International Journal of Social Robotics (2022) 14:1493–1506
https://doi.org/10.1007/s12369-022-00893-y

Towards Culture-Aware Co-Speech Gestures for Social Robots

Ariel Gjaci[1] · **Carmine Tommaso Recchiuto**[1] · **Antonio Sgorbissa**[1]

Accepted: 18 May 2022 / Published online: 3 June 2022

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, the definition of *natural* communicative gestures may not be trivial, since it strictly depends on the culture of the person interacting with the robot. The proposed work investigates the possibility of generating culture-dependent communicative gestures, by proposing an integrated approach based on a custom dataset composed exclusively of persons belonging to the same culture, an adversarial generation module based on speech audio features, a voice conversion module to manage the multi-person dataset, and a 2D-to-3D mapping module for generating three-dimensional gestures. The approach has eventually been implemented and tested with the humanoid robot Pepper. Preliminary results, obtained through a statistical analysis of the evaluations made by human participants identifying themselves as belonging to different cultures, are discussed.

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 [1–3], which have analyzed how conversational hand gestures may convey semantic information, and how gestures are involved in the conceptual planning of messages.

A broadly adopted classification of co-speech gestures has been introduced in [2]. Based on this classification, gestures may be *metaphoric*, describing abstract content, *iconic*, for illustrating physical actions or properties of an element, *deictic*, when pointing at a specific object in space, *beat*, rhythmic gestures which are in tune with the speech and do not carry any speech content, *adaptors*, which are hand movements directed to other parts of the body, and finally *emblems* or *symbolic*, in case gestures have a conventional meaning that often depends from the culture.

Indeed, the influence of culture in the generation of speech-accompanying gestures has been the subject of previous research works. In [4], 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 import from other cultures. About this, the documentary *A World of Gestures*, developed by the University of California in 1992 [5], offers a clear view of the extraordinary cultural diversity inherent to non-verbal communication. More recently, the work of [6] has reviewed the Literature on the cross-cultural variation of gestures, by identifying four factors that drive those variations, among which *culture-specific conventions* for form-meaning associations (e.g., the symbolic gesture constituted by a ring formed by the thumb and the index finger means "OK" in many European countries, but has a completely different meaning in other cultures), and *culture-specific gestural pragmatics* (i.e., concerning the politeness of the gesture use, the role of gesture in conversation, the gesture rate) are particularly relevant.

✉ Carmine Tommaso Recchiuto
carmine.recchiuto@dibris.unige.it

Ariel Gjaci
argjaci.95@gmail.com

Antonio Sgorbissa
antonio.sgorbissa@unige.it

[1] Department of Informatics, Bioengineering, Robotics and System Engineering, Universitá degli studi di Genova, via all'Opera Pia 13, Genova 16145, Italy

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 Interaction. Answers to this question have been given in [7]. 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 [8]: 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.

Based on this analysis, this article deals with the problem of embedding humanoid robots with the capabilities of using co-speech gestures during their verbal interaction with users. In particular, given the importance of culture in this context, this work investigates the possibility of generating culture-dependent gestures for social robots, i.e., gestures that are coherent with the expectations of people self-identifying with a given culture. To this aim, a dataset composed of audio features and poses of persons belonging to the same culture has been created. This culture-specific dataset has been used to train a Generative Adversarial Network (GAN), capable of generating 2D samples, which are then fed to a Neural Network to map the 2-Dimensional generated poses to a 3D space, and finally to the robot's joints.

Differently from similar works, which usually aim to learn the mapping between common and uncommon poses of one person, and his/her audio or speech features, here the dataset is composed of different individuals that, even if belonging to the same culture, may be quite different in terms of physical and speech characteristics. If, on the one hand, we need to have variations in movements when training generative models to learn a target distribution, on the other hand we want to avoid that elements such as the speaker's height, constitution or dimension in the image plane may negatively affect training or produce results that are not easy to map in the actual robot. Similarly, we are interested in the prosody of the audio, but we do not want that such elements as the pitch or tone of voice of the speaker are considered during training. Then, in this article we also propose to add a normalization step to correctly extract keypoints (i.e., body features significant for co-speech gesture generation, such as head, neck, and arms) and an audio conversion aimed at mapping the articulation of speech sounds of the different speakers to a virtual target speaker (*many-to-one* conversion [9]), also preserving the prosody of the original audio.

Given the absence of culture-dependent approaches for gesture generation in the Literature, and considering the main aim of this work, i.e., proposing and evaluating an integrated methodology to generate gestures perceived as culturally competent, the proposed approach has not been compared with similar methodologies. On the contrary, experimental tests have mainly focused on assessing if subjects identifying themselves with the same culture used as a reference for the dataset (experimental group) will evaluate gestures with a higher score than subjects not identifying themselves with that culture (control group). Moreover, tests have been performed also with two other methods, that have not been designed to be culturally dependent, to assess if possible differences between the experimental and the control group appear only when using our model for gesture generation or also using other models.

Finally, in the current implementation a single culturally-homogeneous dataset has been used for training the model: hence, currently, the robot is only capable to reproduce gestures for a specific culture, i.e., culturally-dependent gestures. However, in principle, different datasets could be used, allowing the robot to adapt its communicative gestures to different human cultures. The present work can be interpreted as an explorative analysis towards implementing a fully culture-aware system for co-speech gesture generation.

The paper is structured as follows: Sect. 2 reviews the recent Literature, by describing the most common approaches for autonomously generating communicative gestures, considering humanoid robots and, more generally, artificial agents. Sect. 3 describes the proposed approach, by giving an overview of the methodology adopted and discussing its implementation. Finally, Sect. 4 details the evaluation procedure carried out to assess the validity of the approach, and Sect. 5 draws conclusions.

2 Related Work

2.1 Generation of Communicative Gestures

The generation of communicative gestures for humanoid robots (and artificial agents in general), although being a quite recent field of research, is gaining growing attention in the scientific community [10,11]. The most common approaches used in this context may be classified into two main areas: *rule-based* approaches, and *data-driven* approaches [11].

Rule-based speech-gestures approaches refer to those methods in which the rules for mapping the speech to the gestures are handcrafted. It is the most widespread approach in commercial robots (e.g., the *Animated Speech* library of [12]) since it is relatively simple, robust, and gives the programmers the possibility of controlling the whole set of motion, by manually defining rules mimicking human movements. However, as a drawback, this approach works well only when considering a limited number of gestures, which may result in repetitive and boring behaviors in the long term. Given their simplicity, these approaches have been widely adopted

also in a number of research works: for example, in [13,14], the humanoid robot NAO has been made able to use a set of communicative gestures while telling stories.

Data-driven approaches may be further classified into *probabilistic* and *end-to-end* approaches. The first class corresponds to methodologies that use a probabilistic model to create a mapping between features, for example, gestures and speech. In [15], a predefined set of gestures and audio features was needed for creating an inference layer to analyze speech features, finally producing a distribution over a set of hidden states. This approach has some important limitations: as an example, the training set for the inference layer must be extensive enough to contain a representative sample of significant audio prosody features and prosody-gesture associations.

