Conflict management training in the workplace through simulation with socio-affective embodied conversational agent

Alice Delbosc, Magalie Ochs, Nicolas Sabouret, Brian Ravenet, Stéphane Ayache

Conflict management training in the workplace through simulation with socio-affective embodied conversational agent

Alice Delbosc*†
Davi, The Humanizers
Puteaux, France
alice.delbosc@lis-lab.fr

Magalie Ochs
CNRS, LIS, Aix-Marseille University
Marseille, France
magalie.ochs@lis-lab.fr

Nicolas Sabouret
CNRS, LISN, Paris-Saclay University
Orsay, France
nicolas.sabouret@universite-paris-saclay.fr

Brian Ravenet
CNRS, LISN, Paris-Saclay University
Orsay, France
brian.ravenet@limsi.fr

Stéphane Ayache
CNRS, LIS, Aix-Marseille University
Marseille, France
stephane.ayache@lis-lab.fr

ABSTRACT

Workplace conflicts hold significant importance and can greatly impact the overall dynamics of an organization. Effectively managing these conflicts can contribute to a healthier and more harmonious work environment and potentially leading to increased productivity levels. Consequently, it's imperative to provide training in conflict management skills for addressing and resolving workplace conflicts. One method largely used for social skill training consists in role-playing scenarios. Given the cost of such sessions, virtual agents appear to be an interesting tool to set them up. Today, one of the main limitations of using virtual agents is the lack of believability in the agents' non-verbal behavior, not allowing interaction to be as natural as with humans and limiting user engagement. This PhD aims to automatically generate the socio-affective non-verbal behavior of a virtual agent, that will be integrated into a tool for conflict management training. We present the identified steps for developing this agent and provide an overview of the work completed thus far.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Animation**.

KEYWORDS

Conflict simulation, non-verbal behavior, embodied conversational agent, neural networks, adversarial learning

1 INTRODUCTION

The more diverse and complex our society becomes, the more differences in points of view or objectives can lead to potential conflicts [18]. In particular, interpersonal conflicts at work are quite common, causing stress and health problems [11, 61]. While they have long had an exclusively negative reputation [34, 54], it is beginning to be accepted that conflicts can bring positive results [4, 33], such as improved autonomy, social cognitive skills or negotiating capacity [42]. It is therefore important to learn how to behave in different conflict situations and identify the opportunities that can arise from them. If this applies to all types of conflict, those that occur in the workplace are especially important, as their proper management can contribute to a healthier, more peaceful working environment, and perhaps even an increase in productivity [67].

Training by practice has been shown to enhance learning [8], one of the reasons why over the past few years, systems based on virtual agents for training purposes have gained in popularity [13]. They enable the reproduction of believable situations in a safe environment [18], with a sense of engagement for the user. Training systems integrating virtual agents exist in a number of fields, for instance in the medical domain [2, 3, 50]. In the field of conflict management, few virtual agents have been deployed [30, 43].

Davis [19] has shown that virtual agent gestures have positive effects on knowledge transfer and information retention. Other studies have also shown that head movements improve the way a virtual agent is perceived in general [14, 45], and "uncanniness" is increased for a character whose facial expressions are perceived as insufficient [64]. However, current training systems for virtual agents frequently rely on Wizard of Oz methods, which lack the advantages offered by fully autonomous systems, such as faster access to training, adaptability, and reduced training costs [60]. Despite their potential, systems using virtual agents for training are therefore not widely adopted in practice. The interactions with virtual agents are not as natural as it is with humans, a major obstacle for user engagement.

This PhD aims to address this weakness by creating a model for the **automatic generation of socio-affective non-verbal behaviors for a virtual agent, within the context of workplace conflict management.** The conflict management use case was chosen for its capacity to simulate several social attitudes. The paper is organised as follows: after reviewing existing works in section 2, we present the use case and the approach adopted to implement it in section 3. In section 4, we present the work we've done so far and our results. Finally, we conclude in section 5.

