A GAN-based Approach for Generating Culture-Aware Co-Speech Gestures

Ariel Gjaci[1], Carmine Tommaso Recchiuto[1] and Antonio Sgorbissa[1]

[1]*Laboratorium - DIBRIS, Università di Genova, via all'Opera Pia 13, 16145, Genova, Italy*

Abstract

Embedding social robots with the capability of accompanying their sentences with natural gestures may be the key to increasing their acceptability and their usage in real contexts. However, it could be argued that the definition of *natural* communicative gestures is not trivial, since it strictly depends on the culture of the person interacting with the robot. The proposed work investigates the usage of Generative Adversarial Networks (GANs) for generating culture-dependent communicative gestures based on speech audio features. To this aim, a custom dataset, only composed of persons belonging to the same culture, has been created, to extract all keypoints and audio features needed to train the network. Then, a generative model, also consisting of a voice conversion module, has been implemented and tested with the humanoid robot Pepper. Preliminary results, obtained through objective measurements and subjective evaluation, show that the proposed approach may be promising for generating culture-dependent communicative gestures for social robots

Keywords

Social Robots, Communicative Gestures, Generative Adversarial Networks

1. Introduction

Humans do not interact with others by only relying on their speech capabilities. Indeed, we unconsciously accompany what we say with non-verbal movements, which are usually called *co-speech* or *non-verbal* gestures. The importance of this kind of non-verbal communication has been confirmed by different studies [1],[2], which have analyzed how conversational hand gestures may convey semantic information, and how gestures are involved in the conceptual planning of messages. The influence of culture in the generation of speech-accompanying gestures has been the subject of many research works. In [3], starting from the assumption that "communicative gestures are not a universal language", the author suggests that, notwithstanding a certain homologation process due to Western cultural hegemony, gestures are still culturally deep, and hardly accessible to imports from other cultures.

Given the importance of communicative gestures in human-human interaction, it is natural to ask oneself if co-speech gestures may play the same important role also in human-robot

The 8th Italian Workshop on Artificial Intelligence and Robotics – AIRO 2021

✉ argjaci.95@gmail.com (A. Gjaci); carmine.recchiuto@dibris.unige.it (C. T. Recchiuto); antonio.sgorbissa@unige.it (A. Sgorbissa)
🌐 https://www.researchgate.net/profile/Carmine-Recchiuto (C. T. Recchiuto); https://www.researchgate.net/profile/Antonio-Sgorbissa (A. Sgorbissa)
iD 0000-0001-9550-3740 (C. T. Recchiuto); 0000-0001-7789-4311 (A. Sgorbissa)

interaction. Answers to this question have been given in [4]. In this work, a monologue has been performed by the humanoid robot BERTI, showing how people paid more attention to the robot when it performed co-verbal gestures. Moreover, in the same scenario, a robotic speech accompanied by gestures has been proven to be better recalled than when gestures are absent. Similar conclusions have been drawn by [5]: additionally, the authors underline here the importance of communicative gestures for a social robot, to improve the engagement and the relational bondings that can arise between humans and robots.

On the basis of this analysis, this work describes a novel approach based on Generative Adversarial Networks (GANs) for embedding humanoid robots with the capabilities of using co-speech gestures during their verbal interaction with users. Given the importance of culture in this context [6, 7, 8, 9, 10, 11, 12], the idea underlying the proposed approach is to autonomously learn communicative gestures for a social robot by relying on a dataset composed by considered persons belonging to the same culture, thus creating a mapping between common and uncommon poses, and their audio features. The assumption is that the synchronization between gestures and speech is of the utmost importance in this context.

2. Methodology and Implementation

The proposed approach aims at developing a model able to learn the underlying mapping between audio features and gestures of a group of persons identifying themselves with the same culture. The model shall rely on a dataset composed of different speakers of the same cultures, but with different physical and audio properties. Moreover, it shall be capable, at the same time, of keeping the synchronization between words and gestures as well as word articulation and their affective content. Finally, the model shall be implementable for the automatic generation of gestures during verbal interaction with a social robot.

Concerning the proposed methodology, a scheme representing the translational model, used for training the system, is shown in Figure 1, while Figure 2 shows the architecture used for generating the co-speech gestures of a humanoid robot in run-time. The dataset, including video and audio is first pre-processed for keypoints and audio extraction through software such as OpenPose and many-to-one audio conversion approaches. Then, audio features are fed to the Generator of a GAN to produce a sequence of "faked" keypoints, that are fed to the Discriminator. The Generator and the Discriminator iteratively improve to, respectively, produce fake gestures and to distinguish them from the real ones.