End-to-end approaches usually rely on Deep-Learning models trained on raw data to generate gestures. Some examples include recent works that have overcome the aforementioned limitations in terms of the reduced number of gestures and audio features [16,17]. Both approaches use a personalized dataset composed of YouTube videos, but while in [16] a person-specific dataset is used (thus learning the communication style of a certain person), in [17] a more general dataset based on TED[1] Talks videos has been created, giving the possibility of autonomously generating gestures in a way that is independent of a specific style. The other main difference between the two approaches lies in the data used to train the network: gestures and speech audio in [16], gestures and text in [17]. Both choices carry some advantages and drawbacks: text features allow for giving priority to speech semantics and using different speakers as a data source, but need to be extracted from subtitles (which limits the number of suitable videos) and lose the synchronization with gestures, as well as speech prosody. On the contrary, audio features keep all this information, and may be autonomously extracted from any video, but strictly depend on how the person speaks, thus the generated model will correctly work only with a specific user. On the implementation perspective, the extraction of audio features in [16] is based on the log-Mel spectrogram, which is a logarithmic spectrogram adapted on the Mel-scale, representing the frequencies used while talking and more suited for representing human voice than a common spectrogram.

Concerning the learning model, the approach presented in [16] is based on Generative Adversarial Networks (GANs) [18] that are capable of learning how to generate samples that are similar to the given dataset without providing an explicit approximation of its density function. GANs consist of two different Neural Networks: the generator (G), whose role is to produce samples according to the distribution of training data, and the discriminator (D), whose role is to estimate the probability that a sample came from G rather than from real data. These two networks are trained to play an adversarial game, where G tries to imitate the training data to fool D, while D tries to recognize if the generated data is real. Differently from [16,17] relies on the usage of Recurrent Neural Networks, RNNs to align speech and gestures. The model addresses one of the common problems of RNNs, i.e. gradient disappearance as more steps are processed.

Even concerning keypoints extraction, different approaches have been used in [16] and [17]. Indeed, since in [16] the dataset is based on one speaker, the average keypoints values and their standard deviations are sufficient to compute the normalization factor for correctly mapping skeleton keypoints. On the contrary, in [17] the normalization step also takes into account that, in this approach, the videos composing the dataset are taken using different view-angles, different cameras, different resolution and frame rates, and also speakers are different. In this case, the normalization process consists in subtracting to all keypoints of a given frame the x, y coordinates of the neck keypoint, dividing them by the distance between the shoulders of the speaker. Also, in [17], the generated 2-dimensional keypoints have been mapped to the joints of a small humanoid robot, NAO (Softbank Robotics): an additional Neural Network has been created to this aim, consisting of a cascade of three fully connected layers with 30, 20, and 7 nodes with batch normalization, while the data used to train the network was coming from CMU Panoptic Dataset [20], which provides highly accurate 3D poses of social activities including many co-speech gestures.

An additional solution has been proposed in [21], where both audio features and written words are used for training the system. The overall network consists of three encoders and a decoder for gesture generation: the generator creates poses frame-by-frame from an input sequence of features containing the speech context. Here the problem of the synchronization was specifically taken into account by incorporating the multimodal context. Concerning speech text, the exact utterance time of words is considered as known, so that word sequences have been tagged with padding tokens. While this approach has been proven to be capable of generating natural and smooth communicative gestures, the different speakers identities are still kept separate: thus, the system is able to learn the different person-specific communication styles, but it cannot generate co-speech gestures representative of the whole dataset. Interesting results have also been presented in [22]: here a feed-forward Neural Network and a simple autoregressive model have been adopted. However, the training requires a dataset composed of $3D$ keypoints and audio features, which increases the difficulty of building a custom dataset.

[1] TED is a set of conferences where speakers give short presentations on different topics, mainly technology, entertainment, and design.

Recurrent motion modeling for speech gesture generation has also been recently adopted in [23], by using an encoder-decoder structure that takes as input prosodic speech features and generates a short sequence of gesture motion. However, as in [16] the model has been proven to work only with a person-specific set of gestures. The work has been later expanded in [24], by also comparing adversarial training to standard regression loss, but still relying on a dataset composed of a single actor.

Finally, an imitation learning approach aimed at generating human-like rhytmic wave gestures has been recently proposed in [25]. The method approximates the gesture trajectory of the rythmic movement by solving the problem in the frequency domain, decomposing the signals as Fourier series. While the approach looks promising, it has been applied only to a very specific class of features, and its extension to co-speech gesture generation does not look trivial.

2.2 Culture and Robotics

The importance of culture in Social Robotics has recently gained awareness, also with a specific reference to robots' gestures. For example, culture-dependent acceptance and discomfort in relation to greeting gestures have been analyzed in [26], in which a comparative study with Egyptian and Japanese participants has been performed. This preliminary work has been further expanded, leading to a greeting selection system for a culture-adaptive humanoid robot [27]. Although analyzing an interesting aspect, the aforementioned works were dealing only with a specific aspect of Human-Robot Interaction, i.e. greeting.

Other more recent works have analyzed the relationship between cultural factors and robotics, but focusing on different aspects: rhetorical linguistic cues [28], navigation [29], interpersonal distance [30]. In all these works, findings suggest that, both concerning verbal interaction and social norms, people of different cultures tend to prefer robots better compliant with their own culture.

To the best of the authors' knowledge, the only attempt of building a culture-aware social robot has been carried out in the context of the international research project CARESSES, which was aimed to making socially assistive robots for elderly assistance *culturally competent*, i.e. able to apply an understanding of the culture, customs and etiquette of the person they are assisting, while autonomously reconfiguring their way of acting and speaking [31]. In this project, cultural aspects were considered and modelled in terms of general social norms that the robot needs to follow to comply with the user's culture [32], of planning and executing actions [33], and of verbal interaction [34]. Concerning the latter, a knowledge-based framework for culture-aware conversation has been developed: the system allows for letting the robot talk about a huge number of topics that may be more or less relevant for different cultures, by properly managing the flow of the conversation. The system, originally conceived for the humanoid robot Pepper, has been recently refactored as a portfolio of Cloud services [35,36].

3 Methodology and Implementation

The proposed work aims at evaluating the possibility of embedding "cultural brushstrokes" in autonomously generated co-speech gestures, by proposing an approach able to learn the underlying mapping between audio features and gestures of a group of persons identifying themselves with the same culture.

We remark here that this work does not intend to propose a novel approach for co-speech gesture generation. On the contrary, the work described in this manuscript has the twofold aim of:

- Implementing an integrated approach for generating communicative gestures starting from a culturally homogeneous dataset composed of different speakers;
- Assessing, through experiments with participants belonging to the target culture (experimental group) or not (control group), if the generated gestures preserve some of the *cultural brushstrokes* that characterize the communicative gestures of the persons composing the dataset.

In other words, the main problems tackled in this work are the integration of cutting-edge techniques for three-dimensional gesture generation based on a culturally homogeneous dataset composed of poses and audio features of multiple speakers, and the definition and implementation of experimental tests with human participants to evaluate if the generated gestures are perceived as culturally dependent.