2 STATE OF THE ART

2.1 The conflict. Several researchers have proposed definitions of conflict. Traditionally, conflict has been defined as opposing interests involving limited resources and differing goals. A well-known definition in the literature is provided by Thomas [63], who defines conflict as "the process that begins when one party perceives that another has frustrated, or is about to frustrate, one of its concerns". This incompatibility of objectives between the agents can be seen as the cognitive dimension of conflict.

Some authors enhance their definition by incorporating a behavioral aspect, proposing that in order for conflict to be present, at least one party must actually behave in a way that interferes with

*Also with CNRS, LIS, Aix-Marseille University.
†Also with CNRS, LISN, Paris-Saclay University.

another party's objective [20, 65]. Similarly, if two individuals have opposing interests, at least one of them must recognize those interests [27]. These definitions highlight two types of conflict: mutual (where all parties involved are aware of the conflict) or unilateral (only one party is aware of the conflict).

Lindner [44] studies conflict by focusing on emotions such as fear or anger, illustrating how they affect conflict and how they are affected by conflict. For Kolb and Putnam [37], who study conflict within organizations, a conflict occurs when there are real or perceived differences that arise in specific organizational circumstances and generate emotions as a result. Emotions are therefore an integral part of the construction of conflict, thereby adding a cognitive dimension to the equation.

These three dimensions: cognitive, behavioral and affective, are at the heart of most definitions of conflict. While some use one or two of them to define conflict, a study by Barki and Hartwick [6] showed that these dimensions must exist together for a real conflict to exist. Emotions are therefore an essential dimension of conflict, and conflict cannot exist without emotions. This affective dimension of conflict is particularly interesting for us. According to Nair [47], emotional expressions occur largely non-verbally through facial expressions, vocal characteristics and body postures. Whittaker [68] highlight that feelings, emotions, and attitudes are often not expressed verbally and should therefore be inferred from non-verbal channels. The presence of non-verbal communication is essential in any interaction, especially one aimed at simulating a conflict.

2.2 Virtual agents and conflict. Conflict management training tools can use a variety of supports like serious games [15, 16] or agent systems [43, 57]. These different supports can take the form of *automated mediation systems* [26] or *conflict simulation systems* [57]. In *an automated mediation system*, an agent helps the parties involved in the conflict to communicate and resolve the situation. They can use speech and emotion analysis techniques. *Conflict simulation systems* use agents to reproduce a conflict situation and ask users to take an active part in resolving it. Our study covers *conflict simulation systems*.

Virtual agent has several advantages. An autonomous virtual agent is always available and doesn't need a human to control it. Once it is programmed, it's a very economical way of training [60]. Virtual agents also provide the opportunity to practice, as often as necessary and according to Blanch-Hartigan et al. [10], this is the key to developing interpersonal skills. The scenarios set up with virtual agents can reproduce situations that are stressful or anxiety-provoking. The secure environment of the simulation increases the sense of challenge while reducing the sense of threat. It's only when a participant feels challenged and not threatened that he or she can go further in learning [23].

However, these systems are not yet widely used. There are still many obstacles to their use, such as technological complexity, high production costs, ethical issues and resistance to change. One of the obstacles we are particularly interested in is the lack of believability, not allowing user engagement [12, 59].

2.3 Automatic generation of non-verbal behavior. In order to structure the state of the art, we present examples of rules-based models; we describe data-driven models; we explore the different possible input and their involvement in the generated behaviors; finally, we discuss the output representation.

Rules-based approaches The first approach explored for automatic generation of virtual agents' behavior was based on sets of rules, one of the first was Cassell et al. [17] with *Animated Conversation*.

The development of new rule-based systems often required the development of a new domain-specific language (DSL). These DSLs were often incompatible with each other [49], this is why, Kopp et al. [38] developed a unified language for generating multimodal behaviors for virtual agents, called behavior Markup Language (BML). BML has become the standard format for rule-based systems and many other works have followed using this format [46, 56].

It is important to point out that rules-based approaches focused on intention. They were highly effective in terms of communication, but not very natural, since they mainly inserted predefined animations [49]. More recent research has therefore begun to explore data-based systems.