The run-time generation process of communicative gestures (Figure 2) starts by converting the sentence that should be pronounced by the robot to speech, through Text-To-Speech (TTS) functionalities: the resulting audio features are then further converted and a set of two-dimensional corresponding keypoints vectors are generated. Two additional steps are needed, to the aim of performing the resulting motion with a humanoid robot: keypoints need to be translated to the three-dimensional space, and finally mapped to the robot.

A custom dataset has been built for developing and testing the system. In the current implementation, the Indian culture and the English language have been chosen as references for the work. Concerning the language, English was an obligatory choice, as we need a language known to researchers, widespread in the world, and allowing us to easily find subjects for

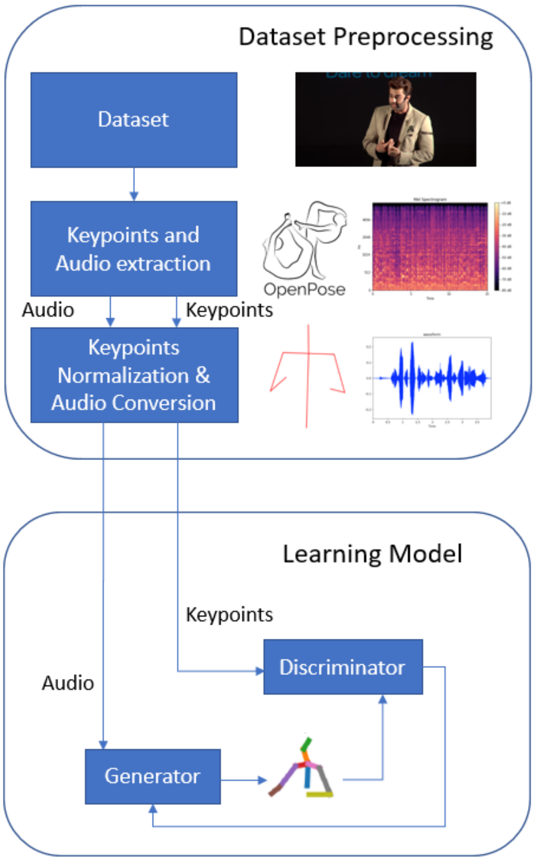

Figure 1: Training model of the proposed approach

the experimental evaluation phase. Concerning the culture, the possibility to find subjects for the experimental evaluation phase, the availability of many videos on the Internet with Indian speakers, and the difference (in terms of co-speech gestures) existing between Indian and Western culture [13] made the choice fall on Indian culture. As a remark, please consider that any language and any culture can ideally be used for training the model.

The dataset has been created by exploiting TED Talks YouTube videos, as done in [14]. Indeed, TED Talks are particularly well suited for the aim of the project (i.e., extracting speakers' skeleton keypoints and audio features): there is a huge number of available videos (including videos of Indian persons speaking English), the speech contents and the speakers are different, the speeches are well prepared (so we may expect that the speakers are using proper hand gestures) and speakers are usually standing and not even partially occluded, which allows for easily extracting their skeleton features.

Considering the Audio Conversion step, please notice that audio segments corresponding to different speakers shall be mapped to the same target speaker. The approach pursued in our work has been inspired by [15], which developed a Neural Network approach based on

