A Methodology for the Automatic Extraction and Generation of Non-Verbal Signals Sequences Conveying Interpersonal Attitudes

Mathieu Chollet, Magalie Ochs, Catherine Pelachaud

▶ **To cite this version:**

Mathieu Chollet, Magalie Ochs, Catherine Pelachaud. A Methodology for the Automatic Extraction and Generation of Non-Verbal Signals Sequences Conveying Interpersonal Attitudes. IEEE Transactions on Affective Computing, 2017, XX, pp.1 - 1. 10.1109/TAFFC.2017.2753777 . hal-01793271

HAL Id: hal-01793271
https://hal.science/hal-01793271

Submitted on 18 May 2018

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.

A Methodology for the Automatic Extraction and Generation of Non-Verbal Signals Sequences Conveying Interpersonal Attitudes

Mathieu Chollet, Magalie Ochs and Catherine Pelachaud

Abstract—Depending on their application, Embodied Conversational Agents (ECAs) must be able to express various affects or social constructs such as emotions or social attitudes. Non-verbal signals, such as smiles or gestures, contribute to the expression of attitudes. Social attitudes affect the whole behavior of a person: as Scherer puts it, they are "characteristic of an affective style that colors the entire interaction" [1]. Moreover, recent findings have demonstrated that non-verbal signals are not interpretated in isolation but along with surrounding signals: for instance, a smile followed by a gaze aversion and a head aversion may signal embarassment rather than amusement [2]. Non-verbal behavior planning models designed to allow ECAs to express attitudes should thus consider complete sequences of non-verbal signals and not only signals independently of one another. However, existing models do not take this into account, or in a limited manner. The contribution of this paper is a methodology for the automatic extraction of sequences of non-verbal signals characteristic of a social phenomenon from a multimodal corpus, and a non-verbal behavior planning model that takes into account sequences of non-verbal signals rather than signals independently. This methodology is applied to design a virtual recruiter capable of expressing social attitudes, which is then evaluated in and out of an interaction context.

Index Terms—Embodied Conversational Agents, non-verbal signals sequences, social attitudes.

1 Introduction

Embodied Conversational Agents (ECAs) are a type of multimodal interface which uses human-like communication modalities, such as speech, gestures, gaze or prosody to communicate with users. ECAs are currently used in various kinds of applications, such as health coaching [3] or social skills tranining [4].

Depending on the kind of applications envisioned, designing such ECAs can be a more or less complex endeavor. For instance, the behavior of ECAs can be planned in advance (scripted) in the case of applications that are fully controlled (*ex.* interaction with the ECA by the use of a menu in a fixed scenario). Conversely, fully simulating social interactions between a user and an autonomous ECA is a much more daunting task. One has to endow the ECA with the capacity to recognize and interpret the signals of the user (*e.g.* detect and understand the user's emotions); the ECA must also be able to perform various communicative functions: participate in the regulation of the interaction (*e.g.* find the right time to take the speaking turn, indicate that it is listening to the user); indicate its thoughts (*e.g.* convey its doubt through its behaviors); it must be able to express the different socio-emotional affects that are relevant to the application (*e.g.* attitudes) at the right time and in a manner that will be perceived and recognized by the user [5].

The different communicative fonctions that we outlined are realized by multimodal signals (*e.g.* gestures, facial expressions, gaze, *etc.*). Behavior planning models for ECAs must therefore take into account numerous communicative functions and consider how the multimodal signals that realize them are interpreted. The standard approach in the domain of ECA behavior planning models is to define a lexicon in which the mapping between communicative function and multimodal signals is defined for all the functions considered in the target application [6], [7], [8]. When one of these communicative functions has to be realized, the behavior planning component can refer to this lexicon to select appropriate non-verbal signals. In order to define such a lexicon, one can use relevant literature on non-verbal behavior [9], [10], [11], [12], analyze multimodal corpora [13], [14] or build models with machine learning techniques [15], [16]. For instance, facial expressions of emotions have been studied by displaying photographs of actors expressing an emotion at its peak [9], [17], [18]: in order to allow an ECA to express these emotions, one can define representations of these expressions in a lexicon [19], [20].

These different methods allow to realize each communicative function specified in the lexicon. However, these realizations are performed independently of one another, *i.e.* the choice of the signals used to realize a specific function is done without taking into account surrounding multimodal signals. Such an approach is insufficient in some cases: for instance, some affects are expressed through complex sequences of facial or multimodal signals [2], [21], [22]. The interpretation of a non-verbal signal can be influenced by surrounding multimodal signals: for instance, a smile does not necessarily convey amusement, but can be a sign of embarrassment if it is followed by a gaze aversion and a head aversion [2]. A behavior planning model that does not take this into account could then inadvertently generate a

- *M. Chollet and C. Pelachaud are with Institut Mines-Telecom, Telecom Paristech, CNRS-LTCI, Paris, France. E-mail: {mathieu.chollet, catherine.pelachaud}@telecom-paristech.fr*
- *M. Ochs is with LSIS, Université Aix-Marseille, Marseille, France. E-mail: magalie.ochs@lsis.fr*

Manuscript received Month 01, 2015; revised Month 31, 2016.

sequence of signals conveying something completely different than what was originally planned.

Similarly, social attitudes, or interpersonal stances, are not expressed by one particular expression at a given time: rather, they are a pervasive affect that influences the whole behavior of a person throughout longer periods of time [1]. A behavior planning model designed for the expression of attitudes, or another similar socio-emotional construct, should thus ensure the coherence of the signals selected to express input communicative functions with respect to the chosen attitude to express. However, while social attitudes are very relevant in many application domains of ECAs, such as counseling [3] or improvisational scenarios [23], [24], and models have been proposed for allowing ECAs to express them [8], [25], none of these models consider sequences of non-verbal signals. On the contrary, they generate multimodal signals independently of other surrounding signals and communicative functions, and thus cannot ensure the coherence of the whole behavior of the ECA regarding an input attitude. Moreover, research on how attitudes are expressed through non-verbal behavior has focused on the study of one modality at a time, and does not provide us information on sequences of multimodal signals.

In this paper, we describe a methodology for automatically extracting sequences of non-verbal signals characteristic of a social attitude from a multimodal corpus. These sequences are then used to build a behavior planning model that computes sequences of multimodal signals, ensuring the coherence of these signals with respect to an input attitude. The job interview domain is very well suited for the study of the expression of social attitudes: thus, we applied our methodology to the study of the expression of social attitudes by sequences of multimodal signals in a corpus of job interviews. We used the resulting behavior planning model for implementing a virtual recruiter, which was integrated in a job interview simulation platform developed in the TARDIS project (*www.tardis-project.eu*). This project aimed at providing tools for youngsters to train their social skills in order to promote youth employment.

In the next section, we start by reviewing social attitudes in more detail: we present how they are expressed through multimodal behavior and describe existing models that have been introduced to allow ECAs to express them. Noting none of these models has considered sequences of non-verbal behavior, we detail in section 3 a novel methodology for the automatic extraction of sequences of multimodal signals which convey a considered socio-emotional affect, and how this dataset of sequences is used to create a behavior planning model. In section 4, we explain how we designed a virtual recruiter capable of expressing social attitudes by using the behavior planning model resulting from our method. An evaluation of the perception of this virtual recruiter's attitudes was performed with a study described in section 5.

2 Social attitudes in human-human and human-computer interaction

In this section, we present the concept of social attitude in more detail, and introduce our chosen representation of this phenomenon. We then describe how attitudes are expressed through non-verbal behavior. Finally, we outline existing models for attitude expression by ECAs.

2.1 Theoretical background

Social attitudes or interpersonal stances have been studied in the fields of social psychology, social linguistics and in social signals processing. As a consequence, many definitions of it have been proposed. Scherer described interpersonal stances as "*characteristic of an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, coloring the interpersonal exchange in that situation (e.g. being polite, distant, cold, warm, supportive, contemptuous).*" [1]. Du Bois proposed this definition: "*Stance is a public act by a social actor, achieved dialogically through covert communicative means, of simultaneously evaluating objects, positioning subjects (self and others), and aligning with other subjects, with respect to any salient dimension of the sociocultural field.*" [26]. A review of the topic was proposed by Chindamo *et al.* [27]. From these definitions we can draw the following conclusions:

- Social attitudes are expressed in order to position oneself towards another and to express an evaluation of someone else [26].
- The expression of a social attitude can either *spontaneous* or *strategic* [1], *i.e.* it is possible to consciously express a certain attitude regarding a particular goal.
- Attitudes are expressed both through verbal and non-verbal behavior [27]. However, they are not expressed at a certain time, but rather they permeate the whole behavior of a person [1].

Argyle proposed a bi-dimensional representation of attitudes [28]. A first dimension is *affiliation*, which can be characterized as the desire for a close relationship with one's interlocutor. Positive values denote a friendly attitude while negative values represent a hostile or unfriendly attitude. The other dimension is *status*, which is used to express the social superiority or inferiority of a person towards their interlocutor. Positive values indicate a dominant person, while negative values are used for submissive individuals. These two dimensions can be represented on an interpersonal circumplex [29]. While other representations of social attitudes have been proposed [30], [31], we use Argyle's representation as it is widely adopted in the field of ECAs [8], [23], [24], [25], and as it allows us to rely on the numerous works that studied the multimodal expression of friendliness and dominance.

2.2 Multimodal behavior and attitudes

The display of multimodal signals influences the perception of the attitude of an individual.

Vocal and verbal behavior: The amplitude of the voice influences the impression of dominance [32]. A lower pitch is correlated with dominance in the case of female speakers, while it is the opposite for male speakers [33]. Hesitations can indicate embarassment and submissiveness [34], and a speaker producing less hesitations and pause fillers is regarded as more dominant [35]. Laugther can be a cue of affiliation and friendliness [36]. Finally, turn-taking behavior influences the perception of attitudes: interruptions are seen

as signs of hostility and dominance [37]. The dominance degree of a person is directly tied to the amount of time he speaks in a conversation [38], [39], [40], [41].