Concerning gesture generation, the problem raises a number of subproblems. Among the others, the fact that the proposed model shall rely on a dataset composed of different speakers identifiable with the same culture but exhibiting different physical and audio characteristics. Also, the model shall be implementable for the automatic generation of gestures during verbal interaction with a social robot. To the best of authors' knowledge, no approaches in Literature address all these problems simultaneously. All existing approaches for co-speech gesture generation are either based on a dataset composed of a single person [16,23] or, when using a dataset composed of multiple persons [17], use text features to train the network. In the latter case, one of the main problems has been identified with the loss of synchronization between speech and generated gestures. However, when generating culture-dependent gestures, keeping the synchronization between words and gestures as well as word articulation and their affective content, is of the utmost importance. Indeed,

the way speech and gestures are synchronized has been found to be highly dependent on the semantic context [37]; hence, we can expect that a lack of synchronization between speech and gesture may negatively impact the production of culture-dependent gestures. For this reason, in this work we propose the usage of speech features, integrating a many-to-one audio conversion module before feeding the generator module, for taking into account the differences in pitch, volume, etc., between individuals considered in the training set while preserving the original speech prosody and articulation of speech sounds.

About the learning approach, data-driven state-of-the-art approaches for gesture generation are mainly based on GANs [16,21,24], or RNNs [17,23]. However, approaches that are not GAN-based have the major drawback of implementing a one-to-one mapping structure, allowing for learning a single-speaker style, without the possibility of capturing the common features that characterize the style of a group of speakers. On the contrary, GAN-based approach are capable to cope with a larger range of movements (as produced by different speakers) when compared with other data-driven models, and the one-to-many mapping structure of GANs can achieve multiple gestures generation for one same speech input with multiple random noises [11]. Please remark that this has been also shown in other application domains, where GANs are quite efficient in learning a source distribution implicitly defined by very diverse samples sharing a common "style", e.g., cultural "brushstrokes" in [38]. Moreover, the absence of an adversarial generator has been proven to cause the generation of static poses that are close to the mean pose, negatively impacting the quality of generated gestures [23]. For this reason, a GAN-based approach has been adopted in the presented work, similarly to the methodology described in [16].

Finally, most existing approaches aim to generate gestures for two-dimensional animated characters, and therefore are not suitable to be implemented on humanoid robots. The proposed work will adapt the solution proposed in [17], with the integration of a Neural Network to estimate depth values from 2D poses. The alternative solution proposed in [22] has not been taken into account, given the difficulties related to the collection of a dataset composed by three-dimensional keypoints.

Based on all these aspects, the model used for training the system has been designed, and it is shown in Fig. 1, while Fig. 2 shows the architecture used for generating the co-speech gestures of a humanoid robot in run-time.

The dataset, including video and audio (Sect. 3.1), is first pre-processed for keypoints and audio extraction through software such as OpenPose, a keypoints normalization step, and a many-to-one audio conversion approach (Sects. 3.2 and 3.3). Then, audio features are fed to the generator of a conditional GAN to produce a sequence of "faked" key-

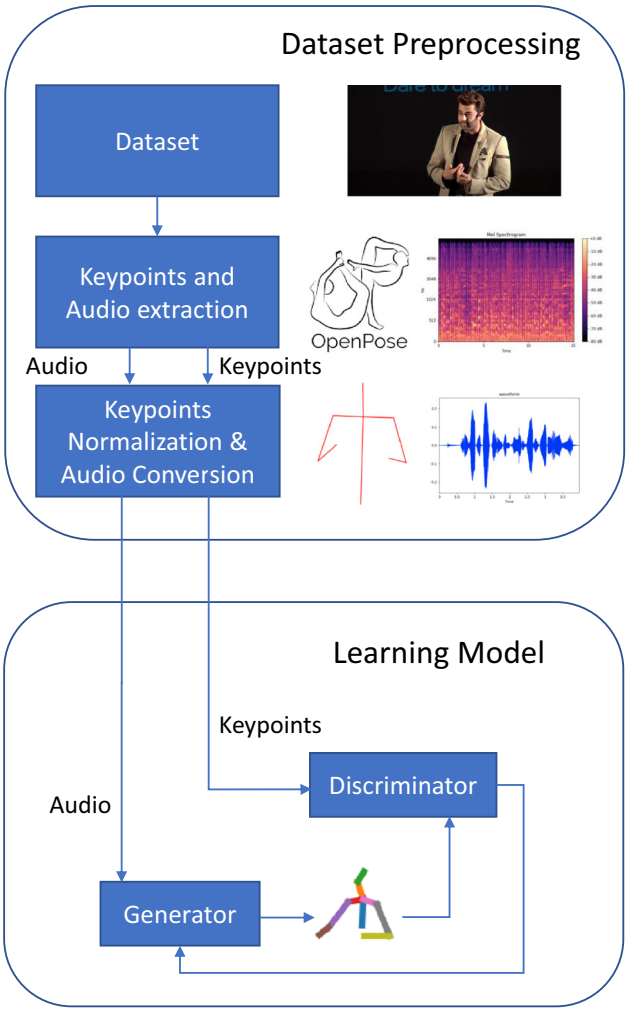


Fig. 1 Training model of the proposed approach

points, 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 (Sect. 3.4).

The run-time generation process of communicative gestures (Fig. 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 processed (see Sect. 3.3), and fed to the GAN, to produce a set of two-dimensional corresponding keypoints vectors (Sect. 3.4). 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 (Sect. 3.5), and finally mapped to the robot (Sect. 3.6).

In the following, all the elements represented in the two figures will be described in detail.

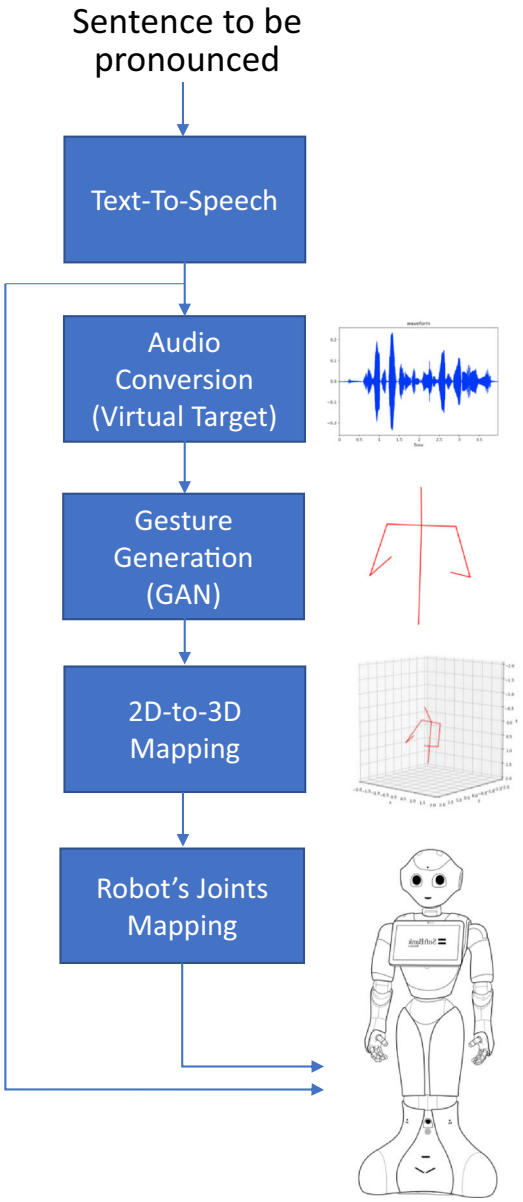


Fig. 2 Run-time model for generating co-speech gestures

3.1 Dataset

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 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 [39] made the choice fall on the 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 [17]. 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.