Data-driven approaches Data-driven approaches do not depend on experts in animation and linguistics. Recently, Yang et al. [69] proposed a motion graph-based statistical system that generates gestures for dyadic conversations. However, these statistical approaches are still based on an animation dictionary, limiting the diversity of the generated movements. There is only one motion sequence for an input signal. It supports the hypothesis of an injective speech-motion correspondence, even though it is "One-To-Many".

Kucherenko et al. [40] proposed an encoder-decoder speech to motion. Even though the generated behaviors are not based on a dictionary, this approach is deterministic and tend to generate average motion representations. To address this issue, researchers have explored the integration of probabilistic components into their generative models. In particular, Generative Adversarial Networks (GANs) have been employed [31, 58, 62]. GANs can convert acoustic speech features into non-verbal behaviors while preserving the diversity and multiple nature of the generated non-verbal behavior. Given the performance of GANs in the area of non-verbal behavior generation, we choose to explore adversarial approaches.

In comparison with rule-based approaches, data-driven approaches have made advancements in terms of naturalness of gestures, by generating continuous and fairly natural gestures, but the gestures are significantly less communicative. Recent research has attempted to preserve the advantages of both methods by creating hybrid systems such as Zhuang et al. [73]. In the context of the PhD, the use of such approaches will probably be necessary.

Inputs of the models behavior generation models can take audio input [31, 39], textual input [9, 71], or both [1, 25, 70]. Approaches using only audio produced usually well-synchronized movements, which correspond to rhythmic gestures, but the absence of text transcription implies that they lack of semantic meaning.

Other forms of input are used, such as non-linguistic modalities (e.g., interlocutor behavior) [35, 48] or control input (e.g., style parameters transmitted during model inference) [25, 28]. The ability to control body motion based on a specific input signal such as their emotional state or a social attitude can significantly improve the usability of the method [28]. In our context of developing a conflict management tool with a virtual agent, it is crucial to possess the ability to simulate various social attitudes.

Outputs of the models Even if facial expressions and head movements are all connected and synchronised with speech [17], most of the previous works only generate facial animations *OR* head movements. As far as we know, only the work of Habibie et al. [29] proposed the automatic generation of facial expressions and head movements jointly from an adversarial approach. Inspired by their work, we generate facial expressions and head movements in a combined way, changing the representation of facial expressions.

While body and head movements are generated with 3D coordinates, facial expressions can be generated in various ways. They can be generated directly with the 3D coordinates of the face [29, 36] or describe using a model [35, 53]. In our work, we represent the facial expressions using action units (AUs) based on the well-known Facial Action Coding System (FACS) [22]. This choice is motivated by the objective to obtain interpretable and explainable results and therefore be able to manipulate the generated facial expressions, for instance to express a particular social attitude [21, 66].

3 THE USE CASE

3.1 The scenario. The conflict management use case was chosen for its ability to express a range of social attitudes. We aim at reproducing a conflict between two people, for example agent A who acts inappropriately towards his/her colleague represented by another agent B. Various scenarios could be simulated, such as discriminatory or aggressive behavior. The user's task is to interact with the virtual agent A and persuade him/her that it has done something inappropriate.

There are therefore two conflicts: the first between agents A and B, and the second between agents A and the user. The first conflict can be pre-recorded or described to the user before the role-playing with the virtual agent. Following the typology presented in Barki and Hartwick [7] on the levels of conflict analysis, we are interested in a unilateral interpersonal social conflict in a company.

An agent conflict management training tool has many important dimensions. We will limit our work to some of them. We will work on the generation of the virtual agent's non-verbal behavior in an interaction situation, i.e. by taking into account the user's non-verbal and verbal input signals. For this purpose, we will build a system able to detect in real time the user's multimodal signals (facial expressions, voice, intonation, etc.) and adapt the agents' behaviors accordingly. However, we won't be working on the textual dimension of the scenario. To avoid tackling the issue of automatic speech recognition, the scenario will be scripted. We won't be working on automatic socio-emotional signals detection either.

To simulate the agent's behavior, we will consider the following inputs: user signals, audio and textual features and social attitude labels. The output will be a sequence of AUs, head movements and gaze direction. These behavioral characteristics could then be represented in a BML or directly simulated on a virtual agent.

