Exploiting evolutionary algorithms to model nonverbal reactions to conversational interruptions in user-agent interactions

Angelo Cafaro, Brian Ravenet, and Catherine Pelachaud

Abstract—In social interactions between humans and Embodied Conversational Agents (ECAs) conversational interruptions may occur. ECAs should be prepared to detect, manage and react to such interruptions in order to keep the interaction smooth, natural and believable. In this paper, we examined nonverbal reactions exhibited by an interruptee during conversational interruptions and we propose a novel technique driven by an evolutionary algorithm to build a computational model for ECAs to manage user's interruptions. We propose a taxonomy of conversational interruptions adapted from social psychology, an annotation schema for semi-automatic detection of user's interruptions and a corpus-based observational analysis of human nonverbal reactions to interruptions. Then we present a methodology for building an ECA behavioral model including the design and realization of an interactive study driven by an evolutionary algorithm, where participants interactively built the most appropriate set of multimodal reactive behaviours for an ECA to display interpersonal attitudes (friendly/hostile) through nonverbal reactions to a conversational interruption.

Index Terms—Nonverbal behaviour, Conversational interruptions, Turn-taking, Interactive Evolutionary algorithms, Embodied Conversational Agents.

1 Introduction

In conversations and social interaction, humans often produce and decipher subtle signals (verbal and nonverbal) to support and ease the interaction. According to Duncan [1], people can use a different voice pitch and body motions to signal their desire to take the floor, and interlocutors quickly react to these signals (either by yielding the floor or competing for it), therefore realizing the turn-taking mechanism of the interaction [2]. This mechanism can be influenced by the social context in which the conversation takes place or by the social relationship (i.e. interpersonal attitude) existing between the participants. A family dinner, for instance, is a context in which overlaps and interruptions are more easily accepted compared to a more formal interaction context [3]. Moreover, interruptions can indicate disagreement but two close friends sharing a positive friendly interpersonal attitude might also feel comfortable and more engaged in the interaction when talking at the same time [4].

An Embodied Conversational Agent (ECA) is a virtual or robotic human-like character that demonstrates many of the same properties as humans in face-to-face conversation, including the ability to produce and respond to verbal and nonverbal communication [5]. In the context of interactions between humans and ECAs, interruptions may also occur frequently, and an ECA needs to be able to handle user's interruptions and react with appropriate verbal and nonverbal behavior when such "unexpected event" (for the agent) occurs in order to keep the conversation smooth and believable for the user.

Existing literature in human-human interaction and social psychology does not focus on the behaviours that people exhibit in case of unexpected conversational interruptions by the interlocutor. Researchers mainly focused on the impact of an interruption on the interruptee in terms of perceived social attitude of the interrupter [6], gender effects [7] and semantic meaning (e.g. overlaps in conversation) [8]. Some researchers in the field of ECAs looked at generating agent's verbal content when handling barge-in user's interruptions [9] or adaptive interruptible speech synthesis [10], [11]. However, none of them looked in detail to the specific nonverbal behaviour performed when a conversational interruption occurs.

In the context of user-agent social interactions, we believe that at least the following aspects should be considered when it comes to manage conversational interruptions: detecting when a user's interruption occurs, understanding when it is more appropriate for an agent to interrupt the user and the effects of an interruption on qualities of the social interaction (e.g. agent's interpersonal attitude towards the user and engagement in the interaction), and also reacting appropriately (i.e. agent's behaviours and mental state) to a user's interruption.

More specifically, in terms of effects of interruptions and impacts of different agent's reactions to a user's interruption, in this work we focus on the link between conversational interruptions and perception of social attitudes. The assessment and proper management of social attitudes in dyadic interactions represents a fundamental aspect for building affective and socially-believable ECAs. We adopted Scherer's definition of interpersonal stance to define what a social attitude is [12], "an affective style that spontaneously develops or is strategically employed in the interaction

- *Brian Ravenet is with the CNRS-LIMSI, University of Paris Saclay, Orsay, France.*
- *Angelo Cafaro and Catherine Pelachaud are with the CNRS-ISIR, Sorbonne University, Paris, France.*

with a person or a group of persons". These attitudes are often represented by two dimensions: friendliness (on an axis ranging from friendly to hostile) and dominance (ranging from submissive to dominant) [13]. We started by focusing on the dimension of friendliness by investigating the perception of friendly and hostile reactions to user's barge-in interruptions. However, our goal is to model ECAs capable of managing and reacting to user's interruptions in a natural manner. Therefore, we examined multimodal reactive behaviours that humans display during interaction as reaction to a conversational interruption, and we adopted those observations to develop a novel methodology that exploits evolutionary algorithms to build a behavioral model for ECAs to manage conversational interruptions.

Common approaches for generating ECAs' behavior that have emerged in the past years can be mainly categorized in **rule-based** vs. **data-driven**. A rule-based approach often consists of deriving rules from human-data observations and social psychology theories for guiding the generation process [14], [15], [16]. Conversely, data-driven approaches are often based on annotated corpora. *Crowdsourcing*, for instance, has been recently used to obtain annotated datasets [17], [18], [19] and delegate to human users the creation of an ECA's behavioral model through a user-friendly tool whose interface allows users to directly configure an ECAs nonverbal behaviors for conveying particular socio-emotional states (e.g. friendliness) [18]. This particular approach allows the participants to design the ECA behaviours based on their perception of the agent, however the participants need to manipulate several parameters for defining the agent's behaviours. This can be a tedious and complex task with an increasing number of parameters and values. Leveraging from this approach, we propose a novel methodology that exploits interactive evolutionary algorithms. The technique, being interactive, still requires users involvement and benefits from direct perception of the ECA, but it also simplifies the participant's task because a genetic algorithm is capable of exploring the space of solutions (i.e. configuration of ECA behaviors in our case) while users only need to observe the behavior performed by the ECA and steer the overall generation process.

In Section 2, we describe the meaning and effects of conversational interruptions, first presenting a taxonomy of interruptions useful to understand and model them in a user-agent interaction (see Section 2.1) and then (in Section 2.2) by quickly reviewing the results of a user study aimed at evaluating interruptions' effects in terms of interruptee and interrupter perceived interpersonal attitudes. The proposed taxonomy and results from this study have been used to develop the automatic detection of user's interruptions as described in Section 3. As described in Section 4, we examine multimodal reactions to different types of conversational interruptions in human-human interactions and, in Section 5, we present an interactive study driven by an evolutionary algorithm where participants interactively built the most appropriate set of multimodal reactive behaviours for an ECA to display friendly/hostile nonverbal reactions to a conversational interruption.

