was used for validation, and recording 1 and 2 for testing. To accommodate for this, the neural network model was simplified and the output dimension was made smaller than in the original version.

B. Generated Beat Gestures

1) Machine Learning Generated (Condition 1): For the first condition we fed the trained model 10 seconds of audio, and concatenated and smoothed the resulting pose positions. To fit the generated poses in the same frame as the other pose positions in other conditions, we normalized and post-processed the resulting skeletons such that the location of the

[1] github.com/Svitozar/Speech_driven_gesture_generation_with_autoencoder
[2] github.com/Phylliade/ikpy

neck was at (0, 0, 0). The poses were also rotated to make the resulting skeleton facing front. [3]

2) Designed Beat Gestures (Condition 2): For our manual gesture conditions, we used a 3D model of the human upper body, as described in section II-C. The start position was with the hands in a resting position, where the hands are close to the hips. To generate beat gestures, we applied a vertical translation from the average resting position [8]. The trajectory is defined with a sine function on the y-axis. The amplitude of the sine function was alternated to generate different types of beat gestures. To make sure that the trajectories of the hands appeared natural, for every position in Cartesian space new joint positions were calculated through inverse kinematics, hence the need for a 3D model. To arrive at natural looking gestures, the x-values for the sine function were drawn from a logarithmic scale, from zero to pi. Beat gestures with different amplitudes were concatenated at random, and combined with audio. To smooth the concatenation of gestures, the input joint positions of a new gesture were the last known joint positions of the previous gesture.

3) Timed Beat Gestures (Condition 3): For the third condition, noisy gestures were sampled (generation of these gestures was similar to that of condition two, but with a very small amplitude). On top of these noisy gestures, a beat gesture was added roughly 400 milliseconds before a pitch in the audio was detected [16]. The onset of several pitches were taken, and the loudest pitch was taken as the pitch to input a beat gesture. Pitch detection and other on-the-fly audio processing was done using Librosa [17].

4) Noisy Gestures (Condition 4): Noisy gestures were generated like our designed beat gesture generation, but with a very small amplitude (to resemble noise on the endpoints). As these are context free, i.e. no speech input is used for the timing and they were not designed to resemble human-like beat gestures, the prediction is that this condition will be ranked lowest.

C. User Study

We used 10 audio samples of 10 seconds, from which we generated 40 videos of 10 seconds, which in turn translates to a video per condition, per audio file. To run pairwise comparisons, we needed in total six pairs per sample, which brought the total number of comparisons to sixty. For every pair, the user was asked to select the video which had the users' preference. A survey was set up using Google Forms, and the order of conditions was counterbalanced, to minimize the chance that two of the same conditions would succeed each other in the survey. The survey was promoted through Amazon Mechanical Turk. To control for the worker's focus, we added a check question, and logged the amount of time it took to complete the survey. Surveys completed in less than 10 minutes were not seen as serious submissions, and not taken into account for our analyses.

[3]C1: https://youtu.be/AJlc54yODPw, C2: https://youtu.be/I5c3FgWgdjY, C3: https://youtu.be/ONehBn8N9a8, C4: https://youtu.be/bXUS3SQBg9w

IV. Results and Discussion

A. User Study

We ran a user study where 41 participants were presented 59[4] video pairs. Of these 41 participants, the average age was 33 years (SD=9.5 years). Nineteen of them were male, twenty-two of them were female. 40 participants were native English speakers, 1 was not.

Since we used pairwise comparisons, we ran a Chi-Square Goodness of Fit test per pair (six in total). For this test, we assumed that if the conditions would be ranked equally, the distribution would be 50/50 per pair. For all six possible pairs, $p < 0.05$, and this assumption was therefore rejected.

A ranking was deduced using the Bradley-Terry Luce model (BTL) [18]. The BTL model provides a prediction p for the outcome of a paired comparison, where this prediction is in the form of the logarithm of the odds, $log(\frac{1}{1-p})$. Logarithm of the odds is a method to map $p[0, 1]$ to $[-\infty, +\infty]$, where a logit less than 0 equals $p < 0.5$. The results of applying this model to our data can be found in table I. Given the results a global ranking of our conditions from the preferences of the users in our user study can be derived, as visible in Figure 2.

TABLE I: Logit of Winning

	Wins			
Losses	Condition 1	Condition 2	Condition 3	Condition 4
Condition 1	-	0.94	1.81	2.23
Condition 2	-0.94	-	0.88	1.30
Condition 3	-1.81	-0.88	-	0.42
Condition 4	-2.23	-1.30	-0.42	-

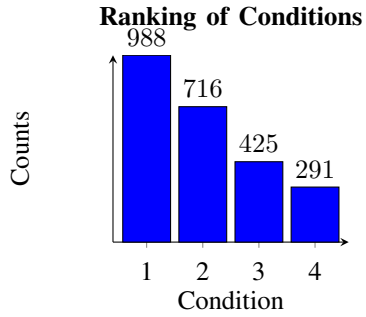


Fig. 2: Ranking based on number of wins (max is 1230).

B. Discussion

Model generated beat gestures are ranked highest among the four conditions, followed by designed beat gestures, timed beat gestures, and noisy gestures. The ranking of the other three conditions seems to indicate a preference for gestures at all, since the amount of beat gestures is highest in the designed condition, followed by the timed beat gestures and noisy gesture condition. We believe that even with a lack of coherent beat gestures, gesticulation of some sort is preferred over noise. Using 2D projections of human poses does not directly translate to robotic movement, and chances are that

[4]Due to an error on our side one pair was left out the survey.

we would get different results when we run this experiment in humanoid robots such as Pepper or NAO. For example, the generated poses from the machine learning model require a robot that is able to move fast with its joints. In combination with the motor noise these humanoid robots often generate, it is very likely that generated beat gestures learned from humans do not look as natural in humanoids as it does in humans.

