UNIVERSITÀ DEGLI STUDI DI GENOVA

DIBRIS

DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

BIOENGINEERING, ROBOTICS AND SYSTEM ENGINEERING

REASEARCH TRACK 2

Autonomous co-speech gesture generation for social robotics

Literature analysis

Authors:

Borelli Simone - S4662264
Gavagna Veronica - S5487110

Professors:

Carmine Recchiuto

s.y. 2022/23

Overview about the Generation of Co-speech Gestures of social robots

Simone Borelli, *Student*, UNIGE
Veronica Gavagna, *Student*, UNIGE,

16 June 2023

Abstract

Human communication involves verbal and nonverbal cues, including gestures, to convey messages, opinions, emotions, and intentions. In human-robot communication, robots need to incorporate natural gestures that accompany speech, but achieving this is complex and dependent on cultural context.
Co-speech gestures are vital for emphasizing words, conveying meaning, and enhancing communication. Robots must integrate content and manner of communication from their partners to generate appropriate responses.
While social robots can mimic human-like gestures, they often require human professionals and effort to establish speech-gesture connections. Teaching co-speech gestures can be done through rule-based or data-driven approaches, with the latter using probabilistic or end-to-end methods.
The aim of our research is to emphasize the importance of natural behaviour in human-robot interactions, with a focus on increasing acceptability and engagement and exploring the potential for robots to be perceived as individuals during conversations.
Although progress has been made in human-humanoid co-speech gesture communication, there is room for improvement, as social robots still lack the naturalness required to fully replicate human gestures.
Bridging this gap and achieving a truly human-like experience will be essential in establishing robots as authentic conversational partners.

Keywords

Social robots, human-robot communication, cultural context, co-speech gestures, gesture generation, anthropomorphism.

Introduction

Humans employ not only verbal language but also nonverbal cues such as tone of voice, body language, and facial expressions to convey messages, opinions, emotions, and gestures [4]. Effective social connections are closely associated with exchanging nonverbal signals alongside verbal communication. This is why, when it comes to communication between humans and social robots, it is crucial to incorporate the ability of robots to accompany their speech with natural gestures[11] [6].
However, the notion of natural communicative gestures may not be straightforward, as it depends on the cultural context of the individual interacting with the robot [2].
During interactions with others, people engage in various actions that are used to emphasize words, convey meaning, provide clarification, or offer vivid descriptions while speaking. These gestures are known as co-speech or co-verbal gestures [11].
There exist four main types of co-speech gestures [10]:

- **Iconic**:
  they have a form, which is visually related to the concept being communicated;
- **Metaphorical**:
  they have an arbitrary relation to the concept they communicate;
- **Deictic**:
  they are used to point out elements of interest or to communicate directions;
- **Beat**:
  they do not carry semantic meaning and are often used to emphasize the rhythm of speech;

Despite the fact that social interaction is an indefinite concept, it can be formalized through two main processes: perception, which involves receiving visual stimuli about other people's behaviours or the interaction's state; action, which involves creating a behaviour while taking into account all aspects of interaction [5]. Hence, combining the content and manner of communication from the interaction partner becomes essential to generate socially appropriate responses for robots [5].

Recent social robots like Pepper and RoboThespian can make co-speech gestures that resemble those

of humans, but human professionals create these ones. Furthermore, it takes a lot of human effort to create links between speech phrases and gestures [11].
For teaching co-speech gestures to a robot, there are two possible approaches: the first one is *rule-based* and the second one is *data-driven* [2].
The former gives rules for mapping the speech to the gestures, which are handcrafted and give the programmers the possibility of controlling the motion, but with the usage of a limited number of gestures, while the latter is classified into two additional categories: *probabilistic* and *end-to-end* approaches[2].
The first class corresponds to methodologies that use a probabilistic model to create a mapping between features, the second one usually relies on Deep-Learning models trained on raw data to generate gestures [2].

Research question

The focus of our research is to highlight the importance of making robots interact with humans using natural behaviour. This has a twofold aim to:

- Increase acceptability among humans and enhance engagement, overcoming possible prejudices
- Explore the potential for robots to be perceived as individuals during conversations

Exploring the complexity surrounding the potential for a robot to emulate human behaviour and be perceived as a person leads us to our research question:

"***Can a robot, while engaging in natural conversation, be perceived as a person?***"

This is a huge question, but the existing literature provides valuable insights into this area of study.

