Multimodal modeling of expressiveness for human-machine interaction

Mireille Fares, Catherine I Pelachaud, Nicolas Obin

▶ **To cite this version:**

Mireille Fares, Catherine I Pelachaud, Nicolas Obin. Multimodal modeling of expressiveness for human-machine interaction. Workshop sur les Affects, Compagnons artificiels et Interactions, Jun 2020, Saint Pierre d'Oléron, France. hal-02933482

HAL Id: hal-02933482
https://inria.hal.science/hal-02933482

Submitted on 8 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multimodal modeling of expressiveness and alignment for human-machine interaction

Mireille Fares
ISIR and IRCAM
Sorbonne Université
Paris, France
fares@isir.upmc.com

Catherine Pelachaud
CNRS-ISIR
Sorbonne Université
Paris, France
catherine.pelachaud@upmc.fr

Nicolas Obin
CNRS-IRCAM
Sorbonne Université
Paris, France
nicolas.obin@ircam.fr

ABSTRACT

Myriad of applications involve the interaction of humans with machines, such as reception agents, home assistants, chatbots or autonomous vehicles' agents. Humans can control the virtual agents by the mean of various modalities including sound, vision, and touch. As the number of these applications increases, a key problem is the requirement of integrating all modalities, to leverage the interaction's quality, as well as the user's experience in the virtual world. In this State-of-the-Art review paper, we discuss about designing engaging virtual agents with expressive gestures and prosody. This paper is part of a work that aims to review the mechanisms that govern multimodal interaction, such as the agent's expressiveness and the adaptation of its behavior, to help remove technological barriers and develop a conversational agent capable of adapting naturally and coherently to its interlocutor.

KEYWORDS

multimodality, speech, gestures, prosody, intelligent embodied conversational agents

1 INTRODUCTION

Human-Human interaction inherently involves the communication through multiple channels. We employ several modalities, both sequentially and in parallel, to communicate in our daily life. The multimodal channels adopted in human communication are verbal and non-verbal[23]. Non-verbal are non-vocal signals that are sent by means of facial expressions, gaze, body postures, head or arm movements. The speech modality involves the speech content (words) and vocal signals such as speech prosody and acoustic signals. Both verbal and non-verbal modalities are essential to send and perceive new information. They both reflect our socio-emotional behavior which includes our psychological state and attitude. During a human-human interaction the alignment phenomena (tone of voice, speed of body movement) are signs of common understanding and engagement in the interaction [11, 31]. Communicative modalities should be taken into consideration when developing Human Computer Interaction (HCI) applications because computers are becoming more integrated in our daily life. As a matter of fact, in the past decade, many HCI applications such as personal assistants, tutoring systems, reception agents, chatbots, smartphones and home assistants are being extremely used in our daily life. The rise of such applications are leading humans to interact with virtual agents.

A key problem in the design of virtual assistants is how to maintain user's engagement [5, 6] during the interaction so that the interaction lasts long and stays fluent. The main limitations concern are on the one hand the agent's weak expressiveness, and on the other hand the agent's weak adaptation of his behavior to the behavior of the user. As a matter of fact (i.e Alexa and Google Home), rendering the interaction very short, and lacking in variety and interest. On the other hand, the agent does not adapt its behavior with respect to the interlocutor's behavior which decreases the engagement of the user in the interaction [11]. The behavior of the agent greatly impacts the user's attentive commitment, the duration of the interaction, as well as the user's understanding of the transmitted messages.

The present paper is part of a thesis work that aims to better understand and model the mechanisms that govern multimodal interaction (voice and gesture) between a human and a machine. We aim to develop an engaging conversational agent (ECA) with expressive gestures and prosody, capable of maintaining the attention of the interlocutor during the conversation, emphasizing important points, and making the interaction last longer by improving its quality. The ECA will be able to reinforce the interlocutor's engagement by adapting its behavior according to the behavior of the interlocutor. As a first step of our work, we present in this paper a State-of-the-Art review of how to design engaging virtual agents with expressive gestures and prosody. In the following sections, we discuss the different points that should be taken into account when designing an ECA capable of detecting the user's engagement level, maintaining and reinforcing it by displaying an appropriate behavior. We also discuss some of the existing models of the agent's behavior, based on different scenarios. An emphasis is placed on gestural and prosodic expressivity of the agent.