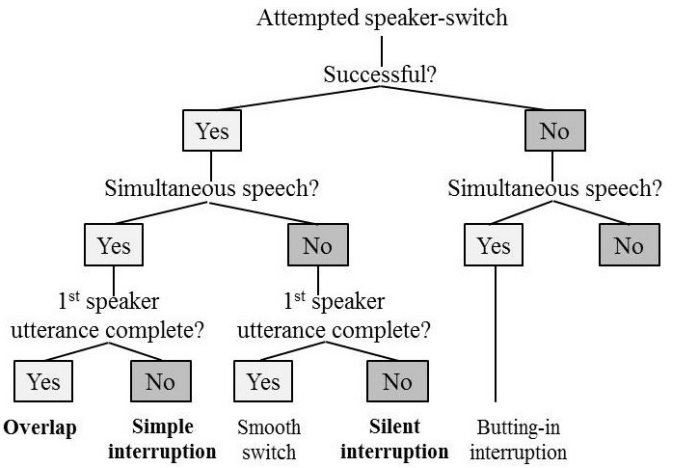


Fig. 1. Adaptation of Beattie's taxonomy of interruption types

2 Interruptions, their meaning and effects

2.1 A Taxonomy of Conversational Interruptions

Interruptions have long been associated with interpersonal dominance [20]. However, researchers have since expanded their views on this subject. Noticing that an interruption can either bring new content or complement the current content of the conversation, researchers have proposed to distinguish *cooperative* interruptions from *disruptive* interruptions based on the content of the intervention [21], [22]. In addition, interruptions can also be differentiated based on their roles in the turn taking system of the conversation, for instance if they prevented the completeness of the utterances or not [7]. During human conversations it is common for participants to quickly exchange turns, sometimes speaking at the same time. It is therefore necessary to use a proper definition of interruptions. According to Shegloff, an interruption can be described as an intervention by a participant while other participants are holding the floor of the conversation, thus leading to not letting them finish [23]. Also from Shegloff, an overlap is defined by the situation of more than one participant talking at the same time [23]. Interestingly, an overlap does not necessarily implies an interruption and vice-versa.

As interruptions can vary in content and role in the conversation, in order to distinguish the different types of interruptions, in our work we used an adapted version of the taxonomy of attempted speaker-switches proposed by Beattie in [7]. Figure 1 depicts this taxonomy, where we highlighted in boldface the types of interruption being studied in this work. Smooth switches are by definition not considered an interruption and have been studied already in [24], [25], whereas the added complexity of handling butting-in interruptions is out of the scope of this work.

The basic criterion for this classification relies on (un)successful speaker-switches, meaning that the interrupter successfully (or not) takes the floor. This depends on the presence of simultaneous speech and first speaker's utterance completeness. However, back-channels (e.g "yeah", "hmm", "exactly") do not fall into Beattie's definition of interruptions [7].

From previous findings in human social psychology and linguistics, we identified two general categories of interruptions, usually referred to as disruptive and cooperative [21], [26] that we defined as "interruption strategies". According to [21] "cooperative interruptions are intended to help the speaker by coordinating the process and/or content of the ongoing conversation" [26], whereas disruptive interruptions pose threats to the current speaker's territory by disrupting the process and/or content of the ongoing conversation [4], [21], [26].

2.2 A perceptive study on Effects of Interrupting Behaviour

In addition to serve as turn-taking mechanism, interruptions may lead to different perceptions of both the interruptee and interrupter's interpersonal attitude, engagement and involvement in the interaction. Following the proposed taxonomy, in an empirical study we investigated whether different interruption types and strategies (disruptive vs. cooperative) in agent-agent interactions had impacts on user's perceived interpersonal attitude of the agents, as well as their engagement and involvement in the interaction, more details can be found in [27]. Results suggested that the amount of overlaps in reaction to an interruption had significant impact on the users' perception of interpersonal attitudes of both agents in particular with respect to dominance, however changing from a disruptive to a cooperative strategy of interruption increased the interrupter's perceived friendliness and reduced its perceived dominance. Moreover, the interruption strategy had a main effect on the user's perception of engagement and involvement of the interrupter, being a cooperative interrupter was considered as being more engaged and involved in the interaction. The detection of user's interruptions has been built on the proposed taxonomy and these results as described in the following section.

3 User's interruptions detection

Detecting when the user interrupts the agent during its speaking-turn is important to keep the interaction believable and allows the agent to react appropriately and update its mental state's representation of the conversation (e.g. current dialogue turn). It is important to differentiate back-channelling behaviour (i.e. when the user gives feedback while listening to the agent without intent to take the speaking floor) and interrupting behaviour as categorized in our taxonomy (i.e. disrupt vs. cooperate).

In addition to modify (e.g. re-plan) the current dialogue state, the agent in its mental state can build a perception model of the users communicative intention by knowing whether s/he aims at disrupting or cooperating in the interaction, thus providing information about the user's potential level of engagement in the interaction. Furthermore, as focusing in this work, a proper verbal and non verbal reaction should be planned and performed in real-time by the agent both in terms of speech synthesis [11] and exhibited behaviour as described in Sections 4 and 5.

We approached this problem by taking advantage of the data collected for the NoXi database (described in [28]) and envisioning a corpus-based machine learning approach. The idea was to learn from annotated dyadic interactions whether the speaker's acoustic features (e.g. prosody) allow us to discern if the user is interrupting (as opposed to back-channelling) and in this case which is the employed strategy (disruptive or cooperative). We needed to manually annotate the interruption strategies in our data because the strategy is strictly related to the semantic content of the speaker's turn and we were not aware of any existing tool to automate this process. We developed, however, a plugin for the Social Signal Interpretation SSI Framework [29], to speed up this job and obtain automatically from speakers' detected voice activity the annotation of their communicative state (speaking vs. silence) and turn transitions (pause and overlap within and between turns) as defined in [30] and explained in more details as follows.

3.1 Schema for manual annotations of interruptions

In the NoXi database, an Expert and a Novice have been recorded during a screen-mediated interaction on a chosen topic of expertise, for the expert, and interest for the novice. We refer to the **Expert** (**E**) and the **Novice** (**N**) as the two interlocutors in remainder of the paper. Starting from voice activity detection (speech vs. silence), in our annotation schema we indicate the participant who's speaking defined as communicative state layer of the conversation, the temporal relations between the interlocutors' speech defined as transitions layer and the strategy of interruption as well as their temporal characteristics defined as interruptions layer.

Communicative State: this layer describes both interlocutors speaking activity during the conversation, thus resulting in four states:

1) **NONE:** no one speaks;
2) **EXPERT:** E speaks;
3) **NOVICE** N speaks;
4) **BOTH:** both speak.

Transitions: this layer represents a transition event in the conversation that can go from speech to silence and vice-versa for the same speaker (within turn) or between the two speakers (between turns):

1) **Pause within (PAUSE.W):** a (long) silence within a speaking turn of E (N) without speaker switch;
2) **Pause between (PAUSE.B):** a speaker switch from E to N (or vice-versa) with a silence in between;
3) **Perfect:** a speaker change without a silence nor an overlap in between;
4) **Overlap within (OVERLAP.W):** an overlap without speaker switch;
5) **Overlap between (OVERLAP.B):** an overlap with a speaker change;

Interlocutors: when two interlocutors are interacting, an interruption entails an interruptee and an interrupter:

1) **Interruptee:** the interlocutor (speaker) being interrupted.
2) **Interrupter:** the interlocutor (addressee) that interrupts the current speaker.

Interruptions: this layer represents either an interruption of a given **strategy** (i.e cooperative vs. disruptive) or a back-channel:

1) **COOPERATIVE, DISRUPTIVE:** these are interruptions that are distinguished based on the semantic content of the utterance.
2) **BACKCHANNEL:** E (N) provides verbal backchannel (n.b. this is not considered as an interruption in our framework).