3.2 Key challenges. Among the key challenges in generating socio-affective non-verbal behavior in interactions, certain issues consistently arise within the research community. To begin with, the availability of high-quality datasets containing the necessary features is limited. Even when available, the quality may be insufficient. Another important topic is the limited and sometimes unreliable nature of objective evaluation methods in this field, as demonstrated by the findings of Nyatsanga et al. [49].

Our objective is to generate high-quality movements that are both rhythmically and semantically consistent. However, the GENEA challenges [72] reveals that while believability levels in motion generation can reach those of motion capture, the coherence of the generated motion often only slightly exceeds chance. Furthermore, few works take interlocutor features into account. The challenge lies in enabling the virtual agent to react coherently during interactions and consider the user's mental states in real time to create a responsive and interactive system.

3.3 Our step-by-step approach. To create a virtual agent, simulating different socio-emotional behavior interactively with the user, different research issues have to be tackled: generating behaviors when the agent is speaking ("speaking" behavior), when the agent is listening ("listening" behavior), taking the user's behavior into account and integrate a social attitude constraint. Our work is organized around four distinct steps, each contributing to the overall system's capabilities. By following this step-by-step approach, we aim at developing a system able to generate nuanced and socio-affective non-verbal behaviors of virtual agents, during interaction scenarios. This methodology enables to compare and select at each stage the most suitable architecture for advancing to the next stage.
The first step focuses on generating rhythmically relevant non-verbal behaviors for the virtual agent as he speaks. This involves creating a model that generates non-verbal features that align with the rhythmic patterns of speech.
The second step aims at generating semantically and contextually relevant non-verbal behaviors for the virtual agent. By aligning non-verbal behaviors with the semantic content of the agent's speech, we will enhance the communicative effectiveness and expressiveness of the virtual agent.
The third step will incorporate the socio-affective dimension. We will introduce a social attitude constraint (aggressiveness, concilience or denial) to guide the generation of non-verbal behaviors.
The fourth step will take into account the behaviors exhibited by the human interlocutor. This will enable the virtual agent to dynamically adjust its non-verbal behavior (in real-time) to align with and effectively engage with the interlocutor. At this stage, we could also focus on "listening" behaviors.

At this time, we have focused on the first step of this approach.

4 PROGRESS AND RESULTS

4.1 Problem formulation. The problem of our first step can be formulated as follows: given a sequence of acoustic speech features, the task is to generate the sequence of corresponding head movements, gaze and facial expressions that a virtual agent should play while speaking. To simulate the generated behaviors on an embodied conversational agent, we use the Greta platform [52].[1]
vspace-1mm
4.2 Dataset and processing of the data. Our task requires a dataset that captures interaction scenarios with a focus on facial recordings. We use the *Trueness* dataset [51], containing 18 interaction scenes of discrimination with 6 different actors. We divide each

[1]This work was performed using HPC/AI resources from GENCI-[CINES/IDRIS/TGCC] (Grant 2022- [AD011014211]).

scene into two parts, representing the perspectives of the first and second persons, resulting in a total of 36 videos with approximately 3 hours and 30 minutes of recording time.

Additionally, we employ the *Cheese* Corpus [55], selecting 10 interaction scenes involving free student conversation, resulting in 20 videos with 20 different speakers, providing approximately 5 hours of recording time. The difference with *Trueness* is that these aren't actors, and they aren't conflict scenes, so their behavior is less expressive. This dataset also differs in terms of shooting conditions, the students are located a little further away from the camera and almost their entire bodies are filmed. Consequently, throughout this article, we will refer to *Cheese*, as having "farther-away shooting conditions" and being "less expressive".

We automatically extract visual and acoustic speech features from the existing videos using state-of-the-art tools [5, 24]. To highlight the distinction between speaking and listening behaviors, we apply adjustments to the extracted data. Including setting the head and gaze coordinates to the center position, and setting all action units intensity to a constant, when the person is not speaking.