2 ENGAGING EMBODIED CONVERSATIONAL AGENTS

Engagement is "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake" as defined by Sidner et al [30]. The design of Embodied Conversational Agents (ECA) capable of rendering the users engaged in the interaction, is essential and critical for Human Agent Interaction applications [13]. Various applications such as tutoring systems [15], ambient assisting living [4, 18], and virtual museum agents [9, 10, 17, 20] convey the importance of maintaining the user's engagement during his/her interaction with the agent. User engagement is marked by nonverbal social behaviors at specific moments of the interaction: it can be feedback signals (to indicate being in phase with the interactant), a form of imitation (i.e. smiling for another smile or making the tone of the voice resemble that of the interactant), or signals synchronized

with those of the interactant (management of speaking turns). In addition to that, the behavior of the agent should adapt to the multimodal behavior of the interlocutor [11, 31]. In other words, it should be capable of adapting its behavior with respect to the behavior of the interlocutor, to strengthen the latter's engagement in interaction.

The two components underlying the engagement process are the attentional involvement and emotional involvement [25]. Short-term engagement is needed for performing a specific task during the interaction, whereas long-term engagement is essential for longer periods of interactions [4]. The development of an engaging ECA must take into account the socio emotional behavior of users which are expressed by means of verbal or non-verbal signals [13]. A socio emotional behavior includes user's social attitudes as well as their emotions. As Scherer defines it, a social attitude is "an affective style that spontaneously develops or is strategically employed in the interaction with a person" [27]. An engaging agent detects the engagement level of its interlocutor and maintains it by displaying an appropriate socio-emotional behavior during the interaction. Some studies prefer to detect the user's level of engagement by analyzing his/her signals (such as the location of his/her face), instead of his/her socio emotional behavior [7]. Other studies prefer to focus on acoustic features such as prosody, voice quality, and spectral features. Prosodic cues are also informative of the engagement level of the users, in the form of global cues (long term patterns), and local cues (short-term)[30]. Furthermore, engagement can be reflected not only in vocal signals, but also in the speech of the interlocutor [30]. User engagement cannot happen if the agent isn't expressive. In the following sections, we review the state of art of gestural expressivity and prosodic expressivity.

3 GESTURAL EXPRESSIVITY

An ECA with expressive gestures is capable of having a varied and coherent expression so that it can maintain the attention of the interlocutor. It is also capable of emphasizing important ideas, while leveraging the quality of the interaction, and lengthening its duration (to exceed one or two speaking turns). A key challenge in designing ECAs with expressive gestures, is to display the right gestures at the right time. For instance, the signification of a ring gesture can change according to other body movements as well as the utterance that are produced simultaneously [32]. In addition to that, a ring gesture can also be used as a kinesic sign, to substitute speech [32]. In this case, other body movement occurring at the same time can change its meaning. For instance, facial expressions of the person doing a ring gesture, can transform the meaning of this gesture: it can be a positive meaning such as "perfect" or "delicious", or a negative one such as "zero" or "worthlessness".

In a human-human interaction, gestures and speech are synchronized in one speaker, or even during an interaction with another speaker [19]. A little change in one gesture will occur at the same time with the beginning of change in another behavior such as a phonological segment. For instance, stressed syllables are frequently in line with gestural strokes and even eye blinks [19]. During an interaction, we tend to mimic each-other unconsciously, following the "Chamelon effect" by matching and mirroring each-other with our postures, facially, and vocally [19]. This way, humans adapt their behavior passively and unintentionally. Gestures and body movements convey emotional intentions [14]. For instance, in [12], they demonstrate that the quantity of motion of the upper body and the velocity of head are sensitive to positive emotional valence of the emotional expression. Developing human-like body language expressions in virtual agents and in robots enhance their expressiveness and improve their sociability.

