Multimodal modeling of expressiveness for human-machine interaction
Mireille Fares, Catherine Pelachaud, Nicolas Obin

To cite this version:
Mireille Fares, Catherine Pelachaud, Nicolas Obin. Multimodal modeling of expressiveness for human-machine interaction. WACAI 2020, Jun 2021, Île d’Oléron, France. hal-02928055

HAL Id: hal-02928055
https://hal-cnrs.archives-ouvertes.fr/hal-02928055
Submitted on 2 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 for human-machine interaction

Mireille Fares
ISIR and STMS Lab
Ircam, CNRS, Sorbonne Université
Paris, France
fares@isi.upmc.com

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

Nicolas Obin
STMS Lab
Ircam, CNRS, 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[25]. 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 [13, 33]. 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 [6, 7] 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 the agent’s behavior is very monomodal and monotonous, 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 [13]. 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 humans and Embodied Conversational Agents (ECA). 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
People’s satisfaction and well-being result from their environment, the decision latitude available to them, as well as the social support that exist in their environment [9]. For instance, a good working life of employees consists of various activities in dynamic environments, the support of co-workers and management, as well as the decisions they are allowed to make. In the context of Human-Computer Interaction, the presence of repetitiveness and routine within a human-machine interaction, and the absence of unpredictability and variation will probably imply a passive and boring interaction.

Personalising interaction is crucial to obtain engagement as well as long-term relationship development. For instance, in [1] they propose a toy companion for children capable of personalizing the interaction based on personal knowledge about the child, in order
to build a long-term relationship, and render the interaction with the child more engaging. Interactions that are not personalized, and that do not contain the experience of control will result in a stressful experience [9].

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 [32]. 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 [15]. Various applications such as tutoring systems [17], ambient assisting living [5, 20], and virtual museum agents [11, 12, 19, 22] 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 [13, 33]. 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 [27]. Short-term engagement is needed for performing a specific task during the interaction, whereas long-term engagement is essential for longer periods of interactions [5]. 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 [15]. 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” [29]. 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 [8]. 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) [32]. Furthermore, engagement can be reflected not only in vocal signals, but also in the speech of the interlocutor [32]. 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 [34]. In addition to that, a ring gesture can also be used as a kinesic sign, to substitute speech [34]. 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 [21]. 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 [21]. During an interaction, we tend to mimic each-other unconsciously, following the “Chameleon effect” by matching and mirroring each-other with our postures, facially, and vocally [21]. This way, humans adapt their behavior passively and unintentionally. Gestures and body movements convey emotional intentions [16]. For instance, in [14], 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 [10], 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 [28], 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 [37], 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 expression 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 [36]. 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) [26]. 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 [2, 30, 31]. 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 [23], 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 [4], 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 [3]. 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 [35]. Human-machine interactions can become as natural as human-human interactions by employing speech with prosodic expressiveness, that is, speech with emotional content [24]. 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 the voice quality as well as prosody parameters (f0, duration, and emotional content of speech can be conveyed by manipulating the mean of speech, in multiple degree of variation, rendering each expression a unique act. Prosody refers to all suprasegmental aspects of speech [36]. 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) [26]. 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 [2, 30, 31]. 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 [23], 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 [4], 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 [3]. 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 [35]. Human-machine interactions can become as natural as human-human interactions by employing speech with prosodic expressiveness, that is, speech with emotional content [24]. 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

5 LIMITATIONS
The previous two sections summarized the latest works related to gestural and prosodic expressivity. Despite the great findings and conclusions that these works draw, further research efforts should be directed at the development of more expressive gestures and prosody. For instance, as we have previously discussed, in [14] they demonstrate that the quantity of head motion displayed by a pianist is sensitive to positive emotional valence of expression, but their analysis is 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 [28], they use GANs to generate gestures using upper body sequence of joint positions which are aligned with each utterance. The limitation of this study is that they generate gesture movements without taking into account the semantic meaning of the spoken text, nor considering emphasis on important words. Moreover, in [37], they present a learning-based co-speech gesture generation system. This system makes repetitive and excessive gestures: the participants in the experiments noted that motions were ‘jerky’, ‘fast’ and ‘jump around from motion to motion’. In addition to that, participants noted that gestures were faster than speech which makes the interaction looks unnatural. The system also lacks the personalization of robots’ gestures: all robots make the same gestures when exposed to one speech context. There is no parameter to control the robots’ expressiveness. Furthermore, prosody is not taken into consideration when generating co-speech gestures: the system lacks prosodic expressiveness.

6 OUR CONTRIBUTION
Our work consists in developing a multimodal model of expressiveness for human-agent interaction. We plan to model the gestural and prosodic expressiveness, while taking into consideration all the points discussed in the previous section. More specifically, we plan to develop a multimodal model to predict facial expressions based on previous head motion, facial expressions, text, and prosody. We will build a multimodal model and it will be implemented using a Sequence to Sequence encoder/decoder framework. This will also allow finding semantic and syntactic alignments. As mentioned in sections 4 and 5, human-agent interactions may become as natural as human-human interactions by employing speech with prosodic and gestural expressiveness. This is the reason why our model will do the prediction based on voice (voice prosody and voice quality) and gestural (facial expressions and head motions) input features. The input speech features that we will use in our model consist of prosodic features, voice quality features, as well as Mel Frequency features. The prosodic features that we will consider in our model are the energy, the fundamental frequency F0, as well F0 contour. As [24] suggests, the voice quality features that can be useful for expressiveness, and which we will use for our model are: Jitter, Shimmer, Harmonic-to-Noise Ratio (HNR), and the hammarberg
index (Hamml). The gestural input features will be the action units of facial expressions as well as head motions. In addition to that, 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. In later work, we will focus on rendering the agent’s behavior coherently aligned with that of its interlocutor by using interactive and imitation learning.

ACKNOWLEDGMENTS
This work was performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d’Avenir programme under reference ANR-11-IDEX-0004-02.

REFERENCES