Gestures: Most gestures are tied to speech [42], and are used to communicate certain intentions (*e.g.* a thumbs up gesture is used to congratulate or show approval) or help regulating speech or the interaction (*e.g.* give the turn, emphasize a word). However they are also tied to the expression of attitudes. *Adaptors* (*e.g.* scratching, manipulating objects) can express stress and are associated to submissive attitudes [12], [43]. Touch gestures are tied to friendly and dominant attitudes [44], [45], [46]. The expressivity of gestures also influences the the perceived attitudes. A person performing larger gestures is seen as more dominant [44], and a person performing numerous gestures is seen as dominant and friendly [43], [44], [47].

Postures: Postures are often adopted unconsciously and are a very revealing signal of a person's mental state [48]. As a general rule, being positioned closer to someone or being oriented towards him is a sign of affiliation [43], [45], [46], [47]. Adopting the same posture as one's interlocutor is also a sign of friendliness [49]. Similarly to the fact that larger gestures are dominance cues, adopting a posture that takes a large amount of space (*e.g.* put the arms behind the head) is a sign of dominance, while a curled postured indicates submissiveness [28], [40], [43], [44], [50].

Gaze: The various functions of gaze have been largely studied by Kendon [10], Argyle and Cook [51]. While gaze is a crucial signal in the regulation of the interaction (turn taking, attention), it also participates in the expression of attitudes [45], [46], [51], [52]. A large quantity of mutual gaze indicates a great affiliation. Generally, looking more at one's interlocutor is a sign of friendliness, unless we look "*too much*", as staring is seen as a hostile and dominant signal. Conversely, averting gaze is seen as a sign of submissiveness.

Head: Head movements and positions can signal many things such as approval or listening [53], as well as displays of attitudes. A head tilted upwards is associated with dominance while a head tilted downwards indicates submissiveness [54], [55]. Tilting one's head sideways (*head canting* or *cocking*) is a cue of friendliness [56], [57] or submissiveness [56], [58], [59]. Head shakes are typically associated with expressions of disapproval, and are more frequently observed in dominant persons [43], [44]. The relationship between head nods and attitude is more unclear as some studies found no link between them [55], [60], while others report a (weak) correlation between the amount of head nods and dominance [47] or a correlation with submissiveness [61].

Facial expressions: Facial expressions of emotions have been largely studied [62], [63]. The tendency of an individual to display or inhibit certain emotions is linked with the attitude he expresses [44], [64]: expressing joy often is associated with friendliness and dominance, anger and disgust are tied with hostility and dominance, while fear and sadness are associated with submissiveness. Facial movements have also been studied independently of emotions: smiles are cues of friendliness and submissiveness [40], [56], [64], [65], [66], [67]. Eyebrow movements seems to be perceived differently depending on one's culture: in western cultures, frowing indicates dominance [65], [68], while Keating *et al.* observed that in rural Thailand, raising eyebrows indicates dominance [65].

These works highlight the particular influence of certain modalities on the expression of attitudes. However, an attitude is expressed by multimodal behaviors that color the whole communication display. Therefore, multimodal signals should not be considered in isolation but as parts of sequences of multimodal signals when planning the behavior of an ECA for attitude expression.

2.3 Related work in attitude expression for ECAs

Various models for the expression of attitudes by ECAs have already been proposed.

In the Demeanour project, Ballin *et al.* [23] proposed a framework for automating the generation of users' avatars' animations in the context of an application for improvising stories. The authors reviewed the literature on interpersonal attitudes and defined rules describing the influence of Argyle's attitude dimensions on various variables, such as relaxation or the amount of occupied space. Using this model, when users changed the attitude of their avatar, these variables would be modified, which in turn would affect their avatar's posture and gaze.

The Laura relational agent was developped in order to sustain long term relationships with users [3]. The authors proposed a relational model that would adjust the behavior of the agent depending on how many times a user interacted with it. As the number of interactions grew over time, the agent would produce more gestures, head nods, facial expressions, and would appear closer to the user. The authors conducted a month-long study where the relational model was compared to a baseline scripted version of the agent. Their results indicated that the users rated the relational agent higher on measures of appreciation, trust and respect.

While most behavior models for ECAs focus on agents acting as speakers or as the direct addressees of a speaker, Lee and Marsella modelled the behavior of agents acting as bystanders or side-participants depending on their interpersonal attitude [24]. They defined a scenario in which the characters held various attitudes towards one another. They then carried out improvisational sessions with actors enacting this scenario in order to gather data. By observing the behaviors of the actors, they defined rules that can be used to select the behavior of side-participants under different interpersonal attitudes.

Cafaro *et al.* studied the expression of attitudes and personalities in the context of first encounters [8]. Using previous works on greetings, proxemics and interpersonal attitudes, the authors defined interpersonal distances at which greeting behaviors can be triggered. By manipulating the appearance of smile, gaze and proxemics behaviors at these distances, the proposed model influences the attitude that an ECA expresses during a greeting situation.

Ravenet *et al.* used a crowdsourcing method in order to build a computational model of attitude expression [25]. They created a web interface where users would be asked to select the behaviors (head movement, gesture with expressive parameters, gaze) of an ECA expressing a communicative intention (*e.g.* ask a question) with a particular attitude expressed in Argyle's bi-dimensional model (*e.g.*

dominant). This method allowed the authors to gather a large amount of different combinations of signals for the same input attitudes and intentions. They then used this data to build a Bayesian network, which can be used to select varied combinations of behaviors for expressing a communicative intention along with an attitude. The authors evaluated their model and found that, when combined with a dialogue model for attitude expression [69], it could express friendliness and hostility successfully.

Though several models have been proposed for the expression of attitudes by ECAs, they present some limitations. Most of these models only consider a limited part of the modalities involved in the expression of attitudes [3], [8], [23], [25]. Some are limited to particular conversational roles [24] or first encouters [8]. Last but not least, none of these models consider sequences of multimodal signals in the expression of attitudes. In this work, we tackle this last limitation by proposing a model that plans sequences of non-verbal signals; signals are not viewed as independent of one another. Indeed, the meaning of non-verbal signals can be changed by surrounding signals. Moreover, attitudes are not expressed at a given moment but affect the behavior of a person on longer periods. Planning sequences of non-verbal signals enables us to ensure the coherence of the different signals it contains regarding the attitude to express.

In order to build such a behavior planning model, we present a methodology which uses a sequential pattern mining technique to automatically extract non-verbal signals sequences expressing different attitude variations from a multimodal corpus. These extracted sequences are then used to train a behavior planning model, which was integrated in the Greta ECA platform [7].

3 Methodology

Many researchers have studied the role of certain modalities in the expression of attitudes [28], [38], [43], [44], [45], [46], [47], as well as in the expression of other affects such as emotions [17], [18]. Unfortunately, these works do not provide us information about the influence of sequences of non-verbal signals on the expression of attitudes. We thus designed a methodology that extracts knowledge automatically from a multimodal corpus and uses it to build a behavior planning model. The first step of this methodology is the acquisition of a multimodal corpus (Section 3.1). An algorithm relying on a sequential pattern mining technique is then applied on this corpus to extract relevant sequences of non-verbal signals (Section 3.2). These steps are represented in Figure 1. Finally, the last step is to build a behavior planning model from these sequences (Section 3.3). Such a model can generate an appropriate sequence of non-verbal signals, according to an input attitude and an input sentence containing communicative functions to be displayed by the ECA. While we designed our methodology around the specific study of social attitudes, it could be applied to other socio-emotional constructs.

3.1 Multimodal corpus acquisition and annotation

The first step in our methodology is to obtain a suitable multimodal corpus. The particularity of our approach is that the corpus is annotated at two levels: the non-verbal behaviors and the socio-emotional phenomena (in our case, social attitudes). For the annotation of attitudes, we propose to use a continuous annotation paradigm. Indeed, attitudes are not expressed at a particular moment but rather are characteristic of an affective style that evolves throughout an interaction. Our methodology could be applied to other socio-emotional affects that can also be represented under continuous dimensions, such as engagement [70], emotions, anxiety [71], and many more.

Corpus acquisition: Before choosing an existing corpus or creating a new one, some key points must be addressed [72]. One must make sure that the recording setup of the corpus allows for a appropriate capture of all the studied modalities, *e.g.* if the posture modality is investigated, then a video feed capturing the participants' full bodies is crucial. It is also important that the context in which the corpus is recorded matches the context of the end application. For instance, displays of emotions can be more or less socially inhibited depending on the social context, such as in a casual situation among friends or in a work meeting. Thus, a corpus used to build an ECA for an application simulating work meetings should not have been recorded in casual situations. Reviews of existing corpora that can be reused for studying multimodal communication can be found in [73] and [74]. A corpus that was elaborated on a similar domain as ours is the HuComTech corpus [75], which contains recordings of role-played job interviews. However the subjects of the corpus were acting as recruiters and were not professional recruiters, and their behaviors may not have reflected behaviors of actual recruiters. Instead, we chose to use a corpus of job interview enactments between human resources practitioners and youngsters (from 18 to 25 years old) that was collected in a study within the TARDIS project. The study was conducted at the Mission Locale Val d'Oise Est, a French job coaching association. It consisted in creating a situation of job interviews between 5 practitioners and 9 youngsters. The setting was the same in all videos. The recruiter and the youngster sat on each side of a table. A single camera embracing the whole scene recorded the dyad from the side, allowing us to observe the recruiters' full body, and thus annotate all the modalities involved in the expression of attitudes. From this study was gathered a corpus of 9 videos of job interviews lasting from 15 to 25 minutes each. The non-verbal behavior of the recruiters, their perceived attitudes and the turn taking were then annotated manually on 3 videos, for a total of slightly more than 50 minutes.