Modeling agent's non-verbal behavior as well as its gestures' expressivity can be performed in multiple ways, based on different scenarios. In [8], they designed a behavioral and computational model based on an evolutionary algorithm for generation behaviors following an interruption. They used an interactive genetic algorithm since it is capable of exploring the space of solutions. In [26], they use Generative Adversarial Networks (GAN) to generate gestures using upper body sequence of joint positions which are aligned with each utterance. They show that high expressivity can be achieved with a low number of degrees of freedom. In [36], they present a learning-based co-speech gesture generation system. Since the co-speech gesture generation problem consists of mapping a sequence of words to a sequence of human poses, this problem is close to the neural machine translation that uses sequence to sequence mapping. Therefore, they chose to use a neural network architecture that is composed of an encoder (for speech-text understanding) and a decoder (to produce a sequence of gestures/frame by frame poses). The model captures speech context by using a bidirectional recurrent neural network, and the results are sent to the decoder to produce gesture motions. They have used a recurrent neural network for decoding, with a soft attention mechanism so that the decoder focuses on specific words instead of the whole utterance when producing poses. The model produces different types of gestures: iconic, metaphoric, deictic, and beat gestures.

4 PROSODIC EXPRESSIVITY

Humans' voice is the "mirror of the soul": it can reflect our emotional states and feelings. Our most personal experience can be expressed by the mean of speech, in multiple degree of variation, rendering each expression a unique act. Prosody refers to all suprasegmental aspects of speech [35]. It involves pitch, duration, amplitude, and voice quality that are used to perform lexical contrasts and convey meaning. As the articulatory functional principle states, speech is a system for transmitting communicative meanings that can be lexical, post-lexical, affective, and social ones. Nowadays, a key problem in the development of Text-to-Speech Systems (TTS) is that they only focus on the physiological (gender, age, intrinsic and co-intrinsic characteristics) speech generation as well as short-term variations of speech parameters (i.e the articulations) [24]. Therefore, TTS systems are very monotonous, and prosody is a real concern since the generated prosodic parameters have a poor variability. Thus, it is very important to consider implementing expressive TTS systems when designing virtual agents or even physical robots. TTS systems used in robotics are very monotonous, lacking prosodic expressivity.

Some of the prosodic tools that can express a speaker's affective and emotional state are non-lexical sounds. Non-lexical sounds are non-phonological sounds that are generated during the speech

or outside the talk turn. These sounds are prosodically relevant, and they can convey the speaker's emotions, intentions, attitudes and mental state [1, 28, 29]. There exist different types of non-lexical sounds such as interjections (i.e. Aie! Ah!), fillers (words used for pausing, i.e. um, okay, uh), grunts (guttural sounds made by humans/animals), and bursts (very brief, discrete, nonverbal expressions of affect in both face and voice as triggered by clearly identifiable events [21], i.e. Wow!). All these non-lexical sounds are pure prosodic tools and could be indicators of the emotional and affective state of the speaker. In [3], they show that these sounds given to a robot in a smart home create a strong "socio-affective glue" with elderly. "Glue" designates the process that is dynamically constructed during an interaction which is considered as an emerging global system whose interactants are not complete subsystems [2]. This socio-affective glue allows speakers to actively build communicative channel during an interaction. Imitation can also produce a socio-affective "glue" in children language acquisition [33]. Human-machine interactions can become as natural as human-human interactions by employing speech with prosodic expressiveness, that is, speech with emotional content [22]. The emotional content of speech can be conveyed by manipulating the voice quality as well as prosody parameters (f0, duration, and energy). Recent work suggest employing both quality voice and prosody parameters to improve the acoustic modelling of expressive speech. In the last few years, research has focused on the Harmonic plus Noise Model (HNM) to generate high quality and versality and perform speech transformation by modifying the speech prosody. For instance, humans can combine different voice pitch with different body motions to signal their desire to take the floor during interactions, and as a result, interlocutors react to their signals and therefore making the turn-taking mechanism happen [16].

In a human-human interaction, speech and movement are rhythmically coordinated in syllables and even smaller units [19]. For instance, humans can combine different voice pitch with different body motions to signal their desire to take the floor during interactions, and as a result, interlocutors react to their signals and therefore making the turn-taking mechanism happen [16].