Two existing YouTube playlists (Josh Talks[2], which has the same concept of TED Talks, but it only contains videos of Indian people, and the official Ted Talks channel[3]) have been used as a starting point for creating the custom dataset, selecting videos with the following criteria:

- the speaker should be representative of the Indian culture. In the case of Josh Talks, this was taken for granted; in the case of TED Talks, information about the speaker has been sought.
- the speaker should speak English. All videos have been carefully checked for this purpose.
- the resolution of the video should be at least 1280x780. Outdated videos of low resolution have been excluded.
- the speaker should be standing and clearly visible. Videos where the speaker's skeleton information was not deductible have been excluded.

At the end of this process, 439 videos, with a variable duration ranging from a few minutes to an hour, and an average frame rate of 25 fps, have been chosen. All videos were autonomously downloaded by developing a *Python* script, which also checks the video duration, format, and resolution.

3.2 Keypoints and Audio Extraction

To extract all necessary information from the selected videos, in a first step all needed poses, scenes, and audio have been acquired; data have been subsequently filtered and processed to create the final dataset.

Concerning poses detection, the OpenPose library [40] has been used for extracting 9 keypoints of the body (head, neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, hip), corresponding to the ones that will

[2] https://www.youtube.com/c/JoshTalksLive/videos

[3] https://www.youtube.com/user/TEDtalksDirector

be mapped on the robot. Face and hands keypoints have not been extracted in the current implementation. Audio files were extracted by using the *ffmpeg* library on the selected videos. Finally, a further processing step has been implemented, aimed at selecting only keypoints belonging to the main speaker and checking if their body is turned back, too small, or too static.

Finally, clips have been split into subsections that last less than 10 seconds and further processed: keypoints that are possibly missing are interpolated, also smoothing the movements between frames. Indeed, since OpenPose cannot perfectly detect all keypoints, sometimes fast joint transitions between frames may occur: a Savitzky-Golay filter has been applied to the computed poses to smoothen the motion.

3.3 Keypoints Normalization and Audio Conversion

This step is necessary given that the dataset processed so far is composed of different persons, possibly filmed at a different distance from the camera and in different positions, and speaking with different audio characteristics. Concerning the keypoints normalization, the process is quite straightforward: taking into account a set of 64 poses, the x, y coordinates of the neck keypoint are subtracted from each keypoint, so as to have a common reference for all frames, and the shoulders length is used as a resize factor for all other body parts detected in each frame.

Things get more complex for the audio features. Indeed in this case the audio extracted and associated with the video clips should be translated in frequency, implementing a many-to-one voice conversion system. In other words, audio segments corresponding to different speakers shall be mapped to an ideal target speaker, in order to obtain an audio that is not dependent on pitch, volume, and tone of voice, by keeyping the original articulation of speech sounds and prosody. The approach pursued in our work has been inspired by [9], which developed a Neural Network approach based on 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 [42]. A phonetic class may refer to senones, words, and phones: in [9] 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 [43], to the final aim of building a many-to-one voice conversion model.

In the proposed implementation, the PPGs of the audio to be converted are extracted, and given as inputs to a Deep Bidirectional Long Short-Term Memory based Recurrent Neural Network (DBLSTM) [44], for generating converted speech, i.e. a resulting audio file with the characteristics of a virtual target speaker.

Differently from [9], 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.

3.4 Learning Model

The dataset built through the described steps, and thus composed of normalized keypoints and speech, taken in different scenes and from different speakers belonging to the same culture, is fed as input for generating culture-related arm and body gesture motions for a humanoid robot. To this aim, similarly to the approach used in [16], the core of the proposed method is composed of a GAN architecture, in which the generator learns the mapping from speech to gesture, while the adversarial discriminator is in charge of rejecting motions that are implausible with respect to the typical motion of the speakers of the same culture.

Considering the generator, 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. Finally, 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 [16,19]. 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.

3.5 2D to 3D Mapping

In order to generate poses that can be executed by a robot, keypoints in the $3D$ space are required. Ideally, training might be performed by using a $3D$ dataset: however, instead of using pre-recorded videos, this would require extracting $3D$ keypoints from actors to collect hundreds of hours of data, an approach that can be hardly implemented (and repeated for different cultures) in practice. For this reason, we decided to rely on a two-dimensional dataset (Sect. 3.1) and add a

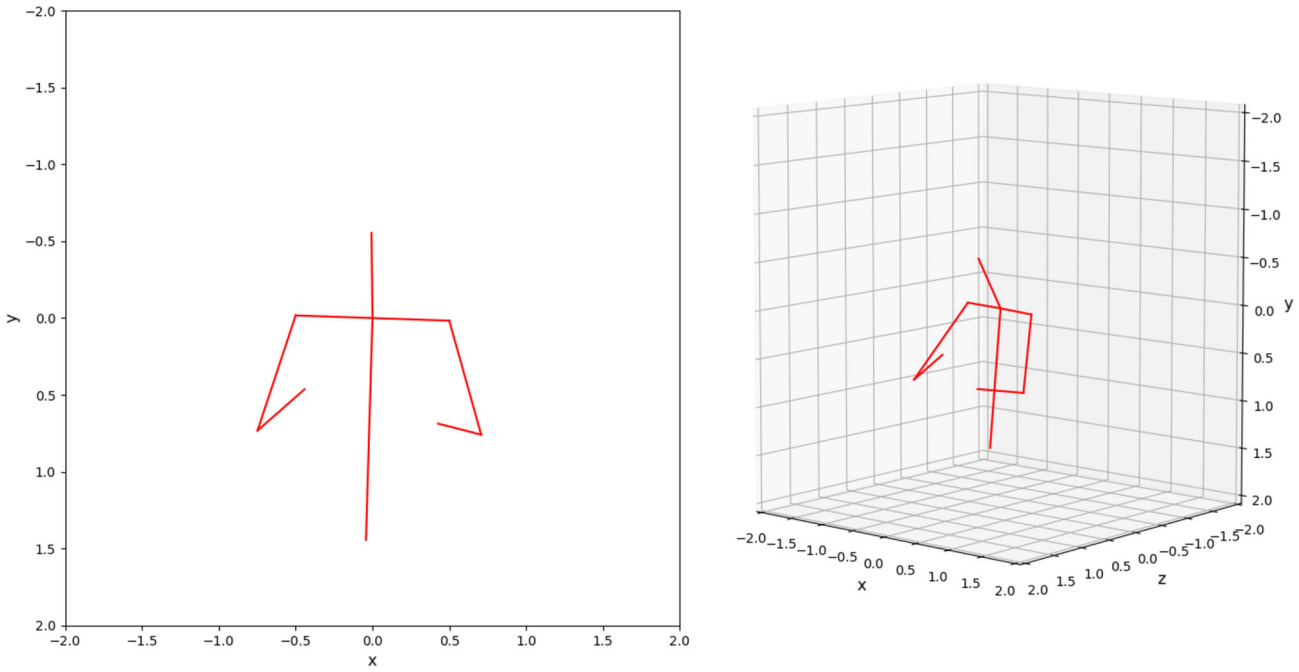


Fig. 3 Plot of a $3D$ pose vector (right) generated starting from a set of $2D$ keypoints (left)

further step to translate generated pose vectors in the three-dimensional space.

In particular, following the approach used in [17], 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 trained network allows for translating the generated keypoints in the $3D$ space, as it may be observed in Fig. 3.