Methodologies

People are influenced by earlier research in psychology, human-robot interaction, and computer animation on gesture modelling. Although psychological research into gesture modelling dates back many years, computational scientists have just recently begun to pay attention to it [4]. Studies of natural gestures, however, only cover a small portion of the pertinent literature [10] [4]. To try to answer the question that has been mentioned before, several papers are compared and described: this section focuses only on methodologies that are used to perform communication between humans-humans and human-humanoid robots.
In the first experiment done in 2013, Cory J. Hayes, Charles R. Crowell, and Laurel D. Riek tried to prove if gestures, made either by humans or humanoid robots, are perceived in the same way by humans [3]. They used the NAO robot, which is a humanoid robot with 6 degrees of freedom per arm; gestures are told to the robot thanks to Choreographe, a development environment provided by Aldebaran. They involved 59 young students, which saw some videos where a human or robot actor performed some gestures, in particular knocking and stirring, accompanied by a human or robot voice. Gestures can be congruent or incongruous with the speech, and two trial was done [3]:

- The first one reported videos with actors doing both gesture and voice in a congruent way (e.g.: human gesture and human voice, or robot gesture and robot voice);
- In the second one, the participants saw other videos where actors doing the gesture could be not the same speaking, and the gesture could also be not congruent with the pitch. Hence, it was possible to have a robot voice with human gestures and vice versa.

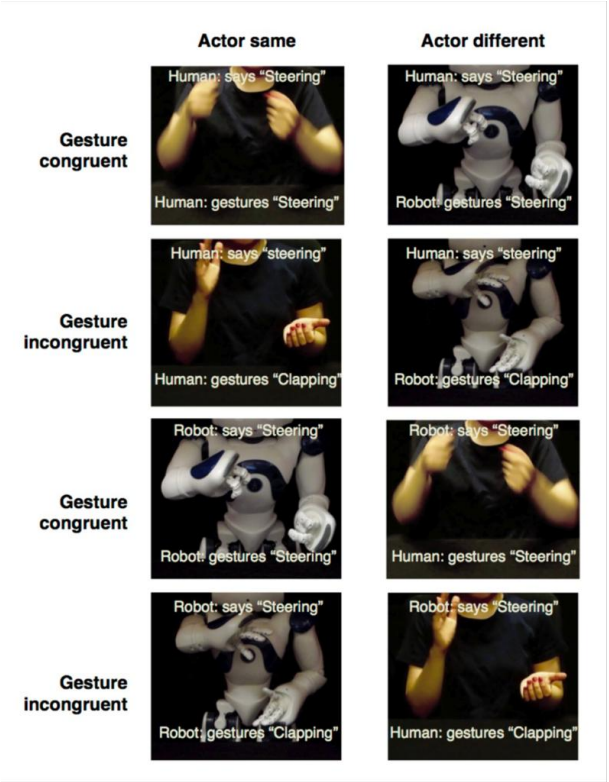

Figure 1: Example still frames of the congruent and incongruent stimuli for the actor-same and actor-different videos

A few years later, in 2015, similar research was made by Paul Bremner, University of The West of England [1], using the same humanoid robot and video stimuli, but here he focused only on beat gestures. Paul Bremner wanted to show if there were differences in speech and gesture integration between human and robot communicators. In order to achieve that, an experiment involved 22 participants, aged 18-55, all Native English speakers.
Two sets of stimuli were recorded, one for humans and another one for the NAO robot. In both sets, there were 10 ambiguous sentences, accompanied by a beat gesture that emphasises the speech: a downward vertical movement of the hand. As technical information, the similarity between robot and human gestures was guaranteed, as close as possible, by analyzing the joint motion profile, recorded by a Kinect [1].
Both researches were based on a common main problem about determining if gesture and speech used together, are processed automatically for language comprehension. This question was discussed by Kelly in [7]: the researchers mentioned earlier focused only on how humans perceived gestures but did not say anything about how can improve the naturalness of gestures learned by robots.

Figure 2: NAO robot

In 2019, Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee conducted research that use videos as a database and employed the NAO robot in real-time [11]. They believed that since humanoid robots have similar appearances and control joints, copying human gestures is a workable technique for them. Since their goal was to teach robots to act naturally, they want to train a model able to generate continuous gestures for any speech text of any length. In the paper [11], the authors present a learning-based approach for generating co-speech gestures using a large dataset of 52 hours of TED talks. Their end-to-end neural network model includes an encoder for speech-text understanding and a decoder for generating a sequence of gestures, which successfully produces various gesture types [11].

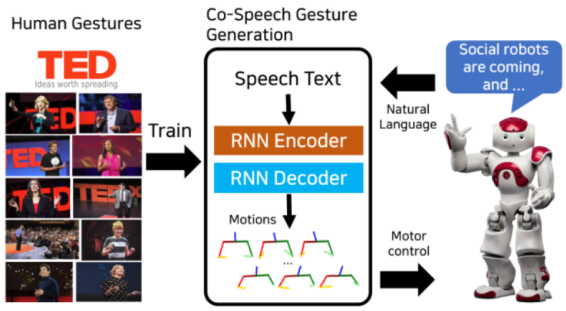

Figure 3: They address a problem of making co-speech gestures for a given speech text. The proposed model generates a sequence of upper-body poses, and it is trained from human gestures in TED talks.