4.3 Model. Following the research conducted during the state of the art, our proposed architecture adopts an adversarial encoder-decoder approach. Our model consists of two neural networks: a generator and a discriminator. The generator learns to produce data that looks like the real distribution, while the discriminator aims to distinguish between the generated and real data. Our generator contains a single encoder and three decoders, each dedicated to a specific feature type (head, gaze, and facial action units). Inspired by Jang et al. [32], the discriminator not only learns from generated and real data, it also learns from false examples designed to enhance the model's understanding of speaking and listening behaviors and improve behavior synchronization with speech. These examples associate audio features of a "speaking" person with behavior features of a "listening" person (and vice versa).

4.4 Research questions and hypotheses. Reflecting the objectives of our first stage, our research question is: which factors influence the model to obtain more or less human-like behaviors and more or less speech-matched behaviors? We make the following assumptions:
(H1) The perception of speech/behavior synchronization will be improved with the addition of our fake examples during training.
(H2) The addition of "less expressive" data during training, will improve the perception of believability.
(H3) The addition of "farther-away shooting conditions" data during training will degrade the perception of synchronization.

Based on our hypotheses, we compare four conditions:
$m1$: architecture presented, trained on *Trueness* dataset.
$m2$: architecture presented, trained on *Trueness* and *Cheese*.
$m3$: model $m1$ without our fake examples during training.
GDS: "Ground Truth Simulated" is the extracted behavior from the data, directly simulated on the virtual agent.

4.5 Evaluation. We conduct both objective and subjective evaluations to assess the performance of our models.
Objective evaluation With objective metrics, we select an optimal architecture for each of our conditions. We use distribution Dynamic Time Wrapping (DTW) to calculate the distance between the ground truth and the generated behaviors. We also use the comparison of acceleration and jerk.

Objective metrics primarily focus on statistical similarity to ground truth signal, rather than contextual appropriateness and coherence with speech. Which is why subjective evaluations play a crucial role in assessing the complexity of social communication.
Subjective evaluation Through user perceptive study, we evaluate two criteria [41]: the believability and the temporal coordination with speech. We randomly generate four sequences around 30 seconds[2], by considering the two evaluated criteria, the four sequences, and the four conditions, we obtained a total of 32 videos. Thirty persons participated in our study. They viewed each of the videos in a random order and rated them.

The results reveal the superiority of $m1$ compared to $m3$ in terms of synchronization ($p < 0.05$) and also in terms of believability ($p < 0.01$). (H1) is significantly validated. We also observe the dominance of $m2$ in terms of believability compared to m1 ($p < 0.01$) and the superiority of $m1$ in terms of coordination ($p < 0.05$). (H2) and (H3) are then significantly validated. An interesting result is that $m1$ tends to outperform *GDS* in terms of synchronization ($p = 0.067$), an uncommon result in the field of behavior generation. We assume that setting "listening" behaviors to a constant and including our fake examples improves the perception of synchronization with speech.

5 CONCLUSION AND PERSPECTIVES

The use of virtual agents for conflict management training offers many advantages. However, they are not widely used today, one of the main obstacles remains the limited interaction they allow with users compared to human-human interactions. To overcome this problem, our work is structured into four steps, each encompasses key aspects of our research and contribute to the development of the socio-affective embodied conversational agent. Until now, our focus has been on the generation of rhythmically relevant and believable non-verbal behavior for the virtual agent as it speaks.

We found that adding data doesn't necessarily increase performances, expressiveness of people within the dataset and shooting conditions are key elements. By employing an adversarial model, we provided the model with incorrect examples that improve its understanding of the synchronization between speech and behaviors, enhancing the performance of it.

Obviously, our findings must be nuanced due to the many limitations, such as the number of people involved in the subjective evaluation or the limitation of the extraction and visualisation tools (limited number of AUs for example).

It is crucial to emphasize that the tool we are developing is not designed to "judge" users' performance. Instead, its purpose is to support users in integrating and enhancing their conflict management skills in a workplace setting. Research by Schmid Mast et al. [60] demonstrated that feedback from human experts is more effective in improving performance compared to feedback from algorithms. Therefore, the tool could be used as part of a training process, beginning with conflict simulation and ending with feedback from human evaluators.