3.6 Robot's Joint Mapping

The last step is obviously strictly related to the specific humanoid platform used. In the proposed implementation, the generated co-speech gestures have been generated for the humanoid robot Pepper, developed by Softbank Robotics [12].

The 9 keypoints generated (Sect. 3.4) and translated into $3D$ space (Fig. 3 and Sect. 3.5), have been remapped into 12 Degrees of Freedom of the robot: 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.

In particular, angles are extracted by filtering generated frames with a frequency of 3 Hz, and applying some corrective parameters, which take into account the limited height of the robot, and the specific configurations of TED talks, where presenters are usually on stage (so they usually look down).

Differently from [17], and also based on the different robot used for the evaluation of the proposed approach, head, and hip angles have been considered for the robot autonomous co-speech gesture generation, to enhance the expressivity of the robot. The synchronization between audio and gestures is ensured by feeding the audio file that will be pronounced by the robot (generated with Text-To-Speech functionalities) in parallel to (i) the system for gesture generation, and (ii) Pepper APIs for sound reproduction.

4 Evaluation and Discussion

An ablation study (Sect. 4.1) has been performed to understand the proposed model in detail, eliminating some components (i.e., adversarial training and many-to-one audio conversion) from the full training model.

After that, the evaluation of the proposed work has been performed relying on subjective evaluation metrics: in particular, a questionnaire based on [45] and [46] has been developed to assess how the cultural background of the user may impact on the way in which the robot's gestures are perceived (Sect. 4.2).

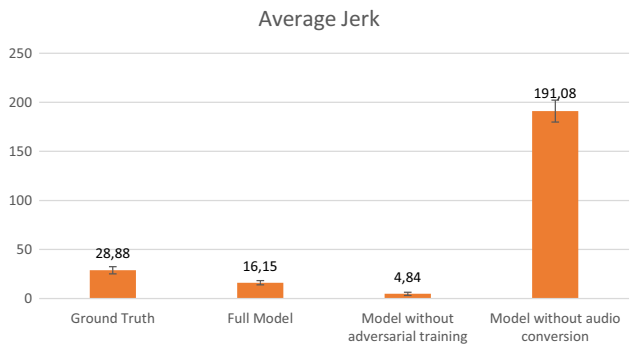


Fig. 4 Average jerk of the whole body keypoints computed from the dataset and with the three different models

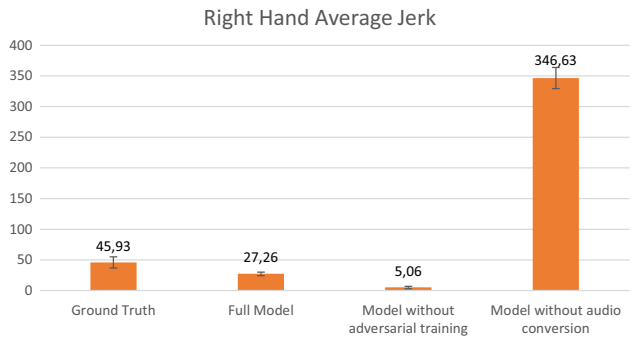


Fig. 5 Average jerk of the right hand keypoints computed from the dataset and with the three different models

4.1 Ablation Study

In order to perform the ablation study, to evaluate the effect of individual components of the proposed model on gesture generation, we have considered the average jerk of keypoints as a metric. The average jerk represents the mean of time derivatives of acceleration norms and allows for getting an idea of the smoothness of movements. It must be here pointed out that one of the main problems emerged in similar research works is the difficulty of finding evaluation metrics, i.e. quantitative measurements to assess the naturalness or dynamism of the generated gestures [45]. However, the average jerk of keypoints is usually considered in Literature as a reliable indicator for evaluating autonomous generated gestures [16,45,47].

The average jerk has been computed ablating different components, in particular the adversarial training (Sect. 3.4), to evaluate the advantages related to the integration of a discriminator in the system, and the many-to-one audio conversion module (Sect. 3.3), to assess the need for audio features conversion.

For the following analysis, the models have been trained in all cases for 300 epochs with 300 steps each, using a batch size of 32: for each model the average jerk of the body and the average jerk of the two hands have been computed (for the sake of brevity only data related to one hand are shown and discussed).

Results are shown in Figs. 4 and 5: the average jerk obtained in the ablation modalities is compared with the full model and the ground truth, averaged on all speakers.

It may be observed how the usage of GANs and the audio conversion step allow for obtaining co-speech gestures closer to the original ones in terms of smoothness. Indeed, jerk values obtained with the full model tend to be closer, both when considering full-body motion (Fig. 4) and only one hand (Fig. 5) with respect to the ground truth movements. It may be additionally pointed out how those values tend to have an extremely high value when training the system without converting the audio features: we may expect very shaky movements in this case, as a result of the highly varying speech audio features which are fed to the network. On the contrary, training the system without a discriminator generates very stiff hand and body movements, suggesting that the usage of GANs may contrast the problem of the regression towards the mean of the resulting motion: since the same utterance can lead to different gestures, the model may learn to predict an *average* motion, which may likely be a very limited movement. The presence of the discriminator ensures that the generated gestures will be representative of the motions identified in the dataset. This analysis confirms the findings of [21], showing how a model without the adversarial scheme tends to generate static poses close to the mean one.

4.2 Evaluation

A pilot test has been designed to evaluate if the proposed approach is able to generate co-speech gestures actually perceived as *culture-dependent*.

To this aim, the model has been trained with a *culture-dependent* dataset as described in Sect. 3.1: the dataset is composed of Indian speakers giving a talk in English. The model has been used for generating the motion of the joints of the humanoid robot Pepper, which was thus able to gesticulate while pronouncing a set of predefined sentences.

The hypothesis is that participants identifying themselves with the Indian culture (experimental group) will appreciate the gestures generated with the proposed approach more than non-Indian participants (control group), since these gestures should somehow embed a *culture-dependent* component, and hence Indian participants may find them more familiar. If confirmed, this hypothesis leads to the conclusion that the proposed model successfully generates gestures that preserve some of the cultural aspects embedded in the dataset.

To evaluate the method, an 80 seconds video has been realized: in the footages, the robot is facing the camera, and it pronounces different sentences. Following the guidelines

proposed in [45], we have used sentences that correspond to different types of gestures: iconic (e.g., for example corresponding to the sentence *Somebody told me that the fundamentals of baseball involve throwing the ball, hitting the ball, and catching the ball.*), metaphoric (e.g., *Playing an instrument is a great way to exercise your body and your mind.*), and beat (*I am really enthusiastic to start this day!*).

To better assess if differences in the evaluations given by the two groups of participants are only caused by the model used for gesture generation, or if possible confounding factors exist, the same aforementioned video has been realized using two other solutions for generating co-speech gestures: a rule-based approach based on the *Animated Speech* library, which exploits a number of predefined animations for accompanying Pepper's sentences with gestures that aim to be similar to the human ones, and a totally random approach (*Random Gestures*), in which Pepper gestures were generated by setting a range for the motion of each joint, and adjusting the speed so as to achieve smooth movements, also limiting the maximum acceleration and the amplitude of the motions.

Very important, please remark that our experimental protocol does not aim to test the hypothesis that one method works better than another: we only want to test if different methods are culture-dependent or not.