In the interruptions layer, we also annotate the interruptee **reaction** to an interruption:

1) **HALT:** the interruptee (i.e. current speaker) halts and yields the turn to the interrupter.
2) **OVERLAP:** while the interrupter grabs the speaking turn, the interruptee gives the turn in the following manner by:
    a) **EXTRASHORT:** stopping to speak during current word;
    b) **SHORT:** arriving until the end of the current word;
    c) **MEDIUM:** arriving until the next word planned within the utterance;
    d) **LONG:** reaching the end of the utterance.
3) **RE-PLAN:** being interrupted and halting the current utterance but immediately responding with a new one, thus keeping the turn.

Following our annotation schema, Figure 2 shows an example of possible labeling of Communicative States (2), Transitions (3) and Interruptions (4) starting from speakers voice activity (1). The communicative state and transitions layers were automatically annotated with our SSI's plugin. The Interruptions layer (4) was manually annotated, thus including both the **strategy** of interruption (cooperative vs. disruptive) concerning the interrupter and the **reaction** concerning the interruptee. However, layers (2) and (3) dramatically reduced the amount of work needed, thus allowing us to focus solely on segments automatically annotated as overlaps and pauses between turns, as they were major representatives of conversational interruptions in the scope of our work.

3.2 Semi-automatic annotation

The semi-automatic annotations of the Communicative States and Transitions layers were supported by a plugin developed within SSI [29]. We first obtained the voice activity of each interlocutor from the recorded speech. These data is then transformed in conversation states markers (i.e. speaker 1 speaks, speaker 2 speaks, no voice activity, both voice are active). Once a conversation state is marked, a Finite State Machine processes it, as depicted in Figure 3, and triggers the appropriate Annotation Event, which is a pair indicating the layer and corresponding label from our annotation schema. From any given state in our FSM, the Output is shown in red, whereas an event capable of triggering a state transition is shown in blue.

The mapping from FSM Outputs to Annotation Events is summarized in Table 1. We abbreviated the FSM Outputs and adopted those abbreviations in both Figure 3 and Table 1 as follows: S stands for Silence, O for Overlap, S1 for Speaker 1, S2 for Speaker 2, S1-S2 for a turn switch from Speaker 1 to Speaker 2, S1-S-S2 for Speaker 1 transition to Silence and then to Speaker 2, and following these we abbreviated all remaining outputs.

After this automatic annotation process supported by SSI and our FSM, we manually annotated the strategy within the Interruptions layer (cooperative vs. disruptive) with the support of speech's semantic meaning.

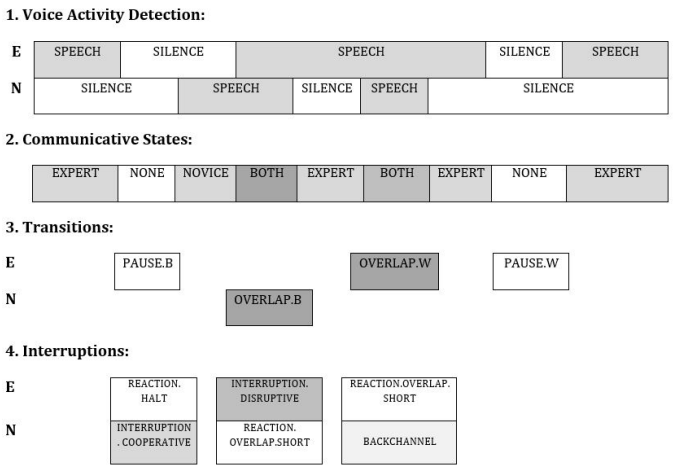


Fig. 2. Example for labeling of Communicative States (2), Transitions (3) and Interruptions (4) starting from speakers voice activity (1).

FSM Output	Annotation Event Layer	Label
UPDATE-END	-	-
S	Communicative State	NONE
O	Communicative State	BOTH
S1	Communicative State	NOVICE
S2	Communicative State	EXPERT
S1-S2	Transitions.Novice	PERFECT
S2-S1	Transitions.Expert	PERFECT
S1-O-S1 O-S1	Transitions.Novice	OVERLAP.W
S2-O-S2 O-S2	Transitions.Expert	OVERLAP.W
S1-O-S2	Transitions.Novice	OVERLAP.B
S2-O-S1	Transitions.Expert	OVERLAP.B
S1-S-S2 S1-S-O	Transitions.Novice	PAUSE.B
S2-S-S1 S2-S-O	Transitions.Expert	PAUSE.B
S1-S-S1	Transitions.Novice	PAUSE.W
S2-S-S2	Transitions.Expert	PAUSE.W

TABLE 1
Mapping from our finite state machine Output and Annotation Event, i.e. layer and a label of our annotation schema for interruptions detection.