V. Related Work

There is ample work on gesture generation, and this can be categorized in either rule-based or data-driven methods. We are primarily interested in how these systems are evaluated. For example, Levine, Theobalt and Koltun modelled prosody and motion using Hidden Markov Models, and evaluated this on virtual avatars while comparing between random synthesis, original motion and generated motion [19]. In their user study they focused on whether movements were timed appropriately and if motions were consistent with speech. Other work by Ng-Thow-Hing, Luo and Okita [20] is rule based, and in their evaluation studies they focused on the expressivity and timing of their generated gestures. Chiu and Marsella, who used a data-driven approach to gesture generation, focused in their evaluation studies on the quality of the generated gestures by comparing them with the original motion capture data [14]. A hybrid approach was taken by Sadoughi and Bosso, where an evaluation study was ran with a previous self-established baseline. This approach takes into account both prosody and semantics, making it possible to generate more meaning-full gestures. Work by Salem, Kopp, Wachsmuth, Rohlfing and Joublin shows that non-verbal behaviors and arm gestures displayed during speech acts make for a higher rating of a robot, even when there is no semantic match between gesture and speech [21]. In their user study the comparison was made between either using deictic gestures in a multi-modal setting or just using speech for instructions. One of the findings was that even with using non semantic matching gestures, the rating of the robot was still better than in the no-gesture condition, which is confirmed by our user study as well. However, we want to highlight that there is a lack of comparisons of the results between different studies, and that the settings in which the gestures are coupled with speech differ per study, which makes it hard to identify the best system and draw good comparisons.

VI. Conclusions

In this paper, we have presented a user study on generated beat gestures. Four types of beat gestures were compared in a pairwise comparisons. We found that beat gestures generated with a machine learning model scored the best among forty-one participants. Although this is just a first step in the direction of benchmarking different types of gesture generation models, it indicates that data-driven approaches toward gesture generation are fruitful and should be explored further. This is in accordance with earlier studies by Kipp, Neff, Kipp and Albrecht [22]. For future work we therefore aim to evaluate gesture generation models in social robots while taking the interaction partner into account. The latter is not only expected to lead to more naturalistic gesture generation, but is a prerequisite for dyadic communication.

References

[1] C.-M. Huang and B. Mutlu, "Robot behavior toolkit: generating effective social behaviors for robots," in *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 25–32, IEEE, 2012.

[2] A. Mehrabian, *Nonverbal communication*. Routledge, 2017.

[3] 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 *2005 IEEE/RSJ international conference on intelligent robots and systems*, pp. 708–713, IEEE, 2005.

[4] C.-M. Huang and B. Mutlu, "Modeling and evaluating narrative gestures for humanlike robots.," in *Robotics: Science and Systems*, pp. 57–64, 2013.

[5] J. Holler, K. H. Kendrick, and S. C. Levinson, "Processing language in face-to-face conversation: Questions with gestures get faster responses," *Psychonomic bulletin & review*, vol. 25, no. 5, pp. 1900–1908, 2018.

[6] J. Holler and K. Wilkin, "An experimental investigation of how addressee feedback affects co-speech gestures accompanying speakers' responses," *Journal of Pragmatics*, vol. 43, no. 14, pp. 3522–3536, 2011.

[7] P. Hömke, J. Holler, and S. C. Levinson, "Eye blinks are perceived as communicative signals in human face-to-face interaction," *PloS one*, vol. 13, no. 12, p. e0208030, 2018.

[8] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[9] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, and G. Lee, "Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots," in *International Conference on Robotics and Automation (ICRA '19)*, IEEE, 2019.

[10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[11] 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.

[12] Y. Ferstl and R. McDonnell, "Investigating the use of recurrent motion modelling for speech gesture generation," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp. 93–98, ACM, 2018.

[13] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2002.

[14] C.-C. Chiu and S. Marsella, "How to train your avatar: A data driven approach to gesture generation," in *International Workshop on Intelligent Virtual Agents (IVA'11)*, pp. 127–140, Springer, 2011.

[15] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, p. 5, Kobe, Japan, 2009.

[16] A. Kendon, "Gesticulation and speech: Two aspects of the," *The relationship of verbal and nonverbal communication*, no. 25, p. 207, 1980.

[17] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, pp. 18–25, 2015.

[18] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[19] S. Levine, C. Theobalt, and V. Koltun, "Real-time prosody-driven synthesis of body language," in *ACM Transactions on Graphics (TOG)*, vol. 28, p. 172, ACM, 2009.

[20] V. Ng-Thow-Hing, P. Luo, and S. Okita, "Synchronized gesture and speech production for humanoid robots," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4617–4624, IEEE, 2010.

[21] M. Salem, S. Kopp, I. Wachsmuth, K. Rohlfing, and F. Joublin, "Generation and evaluation of communicative robot gesture," *International Journal of Social Robotics*, vol. 4, no. 2, pp. 201–217, 2012.

[22] M. Kipp, M. Neff, K. H. Kipp, and I. Albrecht, "Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis," in *International Workshop on Intelligent Virtual Agents*, pp. 15–28, Springer, 2007.