[2]https://www.youtube.com/playlist?list=PLRyxHB7gYN-Cs127qTMJIR78fsQu_8tZQ

REFERENCES

[1] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. 2020. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1884–1895.

[2] Mohammed Al Owayyed, Myrthe Tielman, and Willem-Paul Brinkman. 2022. Virtual Patients to Train Communication Skills of Healthcare Providers. (2022).

[3] Glenn Albright, Craig Bryan, Cyrille Adam, Jeremiah McMillan, and Kristen Shockley. 2018. Using virtual patient simulations to prepare primary health care professionals to conduct substance use and mental health screening and brief intervention. *Journal of the American Psychiatric Nurses Association* 24, 3 (2018), 247–259.

[4] Allen C Amason. 1996. Distinguishing the effects of functional and dysfunctional conflict on strategic decision making: Resolving a paradox for top management teams. *Academy of management journal* 39, 1 (1996), 123–148.

[5] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[6] Henri Barki and Jon Hartwick. 2001. Interpersonal conflict and its management in information system development. *MIS quarterly* (2001), 195–228.

[7] Henri Barki and Jon Hartwick. 2004. Conceptualising the Construct of Interpersonal Conflict. *International Journal of Conflict Management* 15 (March 2004), 216–244. https://doi.org/10.1108/eb022913

[8] Wendy L Bedwell, Stephen M Fiore, and Eduardo Salas. 2014. Developing the future workforce: An approach for integrating interpersonal skills into the MBA classroom. *Academy of Management Learning & Education* 13, 2 (2014), 171–186.

[9] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*. IEEE, 1–10.

[10] Danielle Blanch-Hartigan, Susan A Andrzejewski, and Krista M Hill. 2012. The effectiveness of training to improve person perception accuracy: A meta-analysis. *Basic and Applied Social Psychology* 34, 6 (2012), 483–498.

[11] Niall Bolger, Anita DeLongis, Ronald C Kessler, and Elizabeth A Schilling. 1989. Effects of daily stress on negative mood. *Journal of personality and social psychology* 57, 5 (1989), 808.

[12] Dario Bombari, Marianne Schmid Mast, Elena Canadas, and Manuel Bachmann. 2015. Studying social interactions through immersive virtual environment technology: virtues, pitfalls, and future challenges. *Frontiers in psychology* 6 (2015), 869.

[13] Merijn Bruijnes, Jeroen Linssen, and Dirk Heylen. 2019. Special issue editorial: Virtual Agents for Social Skills Training. , 2 pages.

[14] Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth Narayanan. 2007. Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE transactions on audio, speech, and language processing* 15, 3 (2007), 1075–1086.

[15] Henrique Campos, Joana Campos, João Cabral, Carlos Martinho, Jeppe Herlev Nielsen, and Ana Paiva. 2013. My Dream Theatre (Demonstration). (2013).

[16] Henrique Campos, Joana Campos, Carlos Martinho, and Ana Paiva. 2012. Virtual Agents in Conflict. In *Intelligent Virtual Agents (Lecture Notes in Computer Science)*, Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker (Eds.). Springer, Berlin, Heidelberg, 105–111. https://doi.org/10.1007/978-3-642-33197-8_11

[17] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. 413–420.

[18] Yun-Gyung Cheong, Rilla Khaled, Corrado Grappiolo, Joana Campos, Carlos Martinho, Gordon PD Ingram, Ana Paiva, and Georgios Yannakakis. 2011. A computational approach towards conflict resolution for serious games. In *Proceedings of the 6th international conference on foundations of digital games*. 15–22.

[19] Robert O Davis. 2018. The impact of pedagogical agent gesturing in multimedia learning environments: A meta-analysis. *Educational Research Review* 24 (2018), 193–209.

[20] Morton Deutsch. 1973. *The resolution of conflict: Constructive and destructive processes*. Yale University Press.

[21] Paul Ekman. 2002. Facial action coding system (FACS). *A Human Face, Salt Lake City* (2002).