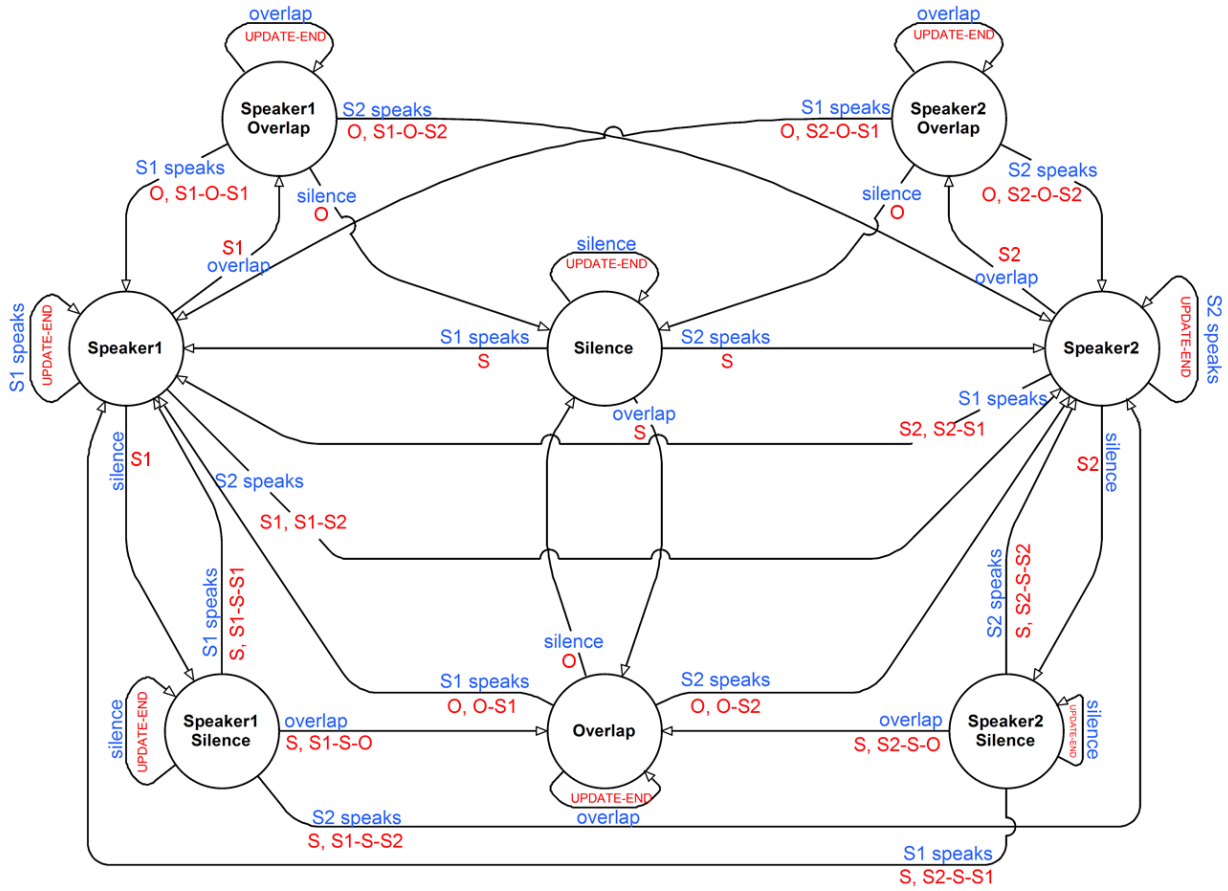


Fig. 3. The Finite State Machine that generates Annotation Events corresponding to Conversation States and Transitions in our annotation schema.

Annotation	Layer	Labels
Automatic	Communicative State	NONE, NOVICE, EXPERT, BOTH
Automatic	Transitions (E & N)	PAUSE.B, PAUSE.W, PERFECT, OVERLAP.B, OVERLAP.W
Manual	Interruptions (E & N)	INTERRUPTION.COOPERATIVE INTERRUPTION.DISRUPTIVE BACKCHANNEL REACTION.HALT REACTION.OVERLAP.EXTRASHORT REACTION.OVERLAP.SHORT REACTION.OVERLAP.MEDIUM REACTION.OVERLAP.LONG REACTION.REPLAN

TABLE 2
Our annotation schema including the main layers and labelling adopted in each one.

4 Observations of Multimodal Reactions to Interruptions

We exploited the data available in NoXi and used our schema to semi-automatically annotate and further analyze dyadic human-human interactions in order to discover relationships between an interlocutor's interruption and another interlocutor's reaction in terms of multimodal nonverbal behaviour. The outcome of these observations was then used to design the genomes of our evolutionary approach described in Section 5.

In this section, we describe the work that we conducted to identify the multimodal reactive behaviours that humans display in a dyadic interaction when reacting to conversational interruptions. We examined 10 French language sessions in NoXi of an average duration of about 30 minutes each. Then we looked at specific occurrences of interruptions as annotated using the methodology described in Section 3.

We carefully observed all the relevant segments automatically marked as interruptions for both expert and novice for a total of 96 minutes for each interlocutor in the selected observations. According to our interruption classification schema (c.f. Table 2), we distinguished 3 types of reactions to an interruption, namely *HALT*, *OVERLAP* and *REPLAN*. For

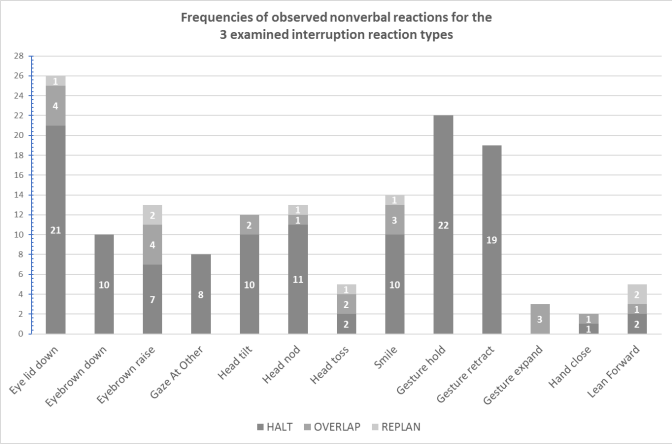


Fig. 4. Frequencies of observed nonverbal reactions for the 3 types of interruption reactions: HALT, OVERLAP and REPLAN.

each of these possible reactions, we observed and noted the instances throughout the sessions of a variety of nonverbal behaviours exhibited by the interruptee when the conversational interruptions occurred. The body parts involved in the reactions that we mainly observed were: eye lids movements (opening or closing), gaze behavior (looking away or to the other interlocutor), eyebrows (eyebrows raising or lowering), head movements (tilts, nods and head toss), smile, gestures and torso movements. More specifically, while reacting to an interruption, interlocutors tended to close/widen the eye lids, raised/lowered the eyebrows, gazed at the interrupter, tilted, nodded or tossed their head, smiled, held, retracted or expanded their current gesture for a variable amount of time, and leaned forward their torso.

These nonverbal behaviours were exhibited either in concert with others or occurred in isolation. The chart in Fig. 4 displays the frequencies of the occurrences for nonverbal reaction to an interruption and for the three different types of reactions every time a speaker (interruptee) was interrupted by the interlocutor.

In case of HALTs, we observed that the interruptee had two distinct gesture behaviours that we named GESTURE HOLD and RETRACT. These correspond to two phases, during the reaction, in which they held their current gesture and, sometimes, they retracted towards a rest pose with hands down by performing another pause (i.e. gesture freeze) of a certain duration that we called GESTURE RETRACT. We did not report exact timings for all these observations because our primary interest was to observe at a first glance the behaviours that emerged during such reactions and not a precise annotation of all aspects of behaviours. In the OVERLAP case, some participants reacted by expanding the current gesture, meaning that while overlapping their speech they enlarged the movement of their hands, thus resulting in a spatial expansion of their current gesture.

These observations, as well as the annotations and insights described in previous sections on the effects of conversational interruptions on perceived interpersonal attitudes, were the inputs to conceive an interactive study driven by an evolutionary algorithm aimed at designing the behaviour model of an ECA when reacting to a conversational interruption.

5 Building the Agent Behavioural Model

In this section we introduce a novel approach to design an ECA behavioral model inspired by evolutionary algorithms, in particular, we present an interactive genetic algorithm.

5.1 An Interactive Genetic Algorithm

A genetic algorithm is an heuristic search method used for finding optimized solutions to search problems based on the theory of natural selection and evolutionary biology [31]. An algorithm starts by generating a random population of solutions (n.b. each solution can be represented as an array of properties or features named genome and each feature is a gene). Then, by implementing bio-inspired operators, such as mutation and crossover, these solutions are selected based on a quality evaluator, named fitness function, in order to generate better ones. The retained solutions are used to breed new solutions by combining and mutating some of their properties using the bio-inspired operators. The algorithm finishes once an optimal solution is found or when the maximum number of iterations is reached. The selection of an optimal solution can be done through the evaluation of any given genome against an objective fitness function, thus incrementally selecting better solutions compared to previously explored/generated ones. Alternatively, the selection can be interactively made by a human participant that becomes integral part of the process.

Our idea was to design the agent's behavioral model through such an interactive evolutionary approach involving users for choosing optimal solutions. In particular, we conceived an interactive study driven by a genetic algorithm and considered the nonverbal behaviour exhibited by an agent in reaction to a conversational interruption as a possible solution of our genetic algorithm. Our ultimate goal was to generate and breed better solutions, in terms of expressing interpersonal attitudes (i.e. friendliness/hostility) until an optimal configuration is obtained. The algorithm, being interactive, required the users to judge the quality of the solution and therefore establish whether a solution (i.e. set of nonverbal signals displayed as reaction to an interruption) is optimal provided a given judgment criterion (i.e. expression of a friendly/hostile reaction).