Non-verbal behaviors annotation: One crucial aspect in the process of annotating a multimodal corpus is the definition of an appropriate coding scheme, and the choice of appropriate tools for performing this annotation. Before starting the annotation, a review should be performed of the existing coding schemes for the studied socio-emotional phenomenon. The chosen coding scheme should include all the modalities involved in the expression of the studied phenomenon, and the granularity level of the scheme should be in line with the quality of the multimodal corpus and the level of detail of the annotated behaviors. We adapted the MUMIN multimodal coding scheme to our task and our corpus, defining categories and modalities to annotate

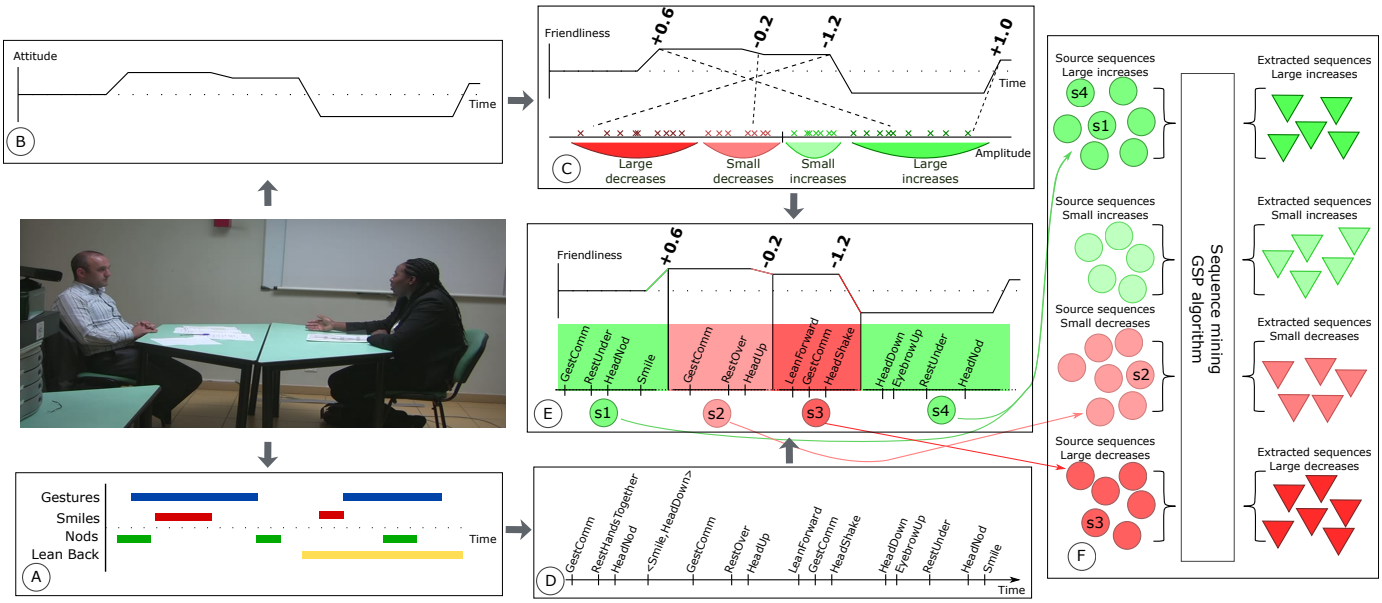


Fig. 1. Schematic representation of our methodology for automatic extraction of non-verbal signals sequences from a multimodal corpus. First, a multimodal corpus is collected and annotated at two levels, the non-verbal behaviors (A) and the attitude (B). Attitude variation events are identified and clustered by variation types (C) and used to segment the non-verbal behavior data (D) into groups of non-verbal signals sequences preceding attitude variations (E). Finally, a sequence mining technique is used to extract frequent sequences from these groups (F).

following our review of the literature on attitude expression (see Section 2). In particular, we added *posture* tags, which are very relevant to attitudes expression, and removed some facial expression tags (*e.g. lips protruded/retracted*) which we did not observe in our videos. The following modalities were considered: gestures (*e.g.* adaptors, deictics, 313 occurrences), hands rest positions (*e.g.* over or under table, arms crossed, 245 occurrences), postures (*e.g.* leaning backwards, 123 occurrences), head movements (*e.g.* nods, 281 occurrences) and head positions (*e.g.* head tilted sideways, 658 occurrences), gaze (*e.g.* looking downwards, 836 occurrences), smiles (91 occurrences) and eyebrow movements (156 occurrences). We used Praat [76] for the annotation of the audio stream and the Elan annotation tool [77] for the visual annotations. Full details on the coding scheme can be found in [78]. A single annotator fully annotated the non-verbal behavior for the three videos. A second annotation on 10% of the total annotated video length was performed one month after the initial annotation to measure the reliability of the coding. Cohen's Kappa measures were computed across the two annotations and were found to be mostly satisfactory (*e.g.* $\kappa = 0.80$ for gestures, $\kappa = 0.93$ for postures). The lowest score was found for eyebrow movements ($\kappa = 0.62$), which we had anticipated considering the video setup (the video camera captured the full scene, thus the faces of the participants were small in the video).

Socio-emotional phenomena annotation: Some socio-emotional phenomena such as attitudes, moods or engagement are not expressed at a specific time by prototypical expressions, rather they are expressed through the whole behavior of a person in longer periods of time. Therefore, we propose to annotate these phenomena with continuous annotations. In our case, this meant annotating the two dimensions of Argyle's attitude model, friendliness and dominance. We used GTrace, successor to FeelTrace [79], a tool that allows for the annotation of continuous dimensions over time. The speech was rendered unintelligible, as we wanted to focus on non-verbal behavior and the content of the recruiters' utterances could have affected the annotators' perception of attitudes. We asked 12 persons to annotate the videos with this tool. Each annotator had the task of annotating one dimension for one video, though some volunteered to annotate more videos. With this process, we collected two to three annotation files per attitude dimension per video.

In a nutshell, the corpus was annotated at two levels: the non-verbal behavior of the recruiters and their perceived attitudes. Our next step was to identify which sequences of non-verbal signals characterize social attitudes. We focused on the non-verbal signals sequences expressed by the recruiters when they are speaking. In the next section, we describe a method for extracting frequent non-verbal signals sequences from a multimodal corpus collected and annotated following our method.

3.2 Mining frequent sequences expressing attitudes

In order to extract significant sequences of non-verbal signals conveying social attitudes from our corpus, we chose to use a *sequence mining* technique. Such techniques have been widely used in tasks such as protein classification [80]. Recently, they were applied in the human-computer interaction domain: Martinez *et al.* used them to find sequences of video game players' key presses correlated with affects such as frustration [81] and Fricker *et al.* mined sequences of multimodal signals in human-robot interaction [82]. The novelty of our method is that it allows to extract patterns of multimodal signals conveying specific socio-emotional phenomena.

Frequent sequence mining techniques require a dataset of sequences of events. Since we investigated which sequences of signals convey attitudes, we decided to segment the non-verbal behavior data using the timestamps in the

annotations files where an attitude dimension begins to vary. We call these instants *attitude variation events*. This means that we obtained segments of time that directly precede a change in perceived attitude: studying these segments allowed us to investigate which sequences of non-verbal signals are relevant for the expression of social attitudes. The attitude traces were smoothed so that very small attitude variations (less than a twentieth of the annotation scale) would not be considered, as these small variations were likely to be noise from the continuous annotation process. When performing the segmentation, all the non-verbal signals starting between the end of an attitude variation event and the end of the next attitude variation event were included in a segment, regardless whether they were finished or not; we hypothesize that a signal does not have to be finished to be able to influence the expressed attitude. Once the data was segmented with these events, we kept only the segments where the recruiter is speaking, which means we did not investigate listening behavior. Our methodology could however be used in the same manner for studying sequences of non-verbal signals while listening.

Attitude variation events came with a wide range of values; therefore, we chose to differentiate between small and strong attitude displays. We used a K-means clustering algorithm with $k = 4$ to identify clusters corresponding to small increases, strong increases, small decreases and strong decreases.

The next step consisted of applying a frequent sequence mining algorithm to the set of segments of each attitude variation type. We used the Generalized Sequence Pattern (GSP) frequent sequence mining algorithm described in [83]. This algorithm extracts sequences without temporal information, *i.e.* it only represents that behaviors happened one after another. Since temporal information is not considered, is not able to differentiate between short and long signals. It also cannot represent simultaneous events (*e.g.* a smile and a nod happening simultaneously). More recent sequence mining techniques exist that take temporal information into account [82], [84]. However, we decided to choose a simpler model and to focus on the sequential representation, as a higher model complexity would require more data to learn from and would be harder to apply to our generation problem. Our model could potentially be complemented by related models considering simultaneous signals, such as [25]. The GSP algorithm requires as input a minimum support, *i.e.* the minimal number of times that a sequence has to be present in the corpus to be considered frequent; its output is a set of sequences along with their support. For instance, using a minimum support of 3, every sequence that is present at least 3 times in the data will be extracted. The GSP algorithm based on the *Apriori* algorithm [85] follows two steps: first, it identifies the frequent individual items in the data and then extends them into larger sequences by iteratively adding other items, pruning out the sequences that are not frequent enough. Having acquired a set of frequent sequences for each type of attitude variation, we characterized each of these sequences with several *quality measures*: *Support*, that is how many times the sequence appears in the data ($[0; \infty] \in \mathbb{N}$); *Confidence*, which represents the proportion of a sequence's occurrences that happen before a particular type of attitude variation ($[0; 1] \in \mathbb{R}$, 1 meaning this sequence only occurs before this attitude variation); *Lift*, which can be seen as how strong the confidence of a sequence is, compared to the random co-occurrence of sequence and the attitude variation given their individual support ($[0; \infty] \in \mathbb{R}$, a higher value representing a stronger association).

	Friendliness		Dominance	
	Segments	Freq. Sequences	Segments	Freq. Sequences
Large Increase	68	86	49	141
Small Increase	66	72	66	244
Small Decrease	77	104	80	134
Large Decrease	36	67	24	361

TABLE 1
Segment counts per cluster and frequent sequences per cluster.