[22] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[23] Andrew J Elliot. 2013. *Handbook of approach and avoidance motivation*. Psychology Press.

[24] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[25] Mireille Fares, Catherine Pelachaud, and Nicolas Obin. 2023. Zero-shot style transfer for gesture animation driven by text and speech using adversarial disentanglement of multimodal style encoding. *Frontiers in Artificial Intelligence* 6 (2023), 1142997.

[26] Sajika Gallege and Eric Lederer. 2010. Oh no you didn't! Mediating Conflict Between Humans and Intelligent Virtual Agents. (2010).

[27] Paulo Gomes and Arnav Jhala. 2013. AI authoring for virtual characters in conflict. In *Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*.

[28] Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. 2022. A motion matching-based framework for controllable gesture synthesis from speech. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.

[29] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 101–108.

[30] Kerstin S. Haring, Jessica Tobias, Justin Waligora, Elizabeth Phillips, Nathan L. Tenhundfeld, Gale Lucas, Ewart J. de Visser, Jonathan Gratch, and Chad Tossel. 2019. Conflict mediation in human-machine teaming: using a virtual agent to support mission planning and debriefing. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1–7.

[31] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. 2018. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 79–86.

[32] Yunseok Jang, Gunhee Kim, and Yale Song. 2018. Video prediction with appearance and motion conditions. In *International conference on machine learning*. PMLR, 2225–2234.

[33] Karen A Jehn. 1995. A multimethod examination of the benefits and detriments of intragroup conflict. *Administrative science quarterly* (1995), 256–282.

[34] Karen A Jehn. 1997. A qualitative analysis of conflict types and dimensions in organizational groups. *Administrative science quarterly* (1997), 530–557.

[35] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. 2020. Let's Face It: Probabilistic Multi-modal Interlocutor-aware Generation of Facial Gestures in Dyadic Settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[36] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.

[37] Deborah M Kolb and Linda L Putnam. 1992. The multiple faces of conflict in organizations. *Journal of organizational behavior* (1992), 311–324.

[38] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents: 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, August 21-23, 2006. Proceedings 6*. Springer, 205–217.

[39] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 97–104.

[40] Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. 2021. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *International Journal of Human–Computer Interaction* 37, 14 (2021), 1300–1316.

[41] Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. Evaluating gesture-generation in a large-scale open challenge: The GENEA Challenge 2022. *arXiv preprint arXiv:2303.08737* (2023).

[42] Brett Laursen and Christopher A. Hafen. 2010. Future directions in the study of close relationships: Conflict is bad (except when it's not). *Social Development* 19, 4 (2010), 858–872. Publisher: Wiley Online Library.

[43] Minha Lee, Jan Kolkmeier, Dirk Heylen, and Wijnand IJsselsteijn. 2021. Who Makes Your Heart Beat? What Makes You Sweat? Social Conflict in Virtual Reality for Educators. *Frontiers in psychology* (2021), 1365. Publisher: Frontiers.

[44] Evelin G Lindner. 2006. Emotion and conflict: Why it is important to understand how emotions affect conflict and how conflict affects emotions. *The handbook of conflict resolution* 2 (2006), 268–293.

[45] Soroosh Mariooryad and Carlos Busso. 2012. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 8 (2012), 2329–2340.

[46] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 25–35.

[47] Nisha Nair. 2008. Towards understanding the role of emotions in conflict: a review and future directions. *International Journal of Conflict Management* 19, 4 (2008), 359–381.