The representation our behavioural model as an interactive genetic algorithm problem is described in the remainder of this section, whereas the design and implementation of our user study driven by such genetic algorithm is described in Section 5.2.

5.1.1 The genomes

In genetic algorithms, a genome is the sequence of genes that is evolving at each iteration and it represents a solution. Therefore, the optimal solution is a genome with a specific configuration (i.e. values) of its genes. In our mapping, a genome is a nonverbal reaction to a conversational interruption and its genes are the specific nonverbal behavior forming such reaction. From the analysis and observations of dyadic interactions in the NoXi database described in Section 4, we identified a total of 16 genes (i.e. nonverbal signals) for our genome (i.e. nonverbal reaction to an interruption) as summarized in Table 3.

Gene	Feature	Range	Step
Head Tilt	Intensity	[0 .. 1]	0.25
	Duration	[0.3 .. 1.3]	0.1
Nod/Toss	Intensity	[-1 .. 1]	0.25
	Duration	[0.3 .. 1.3]	0.1
Lids Close	Intensity	[0 .. 1]	0.25
	Duration	[0.3 .. 1.3]	0.1
EyeBrows	Intensity	[-1 .. 1]	0.25
	Duration	[0.3 .. 1.3]	0.1
Eye Squeeze	Intensity	[0 .. 1]	0.25
	Duration	[0.3 .. 1.3]	0.1
Smile	Intensity	[-1 .. 1]	0.25
	Duration	[0.3 .. 1.3]	0.1
Shoulders Up	Intensity	[0.1 .. 0.4]	0.1
	Duration	[0.3 .. 0.9]	0.1
Gesture Phase Freeze	HOLD, NONE, RETRACT	[-1,0,1]	-
	Duration	[0.3 - 1.3]	0.1

TABLE 3
The genes of our genomes in the genetic algorithm are nonverbal signals, each one is expressed with a pair of discrete features: intensity and duration with ranges and steps as indicated in the table, except for Gesture Phase Freeze which has categorical values instead of intensity.

Each gene (i.e. nonverbal signal) is represented with a pair of discrete features: **intensity** and **duration**. The former is the intensity of the behavior, whereas the latter is the duration.

We have chosen these two features because they were shared by all identified nonverbal signals types, except for gesture freezes that are represented by 3 categorical values as described later. Moreover, intensity and duration manipulations can yield to the generation of a great variety of expressive behavior that, as described in the study design in next section, can express friendly/hostile reactions to interruptions.

We used symbolic ordinal values as ranges of our features, except for gesture freezes that are categorical (i.e. HOLD, NONE, RETRACT), in order to have a defined range for them (i.e. a min and max value) and because continuous values often lead to a space of parameters too big to handle. Furthermore, a human observer might not be able to observe the subtle changes resulting from continuous features that are too close to each other.

Some of the ranges and steps of our features had to be adapted to our problem and to the ECA system used to produce such behaviors, as described later. Shoulders up, for instance, yielded odd visuals when adopting intensities outside the defined range. Features with negative intensities are related to behaviors with symmetrical (i.e. mirrored movements), such as head nods (from -1 to -0.1) and toss (from 0.1 to 1). Gesture Phase Freezes have 3 categorical values (HOLD, NONE and RETRACT) therefore, an intensity set to -1 indicates the duration parameter for a gesture hold, whereas an intensity set to 1 indicates the duration of a gesture retract.

5.1.2 The genetic operators

Typically, in genetic algorithms, *crossover* and *mutation* operators are defined for generating new genomes (i.e. solutions) from a given population of one or more genomes. Because we designed an interactive genetic algorithm, humans are part of the selection process as they judge the quality of a given solution providing a rating and choosing the most optimal ones for evolution. As described in Section 5.2, genomes were presented to human participants in the form of videos displaying an ECA timely exhibiting the nonverbal behaviors as reaction to a conversational interruption. These behaviors were generated by taking the features of each gene in a given genome. Participants had the task of choosing and rating the video, a genome of our algorithm, that in their opinion best corresponded to a friendly (or hostile) nonverbal reaction to an interruption. The selection process yielded a computed score, as described below.

Computer score of a genome. A score was computed for a genome once a user provided his/her rating. The score took into account three factors: the user's satisfaction level, an overall rating for that genome and the number of generations needed to find it. The overall rating was a measure of optimality of the solution for the user, whereas the satisfaction level helped us understanding how satisfactory for the user was the resulting video. The number of generations provided insights on the selection process, indicating the evolution that was needed prior to having the user establishing that a given genome was optimal. We combined these indicators to capture a more fine graded information about a given solution. For instance, consider the case when a user chooses a genome simply to finish the assigned task but without being satisfied of it, or a genome being chosen after only a few number of generations which values less compared to a genome that has been chosen after a greater number of generations.

In sum, let $norm$ be a normalization function, $UserSatisfLevel$ the user satisfaction level, $NumGen$ the number of generations and $StarRating$ the score attributed to a genome, we gave the following weights and computed the score CS of a genome g as follows:

$$CS_g = norm(UserSatisfLevel * 0.15) + norm(NumGen * 0.4) + (StarRating * 0.45)$$

Algorithm start. We initialized the search space of our genetic algorithm by using k-means clustering, thus creating 4 clusters from 1,000,000 randomly generated genomes and taking the 4 centroids as initial genomes. When some computed scores for genomes start to be available (i.e. participants started to select videos according to their given task as described later), the initialization takes only 2 centroids from the clustering and then the top 2 solutions ranked by computed score as initial genomes, thus taking into account also previous solutions found by users in input to the genetic algorithm. Users' chosen genomes (i.e. videos) undergo genetic transformations according to the *crossover* and *mutation* operators.

Crossover operator. Given two or more genomes named *parents* and composed by genes that are pairs of intensity and duration features. The crossover operator randomly picks a gene (i.e. the two features intensity and duration) among the parents and sets the pair of features in the

corresponding gene of the newly born child. The probability of selecting a parent is weighted on computed score values and normalized among all parents, therefore the higher a parent's score the higher is the probability of picking a gene from that parent.

Mutation operator. After crossover new genomes are created by mutating the resulting child. Given a genome, every gene within it has a probability to change of $\frac{1}{4}$ because we grouped genes by nonverbal modality as described below and the algorithm ensures to have for each grouped modality a gene that mutates:

1) **Head movements:** Head tilt (intensity and duration) + Head toss/nod (intensity and duration)
2) **Face expression:** Smile (intensity and duration) + Eyebrows raise (intensity and duration)
3) **Eyes:** Eyelids close (intensity and duration) + Eyelids squeeze (intensity and duration)
4) **Upper body:** Shoulders up (intensity and duration) + Gesture Phase Freeze Type (type and duration)

Shoulders are an exception, we needed to lower the probability of changing this gene's intensity and duration because we intended to have less frequent changes in order to reduce odd visuals, therefore the probability for this gene is: $\frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$.