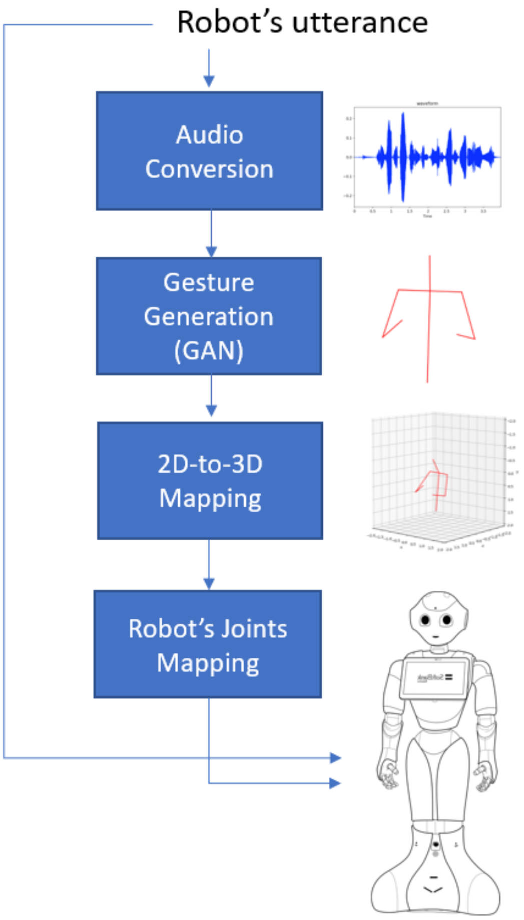

Figure 2: Run-time model for generating co-speech gestures

Phonetic PostreriorGrams (PPGs), capable of representing articulation of speech sounds in a speaker-normalized space. A PPG is a time-versus-class matrix representing the posterior probabilities of each phonetic class for each specific time frame of one utterance [16]. A phonetic class may refer to senones, words, and phones: in [15] senones were used, which consist of clusters of shared Markov states, representing similar acoustic events. It is worth saying that senones are per se speaker-independent, because they mainly rely on the typical sounds used for pronouncing a word: thus, they are particularly suited for building a Speaker-Independent Automatic Speech Recognition (SI-ASR) system [17], to the final aim of building a many-to-one voice conversion model. Differently from [15], the proposed approach may be applied to audio files of different durations: indeed, in the current implementation, the Short Term Fourier Transform has been applied multiple times for each audio file, by adopting a moving window of fixed length (512 samples), also setting a window hop (80 samples), so as to make the approach independent from the length of the audio files.

Speaking about the learning model, the log-Mel spectrogram of the converted audio files is used as input of an audio encoder, which downsamples the spectrograms through a set of convolutions, thus producing a $1D$ signal. At this point, a UNet translation architecture learns the mapping between this signal and a temporal series of keypoints. Indeed, the UNet approach has been proven to give better results than classic Convolutional Autoencoder [18], [19]. The Discriminator takes as input the extracted (and normalized) keypoints and the motion predicted by the Generator, making sure that generated motions that are too different from ground truth poses (thus, also the ones which are too smooth) are classified as fake. Inputs of 64 keypoints vectors and corresponding audio files samples have been used for training, by repeating the training process for 300 epochs, 300 steps each, reaching a total number of 90000 iterations.

Finally, in order to generate poses that can be executed by a robot, a simple neural network that estimates depth values from $2D$ poses has been used. More in detail, the network consists of a cascade of three fully connected layers with 30, 20, and 7 nodes with batch normalization, and it has been trained by using a subsection of the CMU Panoptic Dataset [20]. The $3D$ keypoints generated by the network have been remapped into 12 Degrees of Freedom of the robot (Pepper, by Softbank Robotics [21]): Head (Pitch and Yaw), Hip (Pitch and Roll), Left and Right Shoulders (Pitch and Roll), Left and Right Elbow (Roll and Yaw). The mapping was achieved by computing rotation matrices from pose vectors, and eventually the corresponding Euler angles.

This architecture was practically implemented by using *Tensorflow 2* and *Python 3*.

3. Pilot Experiment and Conclusions

Preliminary tests, involving subjects identifying themselves with different cultures (i.e. Indian and non-Indian), have proven the capability of the system to embed cultural characteristics: indeed, a statistically significant difference has been identified between the evaluations, given by Indian and non-Indian subjects, of the naturalness of the gestures generated with the proposed approach, leading to the conclusion that Indian participants have perceived *GAN-generated Gestures* as *more natural* with respect to non-Indian participants [22]. The same pilot experiment has been carried out also with randomly-generated co-speech gestures and rule-based gestures: in both cases, no statistical differences has been identified in the two groups of subjects.

Still, some improvements may be needed to the proposed approach. For example, even if the current implementation of audio conversion keeps the articulation of words, it may lose some characteristics of the original expression and intonation of the sentence, and needs to be refined. Also, the final mapping of the gestures on the robot should be improved, and adding hands gestures may be of the utmost importance in this context. Finally, the current implementation is only based on audio features, which leads to losing the semantic contents of words: a hybrid method that uses both audio features and written words, or able to mix rule-based and GAN-generated gestures, may possibly lead to better results.

References

[1] R. M. Krauss, Y. Chen, P. Chawla, Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us?, Advances in experimental social psychology 28 (1996) 389–450.