Taking up where we last on the naturalness of gestures and their meaning, it is important to underline that the previous concept is culturally dependent. The researchers Carmine Tommaso Recchiuto, Antonio Sgorbissa and Ariel Gjaci, based their experiment on this aspect, generating gestures which are culture-depended [2].
For example, culture-dependent acceptance and discomfort in relation to greeting gestures have been analyzed in [9], in which a comparative study with Egyptian and Japanese participants have been performed. This preliminary work has been further expanded, leading to a greeting selection system for a culture-adaptive humanoid robot [8].
In their work, they used a custom dataset of Indian people, provided by two YouTube playlists of TED talks made by Josh Talks, to make Pepper robots behave more like humans. A data-driven approach is used during the replication of gestures, which relied on the learning model ”Generative Adversarial Networks” (GANs). For doing that it is used a 2D to 3D mapping module for creating three-dimensional motions, and a speech conversion module to manage the multi-person dataset [2].

Figure 4: Pepper robot, which is used for interacting with people

Results

Based on the research conducted by Cory J. Hayes, Charles R. Crowell, and Laurel D. Riek, it became apparent that participants exhibited faster response times (in voice identification tasks), increased accuracy rates, and reduced brain activity when the actor and voice were matched, irrespective of congruence. However, when confronted with a mismatched and incongruous situation, specifically when gestures were made by a robot while the speech originated from a human voice, participants were unable to automatically process the gestures [3].
Hence, this suggests that robot communicators do not provoke the automatic processing of their co-speech gestures, potentially due to past human experiences. Building upon the previous findings, the experiment conducted by Paul Bremner measured the proportion of high attachment responses, specifically referring to the selection of the first noun as the subject of the relative clause [1]. An interesting outcome emerged: when it comes to verbal emphasis in speech, there is no distinction between humans and robots, as both use the same audio. However, what differs is the influence of beat gestures: in the case of humans, there is an increased probability of attachment [1].
This thing is in contrast with the initial expectation of research, and that can be explained in two possible ways:

- Gender incompatibility between gesture and speech;
- When the robot is replicating a gesture, it does not have a ”communicative intent”, that must be necessary to emphasise a speech.

Since the goal of the research made by Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee, was to develop a model capable of generating continuous gestures, the proposed method outperformed the baselines in the subjective evaluation across all measures of anthropomorphism, likeability, and speech-gesture correlation [11]. Through subjective evaluations, participants reported that the generated gestures were human-like and aligned with the speech content:
”*The participants in the evaluation support this as follows: “positive impressions were human-like movements not stiff, moving freely”; “when the robot arms are not moving in a predictable human fashion, it actually hurts the experience”; “I had a positive impression when the speech correlated with the motions”; and “positive impression when I felt like the motions flowed smoothly with the content of what was being said.”*” [11] The authors also demonstrate the implementation of the model with a real-time NAO robot. Their contributions include the creation of a large-scale dataset with speech transcripts, the development of a novel gesture generation model, and the realization of gestures in a robot prototype. However, some participants expressed dislike for excessive and exaggerated gestures, suggesting that a balance between clear gestures and excessive gesturing should be achieved [11]. ”*Several participants disliked the exaggerated motions. They commented as follows: “I got a negative impression if the gesture was too ‘jerky’ or fast,” and “looked much more brash ... jumped around from motion to motion.” One participant suggested that a few but clear gestures are better than incessant gesturing.*” [11].
The study focused solely on speech text and did not consider audio, leading to a potential mismatch between generated gestures and speech audio. Future directions include incorporating audio for tighter coupling between gestures and speech audio, and exploring personalizing parameters for social robots, such as expressiveness, cultural dependency, and politeness [11].
Regarding cultural dependency, the research made by Carmine Tommaso Recchiuto, Antonio Sgorbissa and Ariel Gjaci, showed positive results [2]: the robot reacted well concerning the ability to incorporate cultural characteristics, but it was noted that the gestures, replicated by this method, received a lower score compared to gestures generated by a rule-based approach; probably this was due to some different aspects [2]:

- The mapping of gestures needs to be improved;
- A system implementation based only on audio features could lead to a loss of semantic meaning, for instance in the case of iconic gestures;
- The system should be trained with different cultural datasets;

These aspects could be an interesting object of discussion in future experiments.