5 DISCUSSION

The previous two sections summarized the latest work related to gestural and prosodic expressivity. Despite the great findings and conclusions that these work draw, further research efforts should be directed at the development of more expressive gestures and prosody. For instance, as we have previously discussed, in [12] they demonstrate that the quantity of head motion displayed by a pianist is sensitive to positive emotional valence of expression. The main limitation of this study is that their analysis and experiments are based on one musician (pianist) playing one instrument (the piano). In addition to that, they didn't investigate all emotional expressive movements, and analyze various modalities of expressions that musicians can do while playing (i.e. gaze, posture, facial expressions, ...). In [26], they use GANs to generate gestures using upper body sequence of joint positions which are aligned with each utterance. They generate gesture movements without taking into account the semantic meaning of the spoken text, nor putting emphasis on important words. In this case, multimodal and soft attention mechanisms [34] could be used to allow the model to focus on specific parts of the text, instead of looking at the whole utterance. This can be developed under an encoder/decoder framework, to allow finding semantic and syntactic alignments. Additionally, visual attention models could be used to pay attention to specific parts of the upper body region and reduce the amount of the information to process. Moreover, in [36], they present a learning-based co-speech gesture generation system. This system makes excessive gestures and participants in the experiments noted that motions were "jerky", "fast" and "jump around from motion to motion". Moreover, participants noted that gestures were faster than speech which makes the interaction looks unnatural. In addition to that, the system lacks the personalization of robots' gestures: all robots make the same gestures when exposed to one speech context. There are no parameters to control the robots' expressiveness. Furthermore, prosody is not taken into consideration when generating co-speech gestures: the system lacks prosodic expressiveness.

6 FUTURE WORK

For future work, we plan to model the gestural and prosodic expressiveness, while taking into consideration all the points discussed in the previous section. Emphasis will be placed on two essential aspects of the modelling. On one hand, we will focus on the development of architectures that are structured on several time scales to improve the modeling of prosodic and gestural variability at the sentence level as well as the whole speech. On the other hand, we will focus on the learning of a coherent multimodal behavior by using multimodal attention mechanisms applied to the synchronicity of the generated prosody and gestures. Furthermore, we will work on making the agent's behavior coherently aligned with that of its interlocutor by using interactive and imitation learning.

ACKNOWLEDGMENTS

We would like to thank Sorbonne Center for Artificial Intelligence (SCAI) for funding this thesis.

REFERENCES

[1] Felix Ameka. 1992. Interjections: The universal yet neglected part of speech. *Journal of pragmatics* 18, 2-3 (1992), 101–118.
[2] Véronique Aubergé. [n.d.]. La glu socio-affective: enjeux et risques du robot «compagnon». *des JA-SFTAG 2014* ([n. d.]), 13.
[3] Véronique Aubergé, Yuko Sasa, Nicolas Bonnefond, Brigitte Meillon, Tim Robert, Jonathan Rey-Gorrez, Adrien Schwartz, Leandra Batista Antunes, Gilles De Biasi, Sybille Caffiau, et al. 2014. The eee corpus: socio-affective" glue" cues in elderly-robot interactions in a smart home with the emoz platform.
[4] Timothy Bickmore and Toni Giorgino. 2006. Health dialog systems for patients and consumers. *Journal of biomedical informatics* 39, 5 (October 2006), 556—571. https://doi.org/10.1016/j.jbi.2005.12.004
[5] Timothy Bickmore, Daniel Schulman, and Langxuan Yin. 2010. Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence* 24, 6 (2010), 648–666.
[6] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and Maintaining Long-Term Human-Computer Relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (June 2005), 293–327. https://doi.org/10.1145/1067860.1067867
[7] Dan Bohus and Eric Horvitz. 2014. Managing Human-Robot Engagement with Forecasts and... Um... Hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction* (Istanbul, Turkey) *(ICMI '14)*. Association for Computing Machinery, New York, NY, USA, 2–9. https://doi.org/10.1145/2663204.2663241
[8] Angelo Cafaro, Brian Ravenet, and Catherine Pelachaud. 2019. Exploiting evolutionary algorithms to model nonverbal reactions to conversational interruptions in user-agent interactions. *IEEE Transactions on Affective Computing* (2019).

[9] Angelo Cafaro, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2016. First Impressions in Human–Agent Virtual Encounters. *ACM Trans. Comput.-Hum. Interact.* 23, 4, Article Article 24 (Aug. 2016), 40 pages. https://doi.org/10.1145/2940325

[10] Sabrina Campano, Chloé Clavel, and Catherine Pelachaud. 2015. "I like this painting too": When an ECA Shares Appreciations to Engage Users. In *AAMAS*.