Sequence	Attitude Variation	*Sup*	*Conf*	*Lift*
BodyStraight → ObjectManip	Friendliness Large Decrease	13	0.31	2.09
HeadNod → Smile	Friendliness Large Decrease	32	0.59	2.09
HeadNod → RestHandsTogether → Smile	Dominance Large Decrease	13	0.31	2.90
EyebrowsUp → RestOverTable	Dominance Large Increase	21	0.33	1.54

TABLE 2
Example sequences obtained from our corpus with sequence mining. Sup = Support. $Conf$ = Confidence.

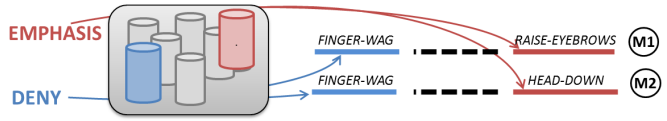
Table 1 presents the amount of segments of data preceding every type of attitude variation, and the amount of frequent sequences extracted them, while table 2 shows examples of extracted sequences. Using a minimum support of 10, we extracted a set of 879 sequences for dominance variations and 329 for friendliness variations. In the next section, we describe an algorithm for generating non-verbal signals sequences conveying attitudes, that makes use of the frequent sequences we presented in this section.

3.3 Planning non-verbal signals sequences conveying social attitudes

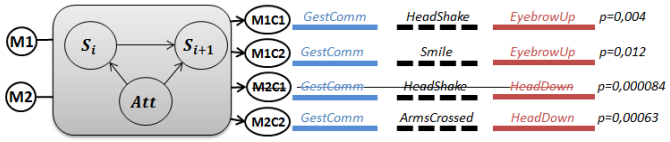
Given an input attitude that an ECA should express and an input utterance tagged with communicative functions that the ECA should perform, our model's objective is to generate a sequence of non-verbal signals that conveys the communicative functions with the desired attitude. We place ourselves within the SAIBA framework [86], where our model fulfills the role of the *Behavior Planner* module, *i.e.* it translates communicative functions into multimodal behaviors and schedules them. Input utterances and functions are defined in the Functional Markup Language (FML) [87], and scheduled multimodal signals are defined in the Behavior Markup Language (BML) format [86].

The main idea behing our algorithm is to first make sure the communicative functions of the FML message are conveyed by enumerating different combinations of signals that can express them, and then to enrich these with additional signals associated with a certain attitude. Our algorithm follows three steps, which are detailed in the following subsections, and represented in Figure 2. First, for each communicative function contained in the input FML message (Fig. 2a), we retrieve all the signals that can express this function using a lexicon approach, and build all the possible combinations of these signals that can express the input's communicative functions (Fig. 2b, section 3.3.1). Secondly, for each of these combinations, the algorithm

(a) Example of an input FML message.

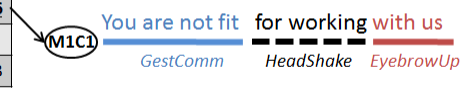


(b) Generation of minimal sequences, using the *behavior sets* of the input message's communicative functions (*i.e. DENY* and *EMPHASIS*). In this example, we obtain two minimal sequences, **M1** and **M2**.



(c) Minimal sequences **M1** and **M2** are enriched and unlikely candidate sequences are pruned using probabilities computed with the Bayesian Networks (*e.g.* **M2C1**).

Seq	Sup	Conf	Sc(s)
M1C1	21	0,6	**12,6**
M1C2	12	0,3	3,6
M2C2	23	0,42	9,43



(d) Selection of the final sequence, *i.e.* the sequence s with the highest score $Sc(s)$.

Fig. 2. Representation of the sequential behavior planning model.

finds all the time intervals where additional signals can be inserted, and builds a set of larger sequences by inserting additional signals in the available time intervals using a probabilistic framework (Fig. 2c, section 3.3.2). These signals will enable the agent to display its social attitude. The third step (Fig. 2d, section 3.3.3) consists of selecting the best sequence out of all these candidate sequences, by using a classification method trained on the frequent sequences that were extracted using the method described in Section 3.2.

3.3.1 Building minimal sequences from the input FML

In a conversation, communicative functions can be expressed through non-verbal behavior as well as through speech. For instance, in Western culture, it is possible to convey uncertainty by squinting the eyelids, tilting the head, or performing a particular hesitation gesture. When emphasizing a word, it is common to make a quick head movement downwards, and to raise one's eyebrows.

The FML language [87] is used to represent such communicative functions. The first step in our algorithm consists of retrieving all the possible non-verbal signals that can be used to express the functions contained in the input FML message. For this purpose, we used Mancini's framework [88], in which each communicative function is characterized by a *behavior set*. A behavior set is the specification of the different non-verbal signals that can be displayed by an ECA to express a communicative function. We can build a non-verbal signals sequence expressing an input message by selecting one signal in the behavior set of each communicative function of the input message. We call such a resulting sequence a *minimal sequence*. In our model, we only consider communicative functions that alter the speech prosody (*i.e.* pitch accents and pauses for emphasis) and communicative functions related to the speech pragmatics (*e.g.* communication of spatio-temporal information may trigger deictic gestures, particular concepts may trigger iconic gestures, *etc*). Once all the minimal sequences have been computed, *i.e.* all the different combinations of signals from the behavior sets have been enumerated, the next step consists of enriching these sequences with additional signals to convey the interpersonal attitude.

3.3.2 Generating new sequences

For every minimal sequence obtained in the previous step, the model starts by looking at all the time intervals where it is possible to insert other signals. For instance, if there is no head movement specified between two specified head positions, we might insert a head nod, or a head shake. Since the signals chosen for the minimal sequences are only related to speech prosody or to certain speech pragmatics, we make the hypothesis that the inserted signals will not conflict with the original communicative functions, and that the inserted additional signals will only contribute to expressing attitudes.

In order to choose these additional signals, we used the extracted frequent sequences dataset (Section 3.2) to learn a Bayesian Network's probabilities (BN). The nodes of the network represents the non-verbal signals and the social attitudes. The edges define a conditional dependence between two variables. The BNs enable us to represent the causal and non deterministic relation of the attitudes on the signals (*e.g.* there may be more smiles before friendliness increases) and the sequences of signals (*e.g.* hands rest pose changes may appear after a gesture). We assume that $P(S_{i+1}|S_i, S_{i-1}, ..., S_1, A) = P(S_{i+1}|S_i, A)$, where S represents signals, i is the index of a signal in the sequence, and A is the chosen attitude variation, *i.e.* the probability of a signal occurring in the sequence only depends on the input social attitude and the previous signal.

An interesting feature of this model is that non-verbal signals sequences absent from our corpus can still be generated, and their likelihood can be evaluated. Indeed, the representation of the sequences may lead to new sequences in the network. These new sequences are valuable as they can improve the variability of the recruiter's behavior beyond the sequences that were observed in the corpus. We trained a BN for dominance and another BN for friendliness, considering the two dimensions independently. We used the Weka open-source machine learning software [89] to train the networks.

The generation of new sequences starts with the minimal sequences obtained after the previous step (Section 3.3.1), and uses the BNs to add new signals in the available intervals. Thus, it is ensured that every generated sequence contains signals that express every input communicative function. The maximum sequence length is the amount of functions contained in the original FML message plus the amount of time intervals between them. In order to reduce computing time and to sort out sequences that are too unlikely, we compute the overall probability of every generated sequence, and only keep those whose probability is above a certain threshold λ. For our evaluation, we chose λ to be equal to $P(minimalsequence) * \alpha$ where

$P(minimal sequence)$ is the probability of the original minimal sequence and α is a coefficient. The alpha value was found after trial-and-error; we found the value 0.005 to be an adequate compromise between the amount of generated sequences and computing time. The generated sequences that are left after this pruning process are called *candidate sequences*. Having computed all the candidate sequences, the final step consists of selecting the one that is most likely to convey the input attitude.

3.3.3 Selecting the final sequence

Once the set of candidate sequences has been generated, we determine which sequence is the most likely to express the input attitude. We draw inspiration from a text classification technique that relies on frequent sequences [90]. For each candidate sequence s, we calculate a score $Sc(s)$ which will be used to select the final sequence. By noting $FSeq$ the group of frequent sequences extracted from our multimodal corpus, $\lambda_s = 1$ if $s \in FSeq$, $\lambda_s = 0$ if not, and sub_i one of the k sub-sequences contained in s, $Sc(s)$ is defined as follows:

$$Sc(s) = \begin{cases} Support(s) * Confidence(s) & \text{if } s \in FSeq \\ \sum_{i=1}^{k} \lambda_{sub_i} * Sc(sub_i) & \text{if not} \end{cases}$$

This means that we compute $Sc(s)$ directly from the sequence's quality measures if it was extracted with our methodology, and from it sub-sequences if it was not. The sequence s with the highest score $Sc(s)$ is selected as the final sequence to be displayed. The last step in our methodology is to express this sequence in the BML format [86]. This sequence can be sent to any ECA platform compatible with the BML standard to animate a virtual character. In our work, we use the Greta platform [7].

4 Application to the Design of a Virtual Recruiter

The methodology we presented in the previous section allowed us to build a behavior planning model for expressing social attitudes. This model was integrated in the TARDIS application where an ECA acts as a virtual recruiter.

4.1 Tardis platform overview

The Tardis job interview simulation platform contains four main components:

- Scenemaker [91] is used to create and execute the scenario of the job interview.
- SSI (*Social Signal Interpretation*) [92] recognizes signals from the user, such as gestures, postures, speech from audiovisual sensors.
- The affective core computes the emotions and attitudes that the virtual recruiter should display [93].
- The last module is the virtual recruiter, which we present in the next section.

These modules interact in the following manner: SSI detects the participant's multimodal signals and sends them to the affective core and virtual recruiter modules. Using these signals, the affective core determines if the participant performed well (*e.g.* an open posture and appropriate use of gestures will yield a better performance) and updates the recruiter's attitude depending on this performance (*e.g.* if the participant performs well, the recruiter will become more friendly). Using the same input signals, the virtual recruiter determines when the recruiter takes the speaking turn; Scenemaker then provides it with the next *dialogue act* to perform, *i.e.* a symbolic representation of the different sentences defined in the scenario. The virtual recruiter then computes the appropriate wording and non-verbal behavior to express this *dialogue act* while displaying the attitude computed by the affective core. When the virtual recruiter is not speaking, it also produces backchannels by using the multimodal signals of the participant.