[48] Tan Viet Tuyen Nguyen and Oya Celiktutan. 2022. Context-Aware Body Gesture Generation for Social Robots. In *ICRA 2022 Workshop on Prediction and Anticipation Reasoning for Human-Robot Interaction*.
[49] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. *arXiv preprint arXiv:2301.05339* (2023).
[50] Magalie Ochs, Daniel Mestre, Grégoire De Montcheuil, Jean-Marie Pergandi, Jorane Saubesty, Evelyne Lombardo, Daniel Francon, and Philippe Blache. 2019. Training doctors' social skills to break bad news: evaluation of the impact of virtual environment displays on the sense of presence. *Journal on Multimodal User Interfaces* 13 (2019), 41–51.
[51] Magalie Ochs, Jean-Marie Pergandi, Alain Ghio, Carine André, Patrick Sainton, Emmanuel Ayad, Auriane Boudin, and Roxane Bertrand. 2023. A forum theater corpus for discrimination awareness. *Frontiers in Computer Science* 5 (2023), 1081586.
[52] Catherine Pelachaud. 2015. Greta: an interactive expressive embodied conversational agent. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. 5–5.
[53] Hai Xuan Pham, Yuting Wang, and Vladimir Pavlovic. 2018. End-to-end learning for 3d facial animation from speech. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 361–365.
[54] Louis R Pondy. 1967. Organizational conflict: Concepts and models. *Administrative science quarterly* (1967), 296–320.
[55] Béatrice Priego-Valverde, Brigitte Bigi, and Mary Amoyal. 2022. CHEESE!: Corpus «CHEESE!». *TIPA. Travaux interdisciplinaires sur la parole et le langage* 38 (2022).
[56] Brian Ravenet, Catherine Pelachaud, Chloé Clavel, and Stacy Marsella. 2018. Automating the production of communicative gestures in embodied characters. *Frontiers in psychology* 9 (2018), 1144.
[57] Andrew Robb, Casey White, Andrew Cordar, Adam Wendling, Samsun Lampotang, and Benjamin Lok. 2015. A comparison of speaking up behavior during conflict with real and virtual humans. *Computers in Human Behavior* 52 (2015), 12–21. Publisher: Elsevier.
[58] Najmeh Sadoughi and Carlos Busso. 2018. Novel realizations of speech-driven head movements with generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6169–6173.
[59] Shane Saunderson and Goldie Nejat. 2019. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics* 11 (2019), 575–608.
[60] Marianne Schmid Mast, Emmanuelle P Kleinlogel, Benjamin Tur, and Manuel Bachmann. 2018. The future of interpersonal skills development: Immersive virtual reality training with virtual humans. *Human Resource Development Quarterly* 29, 2 (2018), 125–141.
[61] Michael T Sliter, Shuang Yueh Pui, Katherine A Sliter, and Steve M Jex. 2011. The differential effects of interpersonal conflict from customers and coworkers: Trait anger as a moderator. *Journal of Occupational Health Psychology* 16, 4 (2011), 424.
[62] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. 2017. Speech-to-gesture generation: A challenge in deep learning approach with bi-directional LSTM. In *Proceedings of the 5th International Conference on Human Agent Interaction*. 365–369.
[63] Kenneth W Thomas. 1992. Conflict and conflict management: Reflections and update. *Journal of organizational behavior* (1992), 265–274.
[64] Angela Tinwell, Mark Grimshaw, Debbie Abdel Nabi, and Andrew Williams. 2011. Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Computers in Human Behavior* 27, 2 (2011), 741–749.
[65] Dean Tjosvold. 1998. Cooperative and competitive goal approach to conflict: Accomplishments and challenges. *Applied Psychology* 47, 3 (1998), 285–313.
[66] Michel François Valstar and Maja Pantic. 2006. Biologically vs. logic inspired encoding of facial actions and emotions in video. In *2006 IEEE International Conference on Multimedia and Expo*. IEEE, 325–328.
[67] Evert Van de Vliert and Carsten KW De Dreu. 1994. Optimizing performance by conflict stimulation. *International Journal of Conflict Management* (1994).
[68] Steue Whittaker. 2003. Theories and methods in mediated communication: Steve Whittaker. In *Handbook of discourse processes*. Routledge, 246–289.
[69] Yanzhe Yang, Jimei Yang, and Jessica Hodgins. 2020. Statistics-based Motion Synthesis for Social Conversations. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 201–212.
[70] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.
[71] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. 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, 4303–4309.
[72] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 736–747.
[73] Wenlin Zhuang, Jinwei Qi, Peng Zhang, Bang Zhang, and Ping Tan. 2022. Text/Speech-Driven Full-Body Animation. *arXiv preprint arXiv:2205.15573* (2022).