Once a gene is selected, in order to variate its intensity and duration we applied regular Gaussian noise (mean $= 0$, standard deviation $= 1$) for 2 out of the 4 new genomes being created, and a wider noise in order to increase variability (mean $= 0$, standard deviation $= 2$) for the remaining 2 genomes to create.

5.2 User Study driven by Genetic Algorithm

We designed a user study driven by our interactive genetic algorithm. Four videos were presented to participants, each representing a genome, displaying an ECA expressing interruption reactions behaviors. The behaviors were rendered through the VIB/Greta platform [32].

5.2.1 Design

The initial 4 videos presented to participants were generated through the initialization phase of our algorithm described earlier. The VIB/Greta platform was used to generate the agent's behaviors in the videos. Participants were given the task of selecting the video that was closer to a friendly (resp. hostile) reaction to an interruption. They had the possibility to confirm a video and then score the satisfaction level, or select one or more videos by applying a score indicating closeness to a friendly (hostile) reaction in order to generate a new population of 4 genomes through the genetic algorithm running behind the *crossover* and *mutation* operators. Participants had also the possibility of re-generating 4 brand new solutions when none of the videos presented was selected.

In each video, we simulated an interruption in the following manner. The ECA always began with the following utterance accompanied by appropriate communicative gesturing behavior and facial expressions: "Alice in wonderland is an amazing book about a white rabbit and a little girl". We have chosen Alice in Wonderland as context for consistency with our experimental design already presented in [27]. The interruption came from a human pre-recorded female voice that said "Yeah, yeah I have read that book", in an excited manner. This ensured an appropriate expressiveness. The interruption systematically happened always when the ECA started to utter the word "about". In our classification of interruptions, such interruption can be considered as a tangentialization (i.e. the interrupter is telling the interruptee that s/he already knows the provided piece of information). The 4 videos were played in a synchronized manner when participants pressed a single play button. Mono sound was used to ensure all participants experienced the same stimuli regardless of the audio equipment being used.

In this work we focused on HALT as reaction type in order to simplify the space of solutions to be explored by our algorithm. In future work we will explore different reaction types (e.g. OVERLAP). We focused on the friendly-hostile axis of interpersonal attitudes for the user to judge the agent's reaction, leaving the dominant-submissive axis for future studies.

In sum, the task of participants was to find out which nonverbal behaviours of the virtual agent corresponded best to a friendly or hostile halt, i.e. when the agent reacts to an interruption with a friendly or hostile attitude. Given a reaction (friendly or hostile), a participant was assigned to one of the two conditions for the whole study. The study ended when the participant believed that s/he has reached an optimal solution (i.e., best corresponding behavior to a friendly or hostile halt) by selecting a video. The satisfaction level, the study condition (i.e. friendly or hostile), the genes of the genome representing the solution (e.g. nonverbal behaviors), the score attributed to the final solution (from 1 to 5) and the number of generations needed to reach it were then recorded.

5.2.2 Implementation

The system designed to implement our study is depicted in Figure 5. We set up a client-server architecture. A client web interface was used to display videos and get users' selections and ratings. Selected videos are sent to a Scheduler on the server that dispatched them to the genetic algorithm. A customized version of the VIB/Greta ECA platform [32], deployed on 3 different machines, generated the new videos to be sent back to the web interface. We parallelized the video generation server-side to have faster response times of the study, in particular when multiple participants were running it at the same time.

Our ECA in the VIB/Greta platform was animated via XML languages that can specify its communicative intentions with Function Markup Language (FML) [33], and behaviours with the Behavior Markup Language (BML) [34]. For our study, the VIB/Greta platform receives the set of behaviors to animate the nonverbal reaction to an interruption transforming the genome received from the web interface into a BML file readable by the platform. The genome is used to configure ECA reaction to the interruption and to compute it in real-time to be recorded in a video temporarily stored on the server in a Video Files Storage module. Once a video is generated, our Scheduler sends it back to the client's web interface for being watched by the participant. The

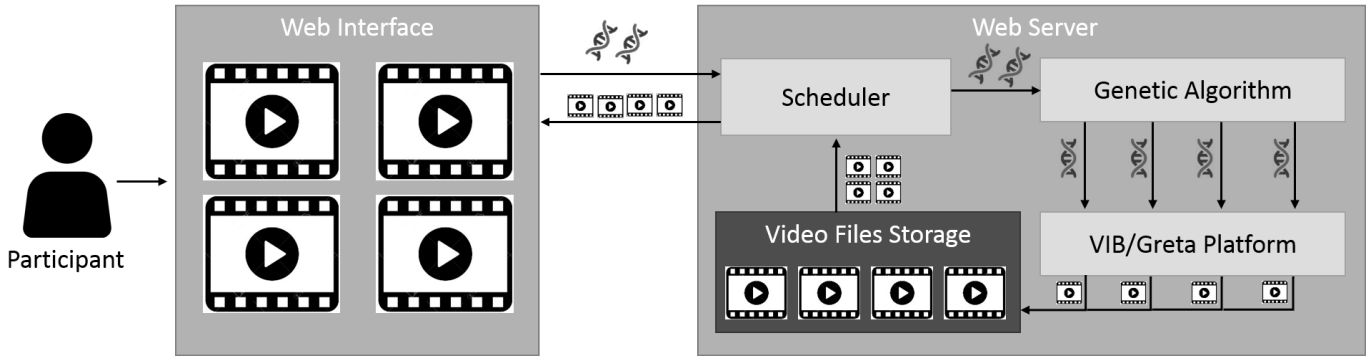


Fig. 5. The design of our study summarized in a diagram: the participant watched and selected the videos showing the ECA's nonverbal reaction to interruptions from a web interface. The genetic algorithm works behind together with the VIB/Greta platform to generate videos from selected genomes as more generations are produced until a satisfactory video is chosen.

whole process of creating a new generation (4 videos) and sending them to the client could take from 5 to 10 seconds.

5.2.3 Procedure

The study was accessible online. The landing page presented the goals and explained the type of data being collected. Then participants read and provided their consent in a separate form. Prior to starting, a video tutorial was shown. Once started, a participant was assigned to a condition (i.e. friendly vs. hostile ECA reaction to an interruption) and then went through exploring/generating the solutions until satisfaction was reached or the study was abandoned.

5.2.4 Demographics

We recruited 180 participants (60% males, 38.3% female and 1.7% undisclosed gender) over mailing lists and from the university students. Participant were mainly aged between 21-30 years (43%) and between 31-40 years (39.8%). Out of these 180 participants, 60 reached the end of the study by selecting a final video as their preferred solution, and 44 of had a satisfaction rating at least of 4 out of 5 points (M = 3.86, SD = 0.83). Participants were mainly from France (46.7%) and were almost exclusively non-native English speakers (98.3%).