4.2 Virtual recruiter architecture

The virtual recruiter's architecture we propose is represented in Figure 3. This architecture includes our behavior planner, allowing the ECA to express social attitudes through sequences of non-verbal signals, and other components, presented below, which perform other functions of multimodal communication (*e.g.* producing backchannels).

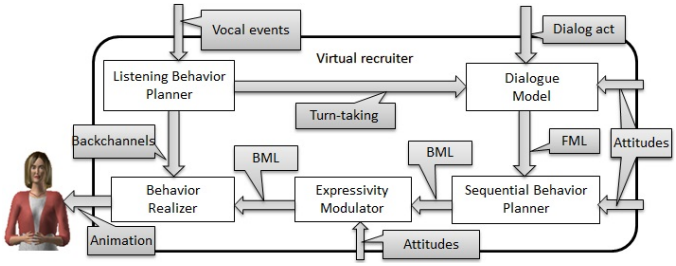


Fig. 3. The virtual recruiter's architecture.

Listening Behavior Planner: This module fulfills two functions: produce backchannels, which has been shown in the past to improve users' engagement [94], and detect the appropriate moment for the recruiter to take the speaking turn. This component receives vocal events from *SSI*. We used Bevacqua's listening behavior model [95], which was developped in the SEMAINE project and produces backchannels and mimicry behavior. For the turn-taking mechanism, we re-used a model developed by Gebhard *et al.* for the Tardis platform [91].

Dialogue model: Once the turn-taking mechanism determines that the recruiter should perform its next turn, Scenemaker is queried in order to return the next dialogue act of the scenario. This dialogue act is then transformed by the *Dialogue model* into an FML file, consisting of a sentence to utter enriched by communicative function tags. For a same dialogue act, the wording can influence the expressed attitude [96]; thus, we chose to use the *Dialogue model* of Callejas *et al.* [69], which uses a set of FML files for the expression of social attitudes in job interview scenarios, for unfriendly, neutral and friendly attitudes. We extended this collection of FML files to include dominant attitudes using the guidelines described in [97].

Expressivity modulator: the FML file chosen by the *Dialogue model* is sent to the *Sequential Behavior Planner*.

Its role is to transform, with respect to a given input attitude (*e.g.* a dominance increase), this FML file into a sequence of non-verbal signals expressed in the BML format. However, as our behavior planner model did not consider the expressiveness of behavior (*e.g.* gesture amplitude), we complemented it with another module, that we called the *Expressivity modulator*. Its role is to adjust the amplitude and intensity of gestures considering the attitude to express. To create this module, we used Ravenet *et al.*'s model [25], which we presented in Section 2.3. For an input attitude and communicative function, we can query Ravenet *et al.*'s model, obtaining gestures' amplitude and intensity parameters.

Behavior Realizer and ECA platform: Finally, the BML message enriched with expressive parameters is sent to the *Behavior realizer* of the Greta platform [7], which transforms it into an animation displayed by its virtual character.

With these components, we created a virtual recruiter able to express various attitudes through its non-verbal and verbal behavior. In the next section, we present an evaluation of our virtual recruiter, which aimed at assessing whether the attitudes it expresses are correctly identified by users.

5 Virtual recruiter evaluation

Our evaluation study pursued several objectives. The first is to validate whether our virtual recruiter can express input attitudes correctly. We also wanted to compare the respective contributions of the non-verbal and verbal modalities to the expression of attitudes, in order to verify previous results indicating that combining both modalities achieves the best results when expressing attitudes [69], [96]. Finally, we investigated whether the perception of attitudes is similar when participants watch videos of the virtual recruiter compared to when they directly interact with it. Indeed, many studies rely on video-based surveys for evaluating models of ECAs [16], [69], [95], [96], [98], while others evaluate these models in interactive settings [91], [94]. However, it has been shown that the level of interactivity can influence perception of ECAs [99]. With this study, we also wanted to investigate whether the perception of attitudes is affected by the level of interactivity. We defined the following research questions:

QR1: Are attitudes expressed by the virtual recruiter correctly identified?

QR2: What combination of modalities (verbal only, non-verbal only, multimodal) allow reaching the highest rate of attitude recognition?

QR3: Is there a difference in attitude recognition between participants interacting directly with the virtual recruiter and participants evaluating the recruiter through videos?

We now present the experimental setup we adopted to answer these three questions.

5.1 Experimental setup

5.1.1 Conditions and independent variables

We defined three independent variables: VI_1 represents the type of expressed attitude: *Dominant*, *Friendly* or *Unfriendly*. We did not consider submissiveness as a preliminary evaluation showed that our behavior planner did not succeed in expressing it [100]. The second variable VI_2 represents the combination of modalities used to express these attitudes: *Speech*, when only the dialogue model is used, *Non-verbal*, when only the behavior planner and expressivity modulator are used, *Multimodal* when both are used, and *Neutral* when none are used. The third variable VI_3 represents the evaluation mode: *Interaction* or *Video*.

One attitude and one evaluation mode were randomly assigned to each participant, *i.e.* one value of VI_1 and one value of VI_3. All participants were exposed to the four combinations of modalities, *i.e.* all values of VI_2, in a counter-balanced order.

In order to compare how attitudes were perceived in interaction or with videos, we built two separate evaluation platforms. The first platform, corresponding to the *Interaction* value of VI_3, consisted of a job interview simulation system, a simplified version of the Tardis platform, where participants interacted directly with our virtual recruiter. The second platform, for the *Video* value of VI_3, was an online web application where participants watched videos of the virtual job recruiter. In both cases, an identical questionnaire was used to assess the virtual recruiter's attitude.

5.1.2 Measures

In order to evaluate the perceived attitude, we chose to use two types of measures: direct scales indicating the perceived dominance and friendliness and adjective scales (inspired from [101]). A seven-point Likert scale was used for all these questions. We reproduce here the questions of the questionnaire:

The virtual recruiter behaves in a – – – – manner:

Q1 -	Not at all dominant	-	Absolutely dominant
Q2 -	Not at all friendly	-	Absolutely friendly

The recruiter is more:

Q3 -	Closed	-	Open
Q4 -	Cold	-	Warm
Q5 -	Shy	-	Confident
Q6 -	Unassuming	-	Bold

5.1.3 Interaction case

Study room - A study room was setup for the *Interaction* case. To improve the participant's level of immersion, we used a large screen allowing us to display the virtual recruiter at a life-size scale (Fig. 4). The 3D scene presented the ECA sitting behind a desk, aligned to its real life counterpart where the participant was sitting. The participant was fitted with a headset microphone, and a computer used to collect the previous measures (Section 5.1.2) during the experiment.

Interaction system - We used a simplified version of the Tardis job interview simulation platform: the SSI software was used in order to retrieve vocal events to drive the Listening Behavior Planner module. The scenario execution module was replaced with an *ad-hoc* scenario module following a linear scenario defined according to the protocol described in the next paragraphs. The affective core was not used, as attitudes and modalities were fixed by the independent variables.

We now give an overview of an evaluation session.

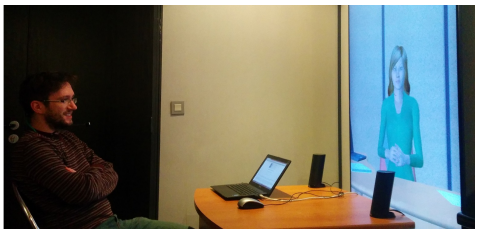


Fig. 4. The study room for the *Interaction* case.

Before the interaction - Participants started by reading and filling out a consent form, and filling out a small demographic questionnaire. They were informed that their participation was voluntary, that no data would be saved except from the questionnaires they would fill out, and that these would be anonymous. The participants were also informed that no speech recognition was performed. They were then given instructions which described the experiment, and were given a chance to ask questions before the experiment.

Habituation phase - Once the participant had finished filling out the questionnaires, we moved on to a habituation phase. The interaction was activated remotely, the participant being alone in the interaction room. The study screen, initially black, would fade to the virtual environment inhabited by a first ECA (Fig. 5, left) which repeated the interaction instructions. This phase gave the participant an opportunity to get used to the situation, and was also an opportunity to detect technical problems (*e.g.* unplugged microphone, system crash). Here, no attitude was expressed by the ECA. Once the instructions had been given, the screen faded to black and another opportunity was given to the participant to ask questions about the experiment.

Fig. 5. The two ECAs used in the *Interaction* case.

Evaluation phase - This phase consisted of a job interview simulation divided into four topics. A second ECA model was used so that participants would not be influenced by their impression of the previous character (Fig. 5, right). In each of these topics, the recruiter asked three questions to the participant on a particular theme. The first topic was about a general presentation of self. The second topic dealt with the participant's education and the third with professional experiences, while the last topic was about interpersonal skills. We chose these themes as they are commonly discussed in job interviews. Once the three questions of each topic had been answered by the participant, the screen faded to black and the participant would fill in a questionnaire with the questions described in the previous section. After the questionnaire had been filled in, a request was sent from the browser to start automatically the next of the 4 topics.

5.1.4 Video case

For collecting participants' data in the *Video* case, we built a web platform. A first web page gave instructions to the participants, then an automated script would assign randomly conditions to the participants. Since the participants in the *Video* case did not interact with the ECA, we did not include an habituation phase. Therefore, after reading the instructions and filling out a small demographics questionnaire, the participants were redirected to the evaluation phase.

For the *Video* case, the same four topics were used in the evaluation phase as in the *Interaction case*: participants saw a series of videos where the recruiter asked the same questions as in the *Interaction* case. After each set of 3 questions, the participants had to answer the same questionnaire as for the *Interaction* case.

5.2 Results

We recruited 48 participants in total (28F, 20M). 24 were assigned to the *Interaction* case, and 24 to the *Video* case. The participants were mostly of French nationality ($91,7\%$), and the mean age of participants was $34,4$ years old ($\sigma = 13,3$).