In other words, we do not aim here at directly confronting our GAN-based gestures with rule-based gestures (*Animated Speech*) and *Random Gestures*, but we aim at investigating how Indian and non-Indian users *react to the same sets of co-speech gestures*. While we expect that the experimental group will appreciate the gestures generated with the proposed approach more than the control group, lower differences are expected to be found when using a method that does not generate gestures in a culture-dependent way.

As an aside, additional reasons suggest how a direct comparison between the proposed approach and the other methods will not be completely fair. Firstly, a fair assessment of rule-based approaches for gesture generation will require a long interaction, which would allow subjects to perceive a possible repetitive behaviour of the robot: in our scenario, the duration of the videos is inevitably short, preventing a fair evaluation of the gestures generated with the *Animated Speech* methodology. About this, in previous trial with older care home residents in UK and Japan, psychologists found during qualitative interviews that users interacting with the robot for a longer time tend to be annoyed by an "overacting" robot [31]. On the other hand, it may be reasonable to expect that random gestures, in which the robot's movements are not synchronized with its utterances, may be evaluated as unnatural and inappropriate.

Finally, the presented approach would need some improvements before a fair comparison with state-of-the-art approaches could be done. For example, hands movements are currently not implemented in the method (differently from the *Animated Speech* rule-based approach). Also, in the proposed approach (Fig. 2 and Sect. 3.6), the same audio file corresponding to the robot's utterance should be fed to the gesture generator and provided to the robot's API for sound reproduction. This was not possible in the described experiments: to have the three methods implemented in the same conditions, the embedded TTS system has been used in run-time, while the robot's voice needed to be manually recorded in advance, to be fed to the gesture generator. This can produce a lack of synchronization between gestures and speech in the proposed approach, which can be cleared by using a TTS system that directly produce an audio file (e.g., Google TTS). All these aspects are important, but have been ignored up to now since the presented work does not aim to develop a new "off-the-shelf" method for co-speech gesture generation. Rather, we are exploring a new approach to generate culture-dependent co-speech gestures, and we believe that preliminary assessement of the system's capabilities to preserve culture-dependent features in the generated gestures may be of the utmost importance to pave the way for future works in this sense.

For the readers' convenience, examples of the videos can be found at the following links [4, 5, 6].

As mentioned before, to take into account the cultural aspects, the experimental study has been conceived with a mixed design, with one factor with two levels (Indian culture vs. non-Indian culture) that is between subjects, and one factor with three levels (the proposed approach, *Animated Speech*, and *Random Gestures*) that is within subjects, for which we counterbalanced the order to account for order effects. Participants have been recruited by means of the online platform Amazon Mechanical Turk[7], which gave the opportunity of splitting the participants based on their self-declared culture. All participants were asked to watch the three videos, presented in a random order, and fill an online questionnaire after each video. 41 Indian subjects and 41 non-Indian (mostly American) subjects, aged 21–69 (average age 31.5) took part in the study. All subjects were able to correctly understand English.

Concerning metrics, the questionnaires to be filled by the participants were taken from [46], being them also compliant with the guidelines defined in [45]. Questions evaluate the following aspects: *coherence with speech* (Do the gestures interpret the verbal information correctly?), *appropriateness* (Did you consider the observed shapes shown at different speech segments appropriate?), *fluency* (Was the movement

[4] Proposed approach: https://youtu.be/iG2-DIz2dw0

[5] Gestures generated with the *Animated Speech* library: https://youtu.be/qJ9I5cs1fik

[6] Random Gestures: https://youtu.be/GnfSAn3vzuc

[7] https://www.mturk.com/

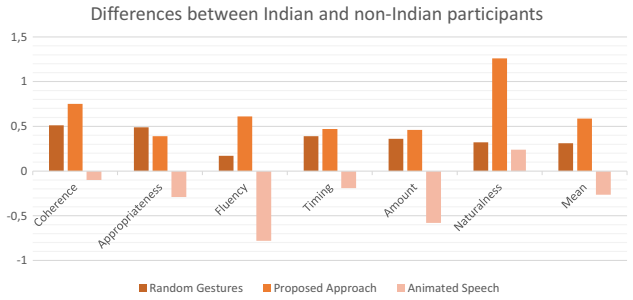


Fig. 6 Each bar reports the average difference between the scores assignment by Indian and non-Indian participants to the six items of the questionnaire and their mean value. Gestures generated with the *Animated Speech* have received higher scores from non-Indian participants (bars with a negative value); on the contrary, Indian participants have assigned higher scores to gestures generated with the other two approaches (bars with a positive value)

fluid?), *timing* (Was the speed and the timing of the represented content appropriate?), *amount of gesticulation* (Was the number of movements sufficient?), and *naturalness* (Was the overall robot's talking gesture natural and humanlike?). The last question was considered as the most relevant, since it depends on all the other aspects, and relates to what people globally perceive from the robot. Participants could answer to all items by using an 11-point Likert scale, where the lowest option corresponds to "Strongly Disagree" and the highest value corresponds to "Strongly Agree". To the best of the authors' knowledge, there are no validated questionnaire to measure the coherence, appropriateness, fluency, timing, amount, and naturalness of gestures for robots. This led us to develop a custom scale for this purpose.

The results collected from the questionnaires are shown in Fig. 6 and Table 1. A statistical analysis has been performed on the collected data: in particular, given that the sample is sufficiently large to be considered as normally distributed, a two-tailed and unpaired T-test with $p < .05$ has been applied to evaluate statistical differences between the evaluation of Indian and non-Indian participants.

The analysis of the results confirms the hypothesis made. In particular, Fig. 6 shows how the scores given by the experimental and control group exhibit the highest difference in the case of the proposed approach: here, an average difference of 0.59 may be observed, while the average differences of the two evaluations are about half in the other cases. This different cultural perception is particularly evident when comparing the evaluation given by the two groups to the *Naturalness* item in the three cases. While randomly generated and rule-based gestures have received a similar evaluation in both cases, i.e., no statistically significant differences can be found in the evaluations of the two groups of participants, a statistically significant difference can be found between Indian and non-Indian participants assessing the proposed approach (Table 1, $p < .05$), being the ratings given by Indian participants definitely higher.

The statistical analysis performed suggests that Indian participants have perceived the proposed approach as *more natural* with respect to non-Indian participants, confirming that the system is able to learn some aspects of communicative gestures that are common to the culture of the speakers used for the dataset, and can eventually be appreciated by people of the same culture.

As an aside, please notice in Table 1 that Random Gestures and Animated Speech may get higher scores that the proposed method, when absolute values are considered. However, as already discussed, our objective is not to directly compare different methods for gesture generation, but rather to evaluate what methods have the most diverse impact on different cultural groups.

Finally, it is interesting to remark that Indian participants have been more generous than non-Indian participants when scoring random-generated gestures, whereas non-Indian participants have been more generous when scoring Animated Speech. This may be due to the manual definition of the *Animated Speech* library, handcrafted by considering design choices that may be accidentally biased toward some specific cultures.

5 Conclusions

Drawing inspirations from previous works aimed at autonomously generating co-speech gestures with a specific style, the work provides an integrated approach for endowing social robots and artificial agents with the capability of gesticulating during verbal communication in a *culturally-dependent* manner. The approach, based on GANs trained with converted audio and normalized keypoints extracted from a selected dataset, has been evaluated with participants identifying themselves with different cultures, giving positive feedback about the capability of the system to embed cultural characteristics.