5.2.5 Preliminary Results

Participants ended the study with a satisfaction level, on average, greater than 4 out of 5 points. However, in our opinion, the many others that abandoned the study were discouraged by the latency time needed to generate new solutions (i.e. new videos). From the remaining data of the participants that ended the study we obtained 60 usable genomes describing the ECA's nonverbal behaviour for a friendly (26 genomes) and hostile (34 genomes) reaction to an interruption. The output was a corpus of configured parameters for the reaction of an autonomous agent ranked by computed scores. From this corpus it was possible to analyze the choices of the participants and the evolution of the generated solutions. Given an interpersonal attitude (i.e. friendly/hostile), our ultimate goal was to build an ECA's interruption behaviour model to express such attitude. We identified this as a classification problem and exploited the collected data by using decision trees. Confirmation of the tendencies that we have found or different classification techniques can be investigated when a greater amount of data will be available. The decision tree generated from our data is depicted in Figure 6. The tree shows the different set of nonverbal behaviors (i.e. configurations) that allows an ECA to react in a friendly or an hostile manner to an interruption. For example, for a friendly reaction, the ECA can smile for longer while raising its eyebrows or, alternatively, it can smile for a shorter duration while retracting its gesture for longer and exaggerating its eye browns raising behavior. On the other hand, for expressing an hostile reaction to an interruption, it can produce short and low intensity smiles along with shorter timing of gesture retract.

6 Conclusion

In this work we examined nonverbal reactions exhibited by an interruptee during conversational interruptions and proposed a novel technique driven by an evolutionary algorithm to build a computational model for ECAs to manage user's interruptions. We proposed a taxonomy of conversational interruptions adapted from social psychology, an annotation schema for semi-automatic detection of user's interruptions and a corpus-based observational analysis of human nonverbal reactions to interruptions. Finally, we presented the design and realization of an interactive study driven by our evolutionary algorithm, where participants interactively built the most appropriate set of multimodal reactive behaviours for an ECA to display interpersonal attitudes (friendly/hostile) through nonverbal reactions to a conversational interruption. Preliminary results of this methodology and insights for exploiting the collected data through decision trees have been described.

The advantage of our methodology, using evolutionary algorithms, is that a complex space of solutions can be interactively explored by a human to converge towards an optimized solution. Furthermore, the ECA's computational model is directly built from users' perception of the ECA's exhibited nonverbal reactions.

We believe that this work is promising for a variety of future research directions. First, interruption reaction behaviours have been rarely studied in the field of ECAs. These behaviours may not be fundamental for the user to understand the dialog with the agent, but they can improve fluidity and naturalness of interaction, leading to more believable ECAs. Secondly, we introduced a novel approach to

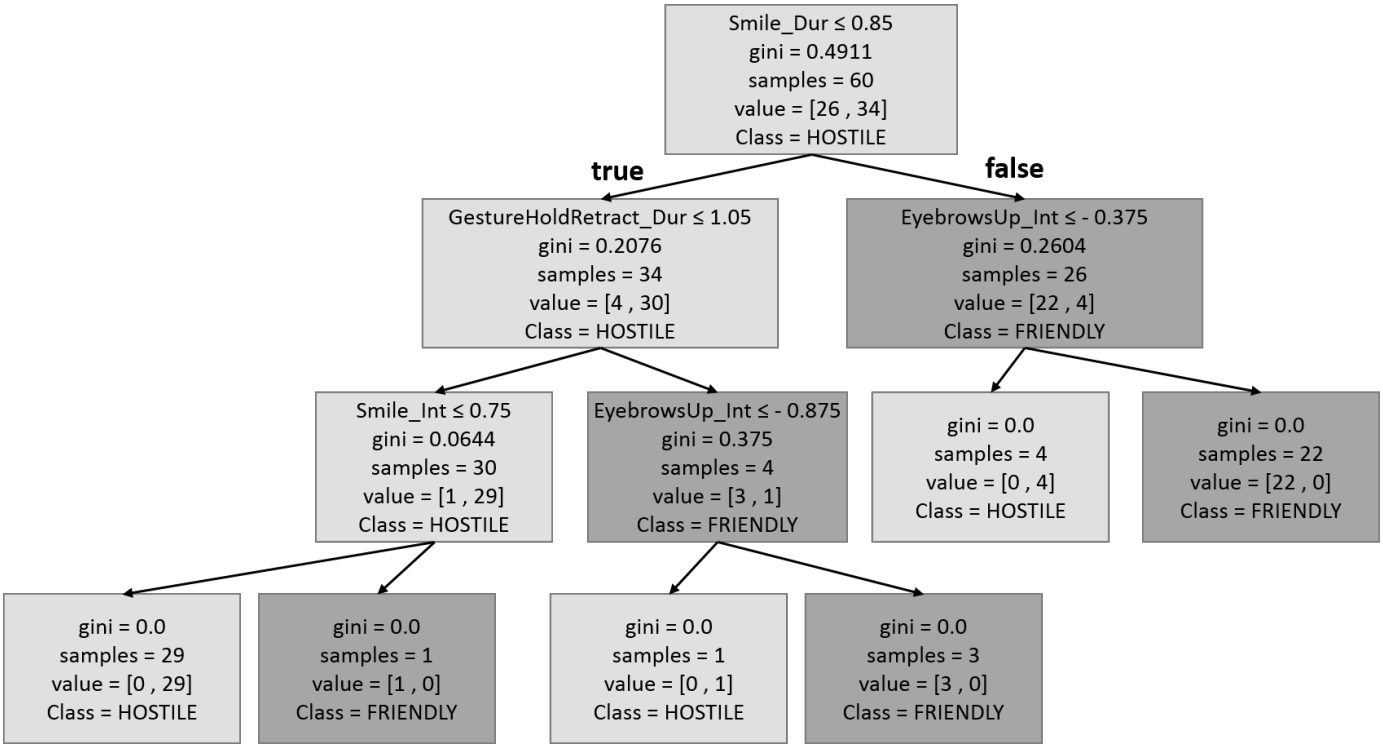


Fig. 6. An example exploitation of the data collected by our interactive genetic algorithm via a decision tree learned from the data.

design the behaviours of an ECA based on interactive evolutionary algorithms. This approach displayed very promising potentials, by adopting state-of-the-art techniques from the literature on genetic algorithms and offering ways to explore complex space with large number of variables. Therefore, an interesting development is to investigate more ways to improve convergence of the genetic algorithm towards optimal solutions.

Our short term plan is to pursue this work and gather more data, also including other types of reactions (OVERLAP and PLAN) and dimensions of social attitudes (e.g. the dominance axis) in order to explore more relationships between the configurations of agent nonverbal behaviours and the perceived attitudes. Moreover, the proposed annotation schema is a first attempt at integrating interruption reaction behaviours in the design of an ECA. While we only considered a subset of interruptions according to our taxonomy, in the near future we aim at including other types of speaker-switch such as smooth-switch and butting-in interruptions. Finally, as described in Section 5.2.4, participants ended the study with a satisfaction level, on average, greater than 4 out of 5 points. However, many others dropped the study, in our opinion, because they were discouraged by the latency time needed to generate new solutions (i.e. new videos). This is certainly an architectural improvement that is needed in order to continuing working with our methodology, thus requiring parallel computing and multiple server instances to process new generations requests more rapidly.

Acknowledgments

We would like to thank Johannes Wagner, Tobias Baur and Elisabeth André from University of Augsburg for their valuable help in developing the NoVa automatic annotation tool. We also thank Stéphane Doncieux from Sorbonne University for his advises on evolutionary algorithms.