In order to analyze the rates of correct attitude identification, we transformed the raw Likert scales data as in table 3: indeed, a correct attitude recognition involved a $Q2$ score higher than 4 for friendliness and lower than 4 for unfriendliness. Therefore, we defined new measures, Q_{Dir} (Dir for *direct* measure), Q_{Adj1} and Q_{Adj2} (Adj for *adjective* scales), which values range between -3 and 3, a positive value indicating a correct attitude recognition.

	Q_{Dir}	Q_{Adj1}	Q_{Adj2}
Dom	$Q_{Dir} = Q1 - 4$	$Q_{Adj1} = Q5 - 4$	$Q_{Adj2} = Q6 - 4$
Frd	$Q_{Dir} = Q2 - 4$	$Q_{Adj1} = Q3 - 4$	$Q_{Adj2} = Q4 - 4$
Hos	$Q_{Dir} = 4 - Q2$	$Q_{Adj1} = 4 - Q3$	$Q_{Adj2} = 4 - Q4$

TABLE 3
Definition of Q_{Dir}, Q_{Adj1} and Q_{Adj2} for the different attitudes

5.2.1 QR1: Attitude recognition in the multimodal condition

We first checked whether attitudes expressed by our full virtual recruiter model (*i.e.* the *Multimodal* condition) were correctly recognized, independently of the evaluation mode. We realized unilateral Student test between the attitude scores (Q_{Dir}, Q_{Adj1} and Q_{Adj2}) and the scale mean ($\mu_0 = 0$).

When considering the three attitudes simultaneously, the Student tests were significant for all the measures Q_{Dir} ($\mu = 0.68, \sigma = 1.18, t(47) = 4.22, p = 0.000$), Q_{Adj1} ($\mu = 0.82, \sigma = 1.43, t(47) = 4.2, p = 0.000$) and Q_{Adj2} ($\mu = 0.95, \sigma = 1.28, t(47) = 5.38, p = 0.000$). We conclude that in general, attitudes are recognized.

Considering dominance alone, we also found that the Student tests were significant: Q_{Dir} ($\mu = 0.68, \sigma = 1.18, t(47) = 4.22, p = 0.000$), Q_{Adj1} ($\mu = 0.82, \sigma = 1.43, t(47) = 4.2, p = 0.000$), Q_{Adj2} ($\mu = 0.95, \sigma = 1.28, t(47) = 5.38, p = 0.000$).

For unfriendliness, tests on Q_{Dir} ($\mu = 0.16, \sigma = 1.20, t(15) = 1.09, p = 0.29$) and Q_{Adj1} ($\mu = 0.10, \sigma = 1.42, t(15) = 0.89, p = 0.39$) were not significant, but testing Q_{Adj2} was ($\mu = 0.625, \sigma = 1.13, t(15) = 2.57, p = 0.02$).

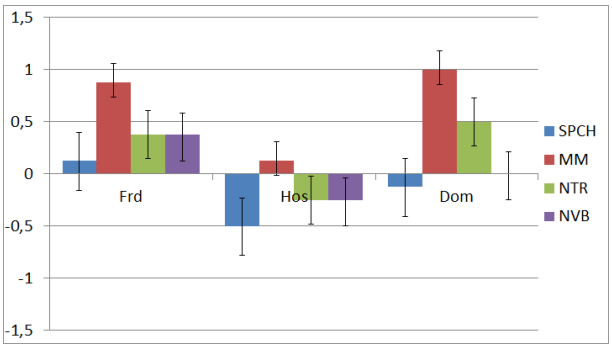


Fig. 6. Score Q_{Dir} per condition for all participants. $SPCH$ = Verbal, MM = Multimodal, NTR = None (control), NVB = Non-verbal. Frd = Friendliness, Hos = Unfriendliness, Dom = Dominance.

While we cannot conclude that unfriendliness was recognized, participants did see unfriendly recruiters as colder.

Finally, considering only friendliness, Student tests on the three measures were significant: Q_{Dir} ($\mu = 0.85, \sigma = 1.25, t(15) = 3.08, p = 0.008$), Q_{Adj1} ($\mu = 0.92, \sigma = 1.42, t(15) = 2.92, p = 0.01$), Q_{Adj2} ($\mu = 0.77, \sigma = 1.13, t(15) = 3.08, p = 0.008$).

5.2.2 QR2: Comparing combination of modalities

We then investigated which combination of modalities led to the best attitude recognition scores. We conducted ANOVA tests between the four different combinations of modalities (*None, Speech, Non-verbal, Multimodal*) on the combined recognition scores of all attitudes and evaluation modes for the three measures Q_{Dir}, Q_{Adj1} and Q_{Adj2}.

For Q_{Adj1} ($F(3, 188) = 2.28, p = 0.08$) and Q_{Adj2} ($F(3, 188) = 2.36, p = 0.07$), results approached the $p < 0.05$ significance threshold. Though this result is not significant, we did notice a trend towards better scores for the *Multimodal* case (in the *Multimodal* case, Q_{Adj1} was on average 0.56 higher and Q_{Adj2} was on average 0.46 higher than in the other cases).

Significance was achieved when testing the Q_{Dir} measure ($F(3, 188) = 4.69, p = 0.003$). In that case, the *Multimodal* combination achieved the highest score (on average, Q_{Dir} was higher of 0.71 compared to the other combinations). Friendliness ($\mu = 0.875$) and dominance ($\mu = 1$) attitudes achieved better recognition scores than unfriendliness ($\mu = 0.1875$), however, the *Multimodal* condition still outperformed the other conditions in that case (Fig. 6).

5.2.3 QR3: Evaluation model

Lastly, we compared attitude recognition scores between interaction models (*Interaction* or *Video*). We regrouped measures independently of the expressed attitude and the combination of modalities (excluding the *None* condition) and realized an ANOVA for the three measures Q_{Dir}, Q_{Adj1} and Q_{Adj2}. Significance was observed for Q_{Dir} ($F(1, 142) = 4.30, p = 0.039$), where higher recognition was achieved in the *Video* case (on average, Q_{Dir} was higher of 0.42). Tests on Q_{Adj1} and Q_{Adj2} did not achieve significance.

We wanted to compare the observed effect on Q_{Dir} of modalities combinations used to express attitudes. To this end, we realized four new ANOVA tests between the different modalities combinations, separating the *Interaction* case and the *Video* case. A significant difference was only observed for the *Speech* condition ($F(1, 46) = 8.59, p = 0.005$), where the difference between recognition scores was much higher in the *Video* case than in the *Interaction* case (on average, Q_{Dir} was 0.89 higher in the *Video* case).

5.3 Discussion

We observed that dominance and friendliness were overall recognized using the full model ($QR1$), while unfriendliness achieved only mixed results (only Q_{Adj2} achieves significance). However, results of $QR2$ indicated the the recruiter expressing unfriendliness in the *Multimodal* condition is seen as much more hostile than with other combinations of modalities. Therefore, we conclude that our virtual recruiter model can express the three attitudes we considered.

As per $QR2$, we observed that combining non-verbal and speech modalities led to better recognition scores, for the three considered attitudes. These results confirmed previous studies' results that observed that combining congruent attitudes expression from the verbal and non-verbal modalities led to better recognition scores [69], [96].

Finally, we observed for $QR3$ that recognition scores were better when participants watched videos than when they directly interacted with the model, this effect being particularly strong when attitudes were only expressed by the dialogue model. This result tends to confirm that the evaluation mode of an ECA model can influence the obtained results [99]. This may have happened because participants interacting with the recruiter experiencde a higher cognitive load than participants watching videos, and could not focus as much on the recruiter's utterances. However, our virtual recruiter model was limited as it did not consider the expression of attitudes through listening and turn-taking behavior, which could have influenced the results. Indeed, the way turn-taking and backchannels are realized do influence the perception of attitudes [98]. As the listening behavior was not observed by participants evaluating videos, this could have affected the results, even if the listening behavior of the recruiter was designed to be neutral.

6 Conclusion

Behavior planning models for ECAs typically consider non-verbal signals in isolation from one another. However, it has been shown that the meaning of non-verbal signals can be altered by surrounding signals [2], [21], [22]. Moreover, social attitudes are not expressed at one given time by prototypical displays of multimodal signals: they affect behavior on longer time spans [1].

We proposed a methodology for building behavior planning models that consider the sequentiality of non-verbal signals, which we applied to the design of a virtual recruiter expressing social attitudes. This method starts with the collection of an appropriate multimodal corpus, which is annotated at two levels. Firstly, non-verbal behavior is annotated with a coding scheme adapted to the task. Secondly, attitudes are annotated with a continuous paradigm. Then, automatic extraction of sequences of non-verbal signals from the corpus is realized using a sequence mining

technique. The extracted sequences are used to build a computational model of behavior planning for the expression of attitudes. This methodology could be applied to other socio-emotional affects that can be expressed under a continuous representation. In such a case, only the first step of our methodology, the multimodal corpus acquisition and annotation, would have to be adapted to fit the studied phenomenon. The two other steps, mining the multimodal corpus and building a computational behavior planning, would remain similar.

We applied our methodology to the design of a virtual recruiter in the context of a broader effort to build a job interview simulation platform in the Tardis project. Our model was complemented with various modules in order to build a complete virtual recruiter. We evaluated the expression of attitudes by our model in two contexts, in a direct interaction and with videos. This evaluation study confirmed that attitudes were correctly recognized by participants, and that our methodology, mining multimodal corpora annotated with non-verbal signals and continuous traces of socio-emotional phenomena for sequences of relevant non-verbal signals, was successful for building a behavior planning model in the case of social attitudes.

Our methodology still holds some limitations; in particular, we do not consider temporal information, such as the duration of particular non-verbal signals or the duration between two signals of a particular sequence. Moreover, while our representation of attitudes is bi-dimensional, our behavior planning model only considers a single dimension at a time when generating sequences. However, we believe that it could be naturally extended to multiple dimensions, by adapting the last step of our behavior planning model to consider multiple dimensions in the computation of a sequence's score.