Beside this finding, it may also be observed how the gestures generated by the proposed method have always obtained lower scores than the gestures generated by the rule-based approach, even when evaluated by Indian participants. However, this is not completely surprising. It has been already mentioned how the final mapping of the gestures on the robot should be improved, adding hands gestures that can be of the utmost importance in this context, and using TTS systems (e.g., Google TTS) that directly produce an audio file, which can be thus straightly fed to the gesture generation model, and reproduced by the robot, ensuring the synchronization between voice and gesture.

Further improvements can be done. Even if the current implementation of audio conversion keeps the articulation

Table 1 Average scores (standard deviation in brackets) of the six items of the questionnaire filled by Indian and non-Indian participants evaluating the three methods. The table also shows the differences between the experimental and the control group, and the related p-values, highlighting in gray values lower than 0.05 (thus related to a statistical significance of 95%)

Questionnaire Results		Coherence with speech	Appropriateness	Fluency	Timing	Amount of gesticulation	Naturalness
Random Gestures	Indian participants	6.49 (2.03)	6.78 (1.92)	6.41 (2.46)	6.76 (2.19)	6.68 (2.35)	6.32 (2.57)
	Non-Indian participants	5.98 (2.66)	6.29 (2.23)	6.24 (2.28)	6.37 (2.33)	6.32 (2.54)	6.00 (2.50)
	Differences	0.51	0.49	0.17	0.39	0.36	0.32
	P-Values	0.3732	0.3582	0.8044	0.5461	0.5449	0.6940
Proposed Approach	Indian participants	6.51 (2.18)	6.34 (1.92)	6.22 (2.29)	6.49 (1.79)	6.51 (1.86)	6.80 (2.20)
	Non-Indian participants	5.76 (2.45)	5.95 (2.53)	5.61 (2.23)	6.02 (2.23)	6.05 (2.57)	5.54 (2.86)
	Differences	0.75	0.39	0.61	0.47	0.46	1.26
	P-Values	0.2079	0.5231	0.2688	0.3825	0.4288	0.0340
Animated Speech	Indian participants	6.83 (1.79)	6.98 (1.99)	6.90 (1.92)	7.15 (1.85)	6.76 (1.77)	7.46 (1.79)
	Non-Indian participants	6.93 (2.66)	7.27 (1.82)	7.68 (2.50)	7.34 (1.85)	7.34 (1.54)	7.22 (1.46)
	Differences	-0.10	-0.29	-0.78	-0.19	-0.58	0.24
	P-Values	0.7091	0.4587	0.1424	0.6036	0.1322	0.5759

of words, it may lose some characteristics of the original expression and intonation of the sentence, and needs to be refined. Also, the current implementation is only based on audio features, which leads to losing the semantic contents of words: in the case of iconic gestures, this may be a great limitation, suggesting that a hybrid method that uses both audio features and written words, or possibly able to mix rule-based and data-driven approaches, may lead to better results.

Finally, the system needs to be trained with different culture-dependent datasets, providing the robots with the capability of adapting in run-time its way of interacting, in terms of co-speech gestures, to different cultural contexts, hence achieveing a complete culture awareness.

Funding Open access funding provided by Universitá degli Studi di Genova within the CRUI-CARE Agreement. The authors state that this work has not received any funding.

Data availibility The datasets generated during the current study are available from the corresponding author on reasonable request.

Declarations

Informed consent Informed consent, constituted by the first page of the online survey, was obtained from all subjects involved in this study. Institutional Review Board approval was not required because responses to the survey were anonymous.

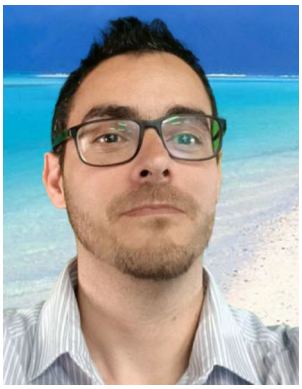
References

1. Krauss RM, Chen Y, Chawla P (1996) Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? Adv Exp Soc Psychol 28:389–450
2. Studdert-Kennedy M (1994) Hand and Mind: What Gestures Reveal About Thought. Lang Speech 37(2):203–209
3. Alibali MW, Kita S, Young AJ (2000) Gesture and the process of speech production: We think, therefore we gesture. Lang Cognit Process 15(6):593–613
4. Archer D (1997) Unspoken diversity: Cultural differences in gestures. Qual Sociol 20(1):79–105
5. Archer D (1992) A world of gestures: Culture and nonverbal communication. video) Berkeley: University of California Extension