References

[1] S. Duncan, "Some signals and rules for taking speaking turns in conversations." *Journal of personality and social psychology*, vol. 23, no. 2, p. 283, 1972.

[2] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in psychology*, vol. 6, p. 731, 2015.

[3] D. C. O'Connell, S. Kowal, and E. Kaltenbacher, "Turn-taking: A critical analysis of the research tradition," *Journal of psycholinguistic research*, vol. 19, no. 6, pp. 345–373, 1990.

[4] J. A. Goldberg, "Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power-and rapport-oriented acts," *Journal of Pragmatics*, vol. 14, no. 6, pp. 883–903, 1990.

[5] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, *Embodied Conversational Agents*, 1st ed. Cambridge: The MIT Press, 2000.

[6] S. D. Farley, "Attaining status at the expense of likeability: pilfering power through conversational interruption," *Journal of nonverbal behavior*, vol. 32, no. 4, pp. 241–260, 2008.

[7] G. W. Beattie, "Interruption in conversational interaction, and its relation to the sex and status of the interactants," *Linguistics*, vol. 19, no. 1-2, pp. 15–36, 1981.

[8] E. A. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Language in society*, vol. 29, no. 1, pp. 1–63, 2000.

[9] N. Crook, D. Field, C. Smith, S. Harding, S. Pulman, M. Cavazza, D. Charlton, R. Moore, and J. Boye, "Generating context-sensitive eca responses to user barge-in interruptions," *Journal on Multimodal User Interfaces*, vol. 6, no. 1-2, pp. 13–25, 2012.

[10] S. Kopp, H. van Welbergen, R. Yaghoubzadeh, and H. Buschmeier, "An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 97–108, 2014.

[11] M. Wester, D. A. Braude, B. Potard, M. P. Aylett, and F. Shaw, "Real-time reactive speech synthesis: Incorporating interruptions," in *Interspeech 2017, 18th Annual Conference of the International*

Speech Communication Association, Stockholm, Sweden, August 20-24, 2017, 2017, pp. 3996–4000.

[12] K. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, 2005.

[13] M. Argyle, *Bodily Communication*, ser. University paperbacks. Methuen, 1988.

[14] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: the behavior expression animation toolkit," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '01. New York, NY, USA: ACM, 2001, pp. 477–486.

[15] T. Bickmore, D. Schulman, and G. Shaw, "Dtask and litebody: Open source, standards-based tools for building web-deployed embodied conversational agents," in *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, ser. IVA '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 425–431. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04380-2_46

[16] A. Cafaro, H. H. Vilhjálmsson, and T. Bickmore, "First impressions in humanagent virtual encounters," *ACM Trans. Comput.-Hum. Interact.*, vol. 23, no. 4, p. 24:124:40, Aug. 2016. [Online]. Available: http://doi.acm.org/10.1145/2940325

[17] B. Rossen and B. Lok, "A crowdsourcing method to develop virtual human conversational agents," *International Journal of Human-Computer Studies*, vol. 70, no. 4, pp. 301 – 319, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S107158191100156X

[18] M. Ochs, B. Ravenet, and C. Pelachaud, "A crowdsourcing toolbox for a user-perception based design of social virtual actors," in *Computers are Social Actors Workshop (CASA)*. Citeseer, 2013.

[19] M. Ochs, C. Pelachaud, and G. Mckeown, "A user perception–based approach to create smiling embodied conversational agents," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 7, no. 1, p. 4, 2017.

[20] J. Youngquist, "The effect of interruptions and dyad gender combination on perceptions of interpersonal dominance," *Communication Studies*, vol. 60, no. 2, pp. 147–163, 2009.

[21] K. Murata, "Intrusive or co-operative? a cross-cultural study of interruption," *Journal of Pragmatics*, vol. 21, no. 4, pp. 385–400, 1994.

[22] S. H. Ng, M. Brooke, and M. Dunne, "Interruption and influence in discussion groups," *Journal of Language and Social Psychology*, vol. 14, no. 4, pp. 369–381, 1995.

[23] E. A. Schegloff, "Accounts of conduct in interaction: Interruption, overlap, and turn-taking," in *Handbook of sociological theory*. Springer, 2001, pp. 287–321.

[24] B. Ravenet, A. Cafaro, M. Ochs, and C. Pelachaud, "Interpersonal attitude of a speaking agent in simulated group conversations," in *International Conference on Intelligent Virtual Agents*. Springer, Cham, 2014, pp. 345–349.

[25] E. Bevacqua, E. De Sevin, S. J. Hyniewska, and C. Pelachaud, "A listener model: introducing personality traits," *Journal on Multimodal User Interfaces*, vol. 6, no. 1-2, pp. 27–38, 2012.

[26] H. Z. Li, "Cooperative and intrusive interruptions in inter-and intracultural dyadic discourse," *Journal of Language and Social Psychology*, vol. 20, no. 3, pp. 259–284, 2001.

[27] A. Cafaro, N. Glas, and C. Pelachaud, "The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2016, pp. 911–920.

[28] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. Andr, and M. Valstar, "The NoXi database: Multimodal recordings of mediated novice-expert interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI 2017. New York, NY, USA: ACM, 2017, p. 350359. [Online]. Available: http://doi.acm.org/10.1145/3136755.3136780

[29] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "The social signal interpretation (ssi) framework: Multimodal signal processing and recognition in real-time," in *Proceedings of the 21st ACM International Conference on Multimedia*. New York, NY, USA: ACM, 2013, pp. 831–834.

[30] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

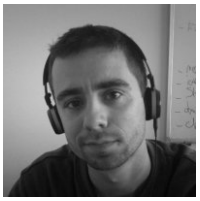
[31] C. M. Anderson-Cook, "Practical genetic algorithms," 2005.

[32] F. Pecune, A. Cafaro, M. Chollet, P. Philippe, and C. Pelachaud, "Suggestions for extending saiba with the vib platform," in *Proceedings of the Workshop on Architectures and Standards for Intelligent Virtual Agents at IVA 2014*, 2014.

[33] D. Heylen, S. Kopp, S. C. Marsella, C. Pelachaud, and H. Vilhjálmsson, "The next step towards a function markup language," in *Intelligent Virtual Agents*. Springer, 2008, pp. 270–280.

[34] S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, and H. Vilhjálmsson, "Towards a common framework for multimodal generation: The behavior markup language," in *International workshop on intelligent virtual agents*. Springer, Berlin, Heidelberg, 2006, pp. 205–217.

Angelo Cafaro was a Research Assistant within the Greta Team of the ISIR-CNRS laboratory. He conducts research in the area of embodied conversational agents with emphasis on social interaction and nonverbal expression of social attitudes. His work is also part of the SAIBA framework, in particular he proposed a unified specification for the Function Markup Language (FML).



Brian Ravenet is an Assistant Professor at University Paris Saclay, conducting his research within the CPU team of the LIMSI-CNRS laboratory. He is particularly interested in the nonverbal capabilities of virtual characters and in their potential to design better communicative agents.

Catherine Pelachaud is a Research Director at ISIR-CNRS laboratory and the head of the Greta Team. She works on all the aspects of human-agent interaction, from the user perception and recognition of emotions to the production of adapted responses (verbal and nonverbal).