ACKNOWLEDGMENTS

This research has been partially supported by the European Community Seventh Framework Program (FP7/2007-2013), under grant agreement no. 288578 (TARDIS).

REFERENCES

[1] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 693–727, 2005.

[2] D. Keltner, "Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame," *Journal of Personality and Social Psychology*, vol. 68, pp. 441–454, 1995.

[3] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Transactions in Computer-Human Interaction*, vol. 12, no. 2, pp. 293–327, 2005.

[4] H. C. Lane, M. J. Hays, M. G. Core, and . Auerbach, "Learning intercultural communication skills with virtual humans: Feedback and fidelity," *Journal of Educational Psychology*, vol. 105, no. 4, pp. 1026–1035, 2013.

[5] R. Beale and C. Creed, "Affective interaction: How emotional agents affect users," *International Journal of Human-Computer Studies*, vol. 67, no. 9, pp. 755–776, 2009.

[6] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: the behavior expression animation toolkit," in *Life-Like Characters*. Springer, 2004, pp. 163–185.

[7] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud, "Greta: an interactive expressive eca system," in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*. International Foundation for Autonomous Agents and Multiagent Systems, 2009, pp. 1399–1400.

[8] A. Cafaro, "First impressions in human-agent virtual encounters," Ph.D. dissertation, Reykjavik University, Iceland, 2014.

[9] P. Ekman and V. Friesen, *Pictures of Facial Affect*. Palo Alto: Consulting Psychologists Press, 1976.

[10] A. Kendon, "Some functions of gaze-direction in social interaction." *Acta psychologica*, vol. 26, no. 1, pp. 22–63, 1967.

[11] I. Poggi, "Mind markers," in *Gestures. Meaning and use.*, M. R. N. Trigo and I. Poggi, Eds. University Fernando Pessoa Press, Oporto., 2003.

[12] M. Knapp, J. Hall, and T. Horgan, *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[13] J.-C. Martin, S. Abrilian, L. Devillers, M. Lamolle, M. Mancini, and C. Pelachaud, "Du corpus vidéo à l'agent expressif : Utilisation des différents niveaux de représentation multimodale et émotionnelle," *Revue d'intelligence artificielle*, vol. 20, no. 4-5, pp. 477–498, 2006.

[14] R. Niewiadomski, S. J. Hyniewska, and C. Pelachaud, "Constraint-based model for synthesis of multimodal sequential expressions of emotions," *IEEE Transaction on Affective Computing*, vol. 2, no. 3, pp. 134–146, 2011.

[15] Z. Deng and U. Neumann, *Data-Driven 3D Facial Animation*. Springer, 2008.

[16] Y. Ding, K. Prepin, J. Huang, C. Pelachaud, and T. Artières, "Laughter animation synthesis," in *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 773–780.

[17] P. Ekman, "The directed facial action task," in *Handbook of emotion elicitation and assessment*. Oxford University Press Oxford, UK, 2007, pp. 47–53.

[18] U. Hess, R. B. Adams Jr, and R. E. Kleck, "Looking at you or looking elsewhere: The influence of head orientation on the signal value of emotional facial expressions," *Motivation and Emotion*, vol. 31, no. 2, pp. 137–144, 2007.

[19] C. Pelachaud, N. I. Badler, and M. Steedman, "Generating facial expressions for speech," *Cognitive science*, vol. 20, no. 1, pp. 1–46, 1996.

[20] I. S. Pandzic and R. Forchheimer, *MPEG-4 facial animation*. John Wiley & Sons, 2002.

[21] S. With and W. S. Kaiser, "Sequential patterning of facial actions in the production and perception of emotional expressions," *Swiss Journal of Psychology*, vol. 70, no. 4, pp. 241–252, 2011.

[22] R. E. Jack, O. G. Garrod, and P. G. Schyns, "Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time," *Current Biology*, vol. 24, no. 2, pp. 187–192, 2014.

[23] D. Ballin, M. Gillies, and I. Crabtree, "A framework for interpersonal attitude and non-verbal communication in improvisational visual media production," in *Proceedings of the 1st European Conference on Visual Media Production*, 2004.

[24] J. Lee and S. Marsella, "Modeling side participants and bystanders: The importance of being a laugh track," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, H. Vilhjlmsson, S. Kopp, S. Marsella, and K. Thrisson, Eds. Springer Berlin Heidelberg, 2011, vol. 6895, pp. 240–247.

[25] B. Ravenet, M. Ochs, and C. Pelachaud, "From a user-created corpus of virtual agents non-verbal behavior to a computational model of interpersonal attitudes," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira, Eds. Springer Berlin Heidelberg, 2013, vol. 8108, pp. 263–274.

[26] J. W. Du Bois, "The stance triangle," in *Stancetaking in discourse: Subjectivity, evaluation, interaction*, R. Englebretson, Ed. John Benjamins Amsterdam, The Netherlands, and Philadelphia, PA, 2007, pp. 139–182.

[27] M. Chindamo, J. Allwood, and E. Ahlsén, "Some suggestions for the study of stance in communication," in *Proceedings of the 4th ASE/IEEE International Conference on Social Computing*. IEEE, 2012, pp. 617–622.

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

[29] T. Leary, *Interpersonal diagnosis of personality*. New York: Ronald, 1957.

[30] W. C. Schutz, *FIRO: A three-dimensional theory of interpersonal behavior.* Oxford, England: Rinehart, 1958.

[31] J. K. Burgoon and J. L. Hale, "The fundamental topoi of relational communication," *Communication Monographs*, vol. 51, no. 3, pp. 193–214, 1984.

[32] K. J. Tusing and J. P. Dillard, "The sounds of dominance," *Human Communication Research*, vol. 26, no. 1, pp. 148–171, 2000.

[33] K. R. Scherer and U. Scherer, "Speech behavior and personality," in *Speech evaluation in psychiatry*, J. Darby, Ed. Grune & Stratton, New York, USA, 1981, pp. 115–135.

[34] C. R. Glass, T. V. Merluzzi, J. L. Biever, and K. H. Larsen, "Cognitive assessment of social anxiety: Development and validation of a self-statement questionnaire," *Cognitive Therapy and Research*, vol. 6, no. 1, pp. 37–55, 1982.

[35] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[36] D. Keltner and J. Haidt, "Social functions of emotions at four levels of analysis," *Cognition & Emotion*, vol. 13, no. 5, pp. 505–521, 1999.

[37] L. Smith-Lovin and C. Brody, "Interruptions in group discussions: The effects of gender and group composition," *American Sociological Review*, vol. 54, pp. 424–435, 1989.

[38] R. Gifford and D. W. Hine, "The role of verbal behavior in the encoding and decoding of interpersonal dispositions," *Journal of Research in Personality*, vol. 28, no. 2, pp. 115–132, 1994.

[39] M. S. Mast, "Dominance as expressed and inferred through speaking time," *Human Communication Research*, vol. 28, no. 3, pp. 420–450, 2002.

[40] N. E. Dunbar and J. K. Burgoon, "Perceptions of power and interactional dominance in interpersonal relationships," *Journal of Social and Personal Relationships*, vol. 22, no. 2, pp. 207–233, 2005.

[41] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[42] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992.

[43] R. Gifford, "A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior." *Journal of Personality and Social Psychology*, vol. 66, no. 2, p. 398, 1994.

[44] D. R. Carney, J. A. Hall, and L. S. LeBeau, "Beliefs about the nonverbal expression of social power," *Journal of Nonverbal Behavior*, vol. 29, no. 2, pp. 105–123, 2005.

[45] J. K. Burgoon, D. B. Buller, J. L. Hale, and M. A. de Turck, "Relational Messages Associated with Nonverbal Behaviors," *Human Communication Research*, vol. 10, no. 3, pp. 351–378, 1984.

[46] Y. Yabar and U. Hess, "Display of empathy and perception of out-group members," *New Zealand Journal of Psychology*, vol. 36, no. 1, p. 42, 2007.

[47] J. K. Burgoon and B. A. Le Poire, "Nonverbal cues and interpersonal judgments: Participant and observer perceptions of intimacy, dominance, composure, and formality," *Communication Monographs*, vol. 66, no. 2, pp. 105–124, 1999.

[48] V. Richmond and J. McCroskey, *Nonverbal Behavior in Interpersonal Relations*. Allyn and Bacon, 2000.

[49] M. LaFrance, "Posture mirroring and rapport," in *Interaction Rhythms: Periodicity in Communicative Behavior*, M. Davis, Ed. New York: Human Sciences Press, 1982, pp. 279–299.

[50] A. Mehrabian, *Nonverbal communication*. Transaction Publishers, 1977.

[51] M. Argyle and M. Cook, *Gaze and Mutual Gaze*. Cambridge University Press, 1976.

[52] N. J. Briton and J. A. Hall, "Beliefs about female and male nonverbal communication," *Sex Roles*, vol. 32, no. 1-2, pp. 79–90, 1995.

[53] D. Heylen, "Listening heads," in *Modeling Communication with Robots and Virtual Humans*, ser. Lecture Notes in Computer Science, I. Wachsmuth and G. Knoblich, Eds. Springer Berlin Heidelberg, 2008, vol. 4930, pp. 241–259.

[54] A. Mignault and A. Chaudhuri, "The many faces of a neutral face: Head tilt and perception of dominance and emotion," *Journal of Nonverbal Behavior*, vol. 27, no. 2, pp. 111–132, 2003.

[55] R. Gifford, "Mapping nonverbal behavior on the interpersonal circle." *Journal of Personality and Social Psychology*, vol. 61, no. 2, p. 279, 1991.

[56] E. Otta, B. B. P. Lira, N. M. Delevati, O. P. Cesar, and C. S. G. Pires, "The effect of smiling and of head tilting on person perception," *The Journal of psychology*, vol. 128, no. 3, pp. 323–331, 1994.

[57] C. Debras and A. Cienki, "Some uses of head tilts and shoulder shrugs during human interaction, and their relation to stancetaking," in *Proceedings of the 4th ASE/IEEE International Conference on Social Computing*. IEEE Computer Society, Sept 2012, pp. 932–937.