[2] M. Studdert-Kennedy, Hand and mind: What gestures reveal about thought., Language and Speech 37 (1994) 203–209.

[3] D. Archer, Unspoken diversity: Cultural differences in gestures, Qualitative sociology 20 (1997) 79–105.

[4] P. Bremner, A. G. Pipe, C. Melhuish, M. Fraser, S. Subramanian, The effects of robot-performed co-verbal gesture on listener behaviour, in: 2011 11th IEEE-RAS International Conference on Humanoid Robots, IEEE, 2011, pp. 458–465.

[5] J. R. Wilson, N. Y. Lee, A. Saechao, S. Hershenson, M. Scheutz, L. Tickle-Degnen, Hand gestures and verbal acknowledgments improve human-robot rapport, in: International Conference on Social Robotics, Springer, 2017, pp. 334–344.

[6] L. Grassi, C. Recchiuto, A. Sgorbissa, Knowledge triggering, extraction and storage via human–robot verbal interaction, Robotics and Autonomous Systems 148 (2022).

[7] L. Grassi, C. Recchiuto, A. Sgorbissa, Knowledge-grounded dialogue flow management for social robots and conversational agents, International Journal of Social Robotics (2022).

[8] C. T. Recchiuto, A. Sgorbissa, A feasibility study of culture-aware cloud services for conversational robots, IEEE Robotics and Automation Letters 5 (2020) 6559–6566.

[9] C. Recchuto, L. Gava, L. Grassi, A. Grillo, M. Lagomarsino, D. Lanza, Z. Liu, C. Papadopoulos, I. Papadopoulos, A. Scalmato, et al., Cloud services for culture aware conversation: Socially assistive robots and virtual assistants, in: 2020 17th International Conference on Ubiquitous Robots (UR), IEEE, 2020, pp. 270–277.

[10] B. Bruno, C. Recchiuto, I. Papadopoulos, A. Saffiotti, C. Koulouglioti, R. Menicatti, F. Mastrogiovanni, R. Zaccaria, A. Sgorbissa, Knowledge representation for culturally competent personal robots: Requirements, design principles, implementation, and assessment, International Journal of Social Robotics 11 (2019) 515–538.

[11] A. Khaliq, U. Kockemann, F. Pecora, A. Saffiotti, B. Bruno, C. Recchiuto, A. Sgorbissa, H.-D. Bui, N. Chong, Culturally aware planning and execution of robot actions, 2018, pp. 326–332.

[12] B. Bruno, N. Chong, H. Kamide, S. Kanoria, J. Lee, Y. Lim, A. Pandey, C. Papadopoulos, I. Papadopoulos, F. Pecora, A. Saffiotti, A. Sgorbissa, Paving the way for culturally competent robots: A position paper, volume 2017-January, 2017, pp. 553–560.

[13] R. Raina, A. Zameer, A study of non-verbal immediacy behaviour from the perspective of indian cultural context, gender and experience, International Journal of Indian Culture and Business Management 13 (2016) 35–56.

[14] Y. Yoon, W.-R. Ko, M. Jang, J. Lee, J. Kim, G. Lee, 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, 2019, pp. 4303–4309.

[15] L. Sun, K. Li, H. Wang, S. Kang, H. Meng, Phonetic posteriorgrams for many-to-one voice conversion without parallel data training, in: 2016 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2016, pp. 1–6.

[16] T. J. Hazen, W. Shen, C. White, Query-by-example spoken term detection using phonetic posteriorgram templates, in: 2009 IEEE Workshop on Automatic Speech Recognition & Understanding, IEEE, 2009, pp. 421–426.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The kaldi speech recognition toolkit, in: IEEE

2011 workshop on automatic speech recognition and understanding, CONF, IEEE Signal Processing Society, 2011.

[18] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, J. Malik, Learning individual styles of conversational gesture, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3497–3506.

[19] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[20] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, Y. Sheikh, Panoptic studio: A massively multiview system for social motion capture, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3334–3342.

[21] A. K. Pandey, R. Gelin, A mass-produced sociable humanoid robot: Pepper: The first machine of its kind, IEEE Robotics & Automation Magazine 25 (2018) 40–48.

[22] A. Gjaci, C. Recchiuto, Sgorbissa, Towards culture-aware co-speech gestures for social robots, International Journal of Social Robotics (2022).