Center for Media and Independent Learning-2000 Center Street. Fourth Floor, Berkeley, California 94704:642–0460
6. Kita S (2009) Cross-cultural variation of speech-accompanying gesture: A review. Lang Cognit Process 24(2):145–167
7. Bremner P, Pipe AG, Melhuish C, Fraser M, Subramanian S (2011, October) The effects of robot-performed co-verbal gesture on listener behaviour. In: 2011 11th IEEE-RAS International Conference on Humanoid Robots. IEEE, p 458–465
8. Wilson JR, Lee NY, Saechao A, Hershenson S, Scheutz M, Tickle-Degnen L (2017, November) Hand gestures and verbal acknowledgments improve human-robot rapport. In: International Conference on Social Robotics. Springer, Cham, p 334–344
9. Sun L, Li K, Wang H, Kang S, Meng H (2016, July) Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In: 2016 IEEE International Conference on Multimedia and Expo (ICME), IEEE, p 1–6
10. Kucherenko T, Jonell P, Yoon Y, Wolfert P, Henter GE (2021, April) A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020. In: 26th International Conference on Intelligent User Interfaces, p 11–21
11. Liu Y, Mohammadi G, Song Y, Johal W (2021, November) Speech-based Gesture Generation for Robots and Embodied Agents: A Scoping Review. In: Proceedings of the 9th International Conference on Human-Agent Interaction, p 31–38
12. Pandey AK, Gelin R (2018) A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. IEEE Robot & Autom Mag 25(3):40–48
13. Le QA, Hanoune S, Pelachaud C (2011, October) Design and implementation of an expressive gesture model for a humanoid robot. In: 2011 11th IEEE-RAS International Conference on Humanoid Robots. IEEE, p 134–140
14. Meena R, Jokinen K, Wilcock G (2012, December) Integration of gestures and speech in human-robot interaction. In 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom). IEEE, p 673–678
15. Levine S, Krähenbühl P, Thrun S, Koltun V (2010) Gesture controllers. In: ACM SIGGRAPH 2010 papers, p 1–11
16. Ginosar S, Bar A, Kohavi G, Chan C, Owens A, Malik J (2019) Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition p 3497–3506
17. Yoon Y, Ko WR, Jang M, Lee J, Kim J, Lee G (2019, May) 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, p 4303–4309
18. Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA (2018) Generative adversarial networks: An overview. IEEE Signal Process Mag 35(1):53–65
19. Ronneberger O, Fischer P, Brox T (2015, October) U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, p 234–241
20. Joo H, Liu H, Tan L, Gui L, Nabbe B, Matthews I, Sheikh Y (2015) Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision, p 3334–3342
21. Yoon Y, Cha B, Lee JH, Jang M, Lee J, Kim J, Lee G (2020) Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Trans on Graph (TOG) 39(6):1–16
22. Kucherenko T, Hasegawa D, Henter GE, Kaneko N, Kjellström H (2019, July) Analyzing input and output representations for speech-driven gesture generation. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, p 97–104
23. Ferstl Y, McDonnell R (2018, November) Investigating the use of recurrent motion modelling for speech gesture generation. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents, p 93–98
24. Ferstl Y, Neff M, McDonnell R (2019) Multi-objective adversarial gesture generation. In: Motion, Interaction and Games, p 1–10
25. Panteris M, Manschitz S, Calinon S (2020, March) Learning, Generating and Adapting Wave Gestures for Expressive Human-Robot Interaction. In: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, p 386–388
26. Trovato G, Zecca M, Sessa S, Jamone L, Ham J, Hashimoto K, Takanishi A (2013) Cross-cultural study on human-robot greeting interaction: acceptance and discomfort by Egyptians and Japanese. Paladyn. J Behav Robot 4(2):83–93
27. Trovato G, Zecca M, Do M, Terlemez Ö, Kuramochi M, Waibel A, Takanishi A (2015) A novel greeting selection system for a culture-adaptive humanoid robot. Int J Adv Rob Syst 12(4):34
28. Andrist S, Ziadee M, Boukaram H, Mutlu B, Sakr M (2015, March) Effects of culture on the credibility of robot speech: A comparison between english and arabic. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, p 157–164
29. Truong XT, Ngo TD (2017) Toward socially aware robot navigation in dynamic and crowded environments: A proactive social motion model. IEEE Trans Autom Sci Eng 14(4):1743–1760
30. Patompak P, Jeong S, Nilkhamhang I, Chong NY (2020) Learning proxemics for personalized human-robot social interaction. Int J Soc Robot 12(1):267–280
31. Papadopoulos C, Castro N, Nigath A, Davidson R, Faulkes N, Menicatti R, Sgorbissa A (2021) The CARESSES Randomised Controlled Trial: Exploring the Health-Related Impact of Culturally Competent Artificial Intelligence Embedded Into Socially Assistive Robots and Tested in Older Adult Care Homes. International Journal of Social Robotics, 1-12
32. Sgorbissa A, Papadopoulos I, Bruno B, Koulouglioti C, Recchiuto C (2018, October) Encoding guidelines for a culturally competent robot for elderly care. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, p 1988–1995
33. Khaliq AA, Köckemann U, Pecora F, Saffiotti A, Bruno B, Recchiuto CT, Chong NY (2018, October) Culturally aware planning and execution of robot actions. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, p 326–332
34. Bruno B, Recchiuto CT, Papadopoulos I, Saffiotti A, Koulouglioti C, Menicatti R, Sgorbissa A (2019) Knowledge representation for culturally competent personal robots: requirements, design principles, implementation, and assessment. Int J Soc Robot 11(3):515–538
35. Recchiuto CT, Sgorbissa A (2020) A feasibility study of culture-aware cloud services for conversational robots. IEEE Robot Automat Lett 5(4):6559–6566
36. Recchuto C, Gava L, Grassi L, Grillo A, Lagomarsino M, Lanza D, Sgorbissa A (2020, June) Cloud services for culture aware conversation: Socially assistive robots and virtual assistants. In: 2020 17th International Conference on Ubiquitous Robots (UR). IEEE, p 270–277
37. Bergmann K, Aksu V, Kopp S (2011) The relation of speech and gestures: Temporal synchrony follows semantic synchrony. In: Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GeSpIn 2011)
38. Zaino G, Recchiuto CT, Sgorbissa A (2022) Culture-to-Culture Image Translation with Generative Adversarial Networks. arXiv preprint arXiv:2201.01565
39. Raina R, Zameer A (2016) A study of non-verbal immediacy behaviour from the perspective of Indian cultural context, gender and experience. Int J Ind Cult Bus Manag 13(1):35–56

40. Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y (2019) OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. IEEE Trans Pattern Anal Mach Intell 43(1):172–186
41. PySceneDetect (2021) PySceneDetect: Intelligent scene cut detection and video splitting tool. Retrieved July 13, 2021, from https://pyscenedetect.readthedocs.io/en/latest
42. Hazen TJ, Shen W, White C (2009, December) Query-by-example spoken term detection using phonetic posteriorgram templates. In: 2009 IEEE Workshop on Automatic Speech Recognition & Understanding. IEEE, p 421–426
43. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Vesely K (2011) The Kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF). IEEE Signal Processing Society
44. Sun L, Kang S, Li K, Meng H (2015, April) Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, p 4869–4873
45. Wolfert P, Robinson N, Belpaeme T (2021) A review of evaluation practices of gesture generation in embodied conversational agents. arXiv preprint arXiv:2101.03769
46. Mlakar I, Kačič Z, Rojc M (2013) TTS-driven synthetic behaviour-generation model for artificial bodies. Int J Adv Rob Syst 10(10):344
47. Kucherenko T (2018, October) Data driven non-verbal behavior generation for humanoid robots. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, p 520-523

Ariel Gjaci is a Ph.D. student in Robotics and Autonomous Systems at the University of Genoa. He has obtained a Bachelor Degree in Biomedical Engineering and a Masters Degree in Robotics Engineering, he is a robotics and tehcnology enthusiast, and he is currently working on a project about Culture-Aware Artificial Intelligence applied to Humanoid Robots. He also has a six-month experience in developing embedded software solutions for military applications. His interests include Machine Learning, Deep Learning, and Social Robotics.



Carmine Tommaso Recchiuto Ph.D is Assistant Professor at the University of Genoa, where he teaches Experimental Robotics, ROS programming, and Computer Science. His research interests include Humanoid and Social Robotics (with a specific focus on knowledge representation and human-robot interaction), wearable sensors, and Aerial Robotics. He is currently the Technical Coordinator of the DIONISO project, a multidisciplinary effort focusing on ICT for intervention in earthquakes, mainly working on analyzing and developing algorithms for human localization and mapping with wearable sensors. He has been the Coordinator of software integration and Head of Software Development for the CARESSES project, aimed at endowing social robots for older adults with cultural competence. Also, he has been the local Coordinator for the BrainHuRo project, developing Brain-Computer Interfaces for humanoid robots' remote control. He is the author of more than 50 scientific papers published in International Journals and conference proceedings, and Associate Editor for Intelligent Service Robotics by Springer.

Antonio Sgorbissa Ph.D is Associate Professor at the University of Genoa, where he teaches Real Time Operating Systems, Social Robotics, and Ambient Intelligence in EMARO+, the European Master In Advanced Robotics. He is the Coordinator of National and EU research projects, among which the H2020 project CARESSES (caressesrobot.org). Also, he is the local Coordinator of the ongoing IENE 10 project, aimed at preparing health and social care workers to work with intelligent robots in health and social care environments. His research focuses on Autonomous Robotics Behaviour, with a focus on Culture Awareness, Knowledge representation, Motion Planning, Wearable, and Ubiquitous Robotics. He is the author of about 150 articles indexed in international databases and has been a member of the Organizing Committee in the top ranked robotic conference as well as Associate Editor for the International Journal of Advanced Robotic System edited by SAGE and Intelligent Service Robotics by Springer.