[58] I. Poggi and C. Pelachaud, "Performative faces," *Speech Communication*, vol. 26, no. 12, pp. 5–21, 1998.

[59] M. Costa, M. Menzani, and P. E. R. Bitti, "Head canting in paintings: An historical study," *Journal of Nonverbal Behavior*, vol. 25, no. 1, pp. 63–73, 2001.

[60] J. Harrigan, T. Oxman, and R. Rosenthal, "Rapport expressed through nonverbal behavior," *Journal of Nonverbal Behavior*, vol. 9, no. 2, pp. 95–110, 1985.

[61] A. M. Von der Puetten, N. C. Krämer, J. Gratch, and S.-H. Kang, ""It doesnt matter what you are!": Explaining social effects of agents and avatars," *Computers in Human Behavior*, vol. 26, no. 6, pp. 1641–1650, 2010.

[62] C. Darwin, *The Expression of the Emotions in Man and Animals*. Harper Perennial, 1872.

[63] P. Ekman and W. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage and coding," *Semiotica*, vol. 1, no. 1, pp. 49–98, 1969.

[64] B. Knutson, "Facial expressions of emotion influence interpersonal trait inferences," *Journal of Nonverbal Behavior*, vol. 20, no. 3, pp. 165–182, 1996.

[65] C. F. Keating, A. Mazur, M. H. Segall, P. G. Cysneiros, J. E. Kilbride, P. Leahy, W. T. Divale, S. Komin, B. Thurman, and R. Wirsing, "Culture and the perception of social dominance from facial expression." *Journal of personality and social psychology*, vol. 40, no. 4, p. 615, 1981.

[66] U. Hess, S. Blairy, and R. E. Kleck, "The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation," *Journal of Nonverbal behavior*, vol. 24, no. 4, pp. 265–283, 2000.

[67] E. Krumhuber, A. Manstead, and A. Kappas, "Temporal aspects of facial displays in person and expression perception: The effects of smile dynamics, head-tilt, and gender," *Journal of Nonverbal Behavior*, vol. 31, no. 1, pp. 39–56, 2007.

[68] J. Aronoff, A. M. Barclay, and L. A. Stevenson, "The recognition of threatening facial stimuli." *Journal of personality and social psychology*, vol. 54, no. 4, pp. 647–655, 1988.

[69] Z. Callejas, B. Ravenet, M. Ochs, and C. Pelachaud, "A computational model of social attitudes for a virtual recruiter," in *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 93–100.

[70] F. Bonin, R. Bock, and N. Campbell, "How do we react to context? annotation of individual and group engagement in a video corpus," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, Sept 2012, pp. 899–903.

[71] R. Cowie, G. McKeown, and E. Douglas-Cowie, "Tracing emotion: an overview," *International Journal of Synthetic Emotions*, vol. 3, no. 1, pp. 1–17, 2012.

[72] J. Allwood, "Multimodal corpora," in *Corpus linguistics. An international handbook*, A. Lüdeling and M. Kytö, Eds. Berlin: Mouton de Gruyter, 2008, pp. 207–225.

[73] H. Gunes and M. Piccardi, "Creating and annotating affect databases from face and body display: A contemporary survey," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, 2006, pp. 2426–2433.

[74] D. Knight, "The future of multimodal corpora," *Revista Brasileira de Linguística Aplicada*, vol. 11, no. 2, pp. 391–415, 2011.

[75] L. Hunyadi, I. Szekrenyes, A. Borbely, and H. Kiss, "Annotation of spoken syntax in relation to prosody and multimodal pragmatics," in *3rd International Conference on Cognitive Infocommunications*. IEEE, Dec 2012, pp. 537–541.

[76] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9-10, pp. 341–345, 2001.

[77] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *Proceedings of the 5th Language Resources and Evaluation Conference*. Language Resources and Evaluation, 2006, pp. 1556–1559.

[78] M. Chollet, M. Ochs, C. Clavel, and C. Pelachaud, "A multimodal corpus approach to the design of virtual recruiters," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 19–24.

[79] R. Cowie, E. Douglas-Cowie, M. McRorie, I. Sneddon, L. Devillers, and N. Amir, "Issues in data collection," in *Emotion-Oriented Systems*, P. Petta, C. Pelachaud, and R. Cowie, Eds. Springer-Verlag, Berlin, Heidelberg, 2011, pp. 197–212.

[80] P. G. Ferreira and P. J. Azevedo, "Protein sequence classification through relevant sequence mining and bayes classifiers," in *Progress in Artificial Intelligence*, ser. Lecture Notes in Computer Science, C. Bento, A. Cardoso, and G. Dias, Eds. Springer Berlin Heidelberg, 2005, vol. 3808, pp. 236–247.

[81] H. P. Martínez and G. N. Yannakakis, "Mining multimodal sequential patterns: a case study on affect detection," in *Proceedings of the 13th International Conference on Multimodal Interfaces*. New York, NY, USA: ACM, 2011, pp. 3–10.

[82] D. Fricker, H. Zhang, and C. Yu, "Sequential pattern mining of multimodal data streams in dyadic interactions," in *IEEE International Conference on Development and Learning*, 2011, vol. 2, pp. 1–6.

[83] R. Srikant and R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements," *Advances in Database Technology*, vol. 1057, pp. 1–17, 1996.

[84] M. Guillame-Bert and J. L. Crowley, "Learning temporal association rules on symbolic time sequences." in *Proceedings of the 4th Asian Conference on Machine Learning*, 2012, pp. 159–174.

[85] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499.

[86] H. Vilhjálmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. Marshall, C. Pelachaud, Z. Ruttkay, K. Thrisson, H. van Welbergen, and R. van der Werf, "The behavior markup language: Recent developments and challenges," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, Eds. Springer Berlin Heidelberg, 2007, vol. 4722, pp. 99–111.

[87] D. Heylen, S. Kopp, S. Marsella, C. Pelachaud, and H. Vilhjlmsson, "The next step towards a function markup language," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, H. Prendinger, J. Lester, and M. Ishizuka, Eds. Springer Berlin Heidelberg, 2008, vol. 5208, pp. 270–280.

[88] M. Mancini and C. Pelachaud, "Dynamic behavior qualifiers for conversational agents," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, C. Pelachaud, J.-C. Martin, E. Andr, G. Chollet, K. Karpouzis, and D. Pel, Eds. Springer Berlin Heidelberg, 2007, vol. 4722, pp. 112–124.

[89] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Exploration Newsletter*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[90] S. Jaillet, A. Laurent, and M. Teisseire, "Sequential patterns for text categorization," *Intelligent Data Analysis*, vol. 10, no. 3, pp. 199–214, 2006.

[91] P. Gebhard, T. Baur, I. Damian, G. Mehlmann, J. Wagner, and E. André, "Exploring interaction strategies for virtual characters to induce stress in simulated job interviews," in *Proceedings of the 13th International Conference on Autonomous Agents and Multi-agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 661–668.

[92] 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*. ACM, 2013, pp. 831–834.

[93] A. Ben Youssef, M. Chollet, H. Jones, N. Sabouret, C. Pelachaud, and M. Ochs, "Towards a socially adaptive virtual agent," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, W.-P. Brinkman, J. Broekens, and D. Heylen, Eds. Springer International Publishing, 2015, vol. 9238, pp. 3–16.

[94] J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. van der Werf, and L.-P. Morency, "Virtual rapport," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, J. Gratch, M. Young, R. Aylett, D. Ballin, and P. Olivier, Eds. Springer Berlin Heidelberg, 2006, vol. 4133, pp. 14–27.

[95] 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.

[96] N. Bee, C. Pollock, E. André, and M. Walker, "Bossy or Wimpy: Expressing Social Dominance by Combining Gaze and Linguistic Behaviors," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, and A. Safonova, Eds. Springer Berlin Heidelberg, 2010, vol. 6356, ch. 28, pp. 265–271.

[97] J. Linssen, M. Theune, and D. Heylen, "Taking things at face value: How stance informs politeness of virtual agents," in *Proceedings of the Workshop on Computers As Social Actors*, ser. CEUR Workshop Proceedings, M. Conci, V. Dignum, M. Funk, and D. Heylen, Eds., 2014, vol. 1119, pp. 71–82.

[98] M. ter Maat, K. Truong, and D. Heylen, "How turn-taking strategies influence users impressions of an agent," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, and A. Safonova, Eds. Springer Berlin Heidelberg, 2010, vol. 6356, pp. 441–453.

[99] B. Weiss, C. Khnel, I. Wechsung, S. Fagel, and S. Mller, "Quality of talking heads in different interaction and media contexts," *Speech Communication*, vol. 52, no. 6, pp. 481 – 492, 2010, speech and Face-to-Face Communication.

[100] M. Chollet, M. Ochs, and C. Pelachaud, "From non-verbal signals sequence mining to bayesian networks for interpersonal attitudes expression," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, T. Bickmore, S. Marsella, and C. Sidner, Eds. Springer International Publishing, 2014, vol. 8637, pp. 120–133.

[101] A. Fukayama, T. Ohno, N. Mukawa, M. Sawaki, and N. Hagita, "Messages embedded in gaze of interface agents - impression management with agent's gaze," in *Proceedings of the 2002 SIGCHI conference on Human factors in computing systems*. New York, New York, USA: ACM Press, 2002, pp. 41–48.

Mathieu Chollet is a post-doctoral scholar at the Institute for Creative Technologies, University of Southern California. His research interests include embodied conversational agents and multimodal interfaces, particularly in the context of social skills training.

Magalie Ochs is a computer scientist, associate professor at Aix-Marseille University, LSIS laboratory. Her research interests include the modeling of emotional and social dimensions in virtual agents, the corpus based analysis of socio-emotional phenomena, and the evaluation of human-virtual agent interaction.

Catherine Pelachaud is director of research at CNRS in the LTCI laboratory, Telecom ParisTech. Her research interest includes embodied conversational agent, nonverbal communication (face, gaze, and gesture), expressive behaviors and socio-emotional agents.