[11] Ginevra Castellano, Maurizio Mancini, Christopher Peters, and Peter W McOwan. 2011. Expressive copying behavior for social agents: A perceptual analysis. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 42, 3 (2011), 776–783.

[12] Ginevra Castellano, Marcello Mortillaro, Antonio Camurri, Gualtiero Volpe, and Klaus Scherer. 2008. Automated analysis of body movement in emotionally expressive piano performances. *Music Perception: An Interdisciplinary Journal* 26, 2 (2008), 103–119.

[13] Chloé Clavel, Angelo Cafaro, Sabrina Campano, and Catherine Pelachaud. 2016. *Fostering User Engagement in Face-to-Face Human-Agent Interactions: A Survey*. Springer International Publishing, Cham, 93–120. https://doi.org/10.1007/978-3-319-31053-4_7

[14] Sofia Dahl and Anders Friberg. 2007. Visual perception of expressiveness in musicians' body movements. *Music Perception: An Interdisciplinary Journal* 24, 5 (2007), 433–454.

[15] Sidney K. D'Mello and Arthur C. Graesser. 2012. AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *TiiS* 2 (2012), 23:1–23:39.

[16] Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology* 23, 2 (1972), 283.

[17] R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A.c. Schultz, and et al. 2005. Designing robots for long-term social interaction. *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2005). https://doi.org/10.1109/iros.2005.1545303

[18] David Griol, José Manuel Molina, and Zoraida Callejas. 2014. Modeling the User State for Context-Aware Spoken Interaction in Ambient Assisted Living. *Applied Intelligence* 40, 4 (June 2014), 749–771. https://doi.org/10.1007/s10489-013-0503-z

[19] Mark L Knapp, Judith A Hall, and Terrence G Horgan. 2013. *Nonverbal communication in human interaction*. Cengage Learning.

[20] Stefan Kopp, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. 2005. A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application. In *Intelligent Virtual Agents*, Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 329–343.

[21] Eva G Krumhuber and Klaus R Scherer. 2011. Affect bursts: dynamic patterns of facial expression. *Emotion* 11, 4 (2011), 825.

[22] Carlos Monzo, Ignasi Iriondo, and Joan Claudi Socoró. 2014. Voice quality modelling for expressive speech synthesis. *The Scientific World Journal* 2014 (2014).

[23] Sigrid Norris. 2004. *Analyzing multimodal interaction: A methodological framework*. Routledge.

[24] Nicolas Obin. 2011. *MeLos: Analysis and modelling of speech prosody and speaking style*. Ph.D. Dissertation.

[25] Christopher Peters, Ginevra Castellano, and Sara de Freitas. 2009. An exploration of user engagement in HCI. In *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*. 1–3.

[26] Igor Rodriguez, José María Martínez-Otzeta, Itziar Irigoien, and Elena Lazkano. 2019. Spontaneous talking gestures using generative adversarial networks. *Robotics and Autonomous Systems* 114 (2019), 57–65.

[27] Klaus R Scherer. 2005. What are emotions? And how can they be measured? *Social science information* 44, 4 (2005), 695–729.

[28] Marc Schröder. 2003. Experimental study of affect bursts. *Speech communication* 40, 1-2 (2003), 99–116.

[29] Marc Schröder et al. 2006. Perception of non-verbal emotional listener feedback. In *Proc. Speech Prosody 2006*. Citeseer.

[30] CL Sidner. 2004. Where to look: A study of human-robot interaction. In *Proc. International Conference on Intelligent User Interfaces (ACM IUI 2004)*. 78–84.

[31] Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. 2004. Where to Look: A Study of Human-Robot Engagement. In *Proceedings of the 9th International Conference on Intelligent User Interfaces* (Funchal, Madeira, Portugal) *(IUI '04)*. Association for Computing Machinery, New York, NY, USA, 78–84. https://doi.org/10.1145/964442.964458

[32] Jürgen Streeck. 2013. Elements of Meaning in Gesture, Geneviève Calbris, John Benjamins Publishing Company (2011), pp. 378+ VIII. Price: EUR 95.00| USD 143.00, ISBN: 978-90-272-2847-5.

[33] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences* 28, 5 (2005), 675–691.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[35] Yi Xu. 2019. Prosody, tone and intonation. *The Routledge handbook of phonetics* (2019), 314–356.

[36] 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.