Conclusions

Overall, the research discussed in the various articles highlights the significance of congruence between speech and gestures in human-robot communication. By achieving better alignment and incorporating cultural and contextual factors, future advancements in gesture generation for robots can enhance their anthropomorphism, likability, and overall effectiveness as communicative entities.
Taking up where we last on the research questions, more improvements are necessary in human-humanoid co-speech gesture communication, because social robots do not have enough naturalness in replicating gestures, also in the case of learning base approach.
Certainly, the current findings are promising: robots are capable of effective communication with humans. However, this alone is insufficient for a human to perceive that they are engaging in conversation with another human being. Although the current results are encouraging, showcasing the capability of robots to engage in effective communication with humans, there is a gap in accomplishing a truly human-like experience. Achieving a higher level of naturalness and seamless integration of gestures and speech will be crucial in closing the distance between human and robot interactions, enabling humans to perceive robots as true conversational partners.

Improvements

To further improve human-robot communication, several key areas can be focused on.

First, future research should concentrate on developing more sophisticated and contextually aware gesture generation models, along with incorporating cultural dependencies and personalizing parameters to enhance the overall human-like perception during conversations with robots. Incorporating contextual and multicultural factors into gesture generation algorithms can also contribute to more socially sensitive and contextually appropriate gestures.

Secondly, there is a need to address the challenge of achieving a truly human-like experience during conversations with robots. This requires seamless integration of gestures and speech to create a cohesive and synchronized communication process. Future research should explore techniques to better align gestures with speech content, ensuring that they complement each other and enhance the overall communication experience. Additionally, personalizing parameters for social robots, such as expressiveness and politeness, can further contribute to creating a more human-like interaction.

Furthermore, when building and creating robot communicators, it is crucial to take human perception into account. Designing efficient interaction systems requires a thorough understanding of how people interpret and react to robot gestures. The user experience can be enhanced by collecting user input and performing user studies, which can reveal important insights into the preferences and expectations of human participants.

Overall, by addressing these areas of improvement, future advancements in gesture generation can enhance the anthropomorphism, likability, and overall effectiveness of robots as communicative entities. Bridging the gap between human and robot interactions requires a continued focus on achieving naturalness, seamless integration of gestures and speech, and understanding human perception, ultimately enabling humans to perceive robots as true conversational partners, for overcoming prejudices.

References

[1] Paul Bremmer. “Speech and Gesture Emphasis Effects For Robotic and Human Communicators-a Direct Comparison”. In: *ACM/IEEE International Conference on Human-Robot Interaction* (2015).

[2] Antonio Sgorbissa Carmine Tommaso Recchiuto and Ariel Gjaci. “Towards Culture-Aware Co-Speech Gestures for Social Robots”. In: *International Journal of Social Robotics* (2022).

[3] Cory J. Hayes, Charles R. Crowell, and Laurel D. Riek. “Automatic Processing of Irrelevant Co-Speech Gestures with Human but not Robot Actors”. In: *ACM/IEEE International Conference on Human-Robot Interaction* (2013).

[4] Yu-Jung Chae, Changjoo Nam, Daseul Yang, Hun-Seob Sin, ChangHwan Kim, Sung-Kee Park. “Generation of co-speech gestures of robot based on morphemic analysis”. In: *ELSEVIER - Robot and Automation Systems* (2022).

[5] Nguyen Tan Viet Tuyen and Oya Celiktutan. “Agree or Disagree? Generating Body Gestures from Affective Contextual Cues during Dyadic Interactions, Nguyen Tan Viet Tuyen and Oya Celiktutan”. In: *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (2022).

[6] Nick DePalma, Sonia Chernova, Jesse Smith, Jessica Hodgins. “Toward a One-interaction Data-driven Guide: Putting co-Speech Gesture Evidence to Work for Ambiguous Route Instructions”. In: *ICRA: International Conference on Robotics and Automation* (2019).

[7] S.Kelly, P.Creigh, J-Bartolotti. “Integration Speech and Iconic Gestures in a Stroop-like Task: Evidence for Automatic Processing”. In: *J. Cog. Neurosci* (2010).

[8] Trovato G, Zecca M, Do M, Terlemez Ö, Kuramochi M, Waibel A, Takanishi A. “A novel greeting selection system for a culture-adaptive humanoid robot”. In: *J Adv Rob Syst* (2015).

[9] Trovato G, Zecca M, Sessa S, Jamone L, Ham J, Hashimoto K and Takanishi A. “Cross-cultural study on human-robot greeting interaction: acceptance and discomfort by Egyptians and Japanese”. In: *Paladyn. J Behav Robot* (2013).

[10] Pieter Wolfert and Tony Belpaeme. “A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents”. In: *IEEE Transactions on Human-Machine Systems* (2022).

[11] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. “Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots”. In: *ICRA International Conference on Robotics and Automation* 52.3 (2019).