An Architecture for a Socially Adaptive Virtual Recruiter in Job Interview Simulations

Atef Ben-Youssef
Univ. Paris-Sud,
LIMSI-CNRS UPR 3251
atef.ben_youssef@limsi.fr

Mathieu Chollet
Telecom-ParisTech
mathieu.chollet@telecom-paristech.fr

Hazaël Jones
SupAgro, UMR ITAP
hazael.jones@supagro.fr

Nicolas Sabouret
Univ. Paris-Sud,
LIMSI-CNRS UPR 3251
nicolas.sabouret@limsi.fr

Catherine Pelachaud
Telecom-ParisTech
catherine.pelachaud@telecom-paristech.fr

Magalie Ochs
Telecom-ParisTech
magalie.ochs@telecom-paristech.fr

ABSTRACT

This paper presents an architecture for an adaptive virtual recruiter in the context of job interview simulation. This architecture allows the virtual agent to adapt its behaviour according to social constructs (*e.g.* attitude, relationship) that are updated depending on the behaviour of their interlocutor. During the whole interaction, the system analyses the behaviour of the human participant, builds and updates mental states of the virtual agent and adapts its social attitude expression. This adaptation mechanism can be applied to a wide spectrum of application domains in Digital Inclusion, where the user need to train social skills with a virtual peer.

Author Keywords

Social Attitudes, Emotions, Affective computing, Virtual Agent, Non-verbal behaviour, adaptation

ACM Classification Keywords

I.6.5 Computing Methodologies: Simulation and Modelling—*Model Development*

INTRODUCTION

Youth unemployment is a significant problem in Europe, with over 22% of young people who are not in employment, education or training (NEETs)[1]. Several inclusion associations over Europe address this problem through specific councelling activities toward youngsters, including job interview simulations. These simulations have a strong impact of on the applicant's self-confidence and presentation in real interviews. However, this is an expensive and time-consuming approach that relies on the availability of trained practitioners as well as a willingness of the young people to discuss their strengths and weaknesses in front of practitioners and often also in front of their peers.

[1] `ec.europa.eu/eurostat`

Digital games offer a promising way of supporting the training and coaching of young people, providing them with a safe and private environment in which they can practice their skills repeatedly. TARDIS[2] [3] is a project funded by the FP7, whose aim is to build a scenario-based serious-game simulation platform that supports social training and coaching in the context of job interviews. It relies on the use of virtual agents that play the role of the virtual recruiter, and post-interview debriefing sessions with a concellor.

Intelligent virtual agents have the ability to simulate and express affects [22, 34] in dyadic interactions. However, the connection of these expressions with the interlocutor's verbal and non-verbal behaviour remains an open challenge. This results into what we can call "scripted" agents whose behaviour does not change in function of the human user's reactions. Our research motivation is that to be credible and to be able to build a relationship, virtual agents should adapt their affective behaviour to the human's [35].

The goal of this paper is to present the TARDIS architecture to adapt the virtual agent's behaviour to the attitude and emotions that are expressed by the user and detected by the system. We show how the system builds affective states and turns them into attitude expressions to enhance the interview simulation. Specifically, the TARDIS game relies on real-time social cue recognition, communicative performance computation and affective computation/decision making by the virtual recruiter. Building a credible job interview simulation involving a reactive virtual agent requires real-time information about the youngster in order to allow for an assessment of the appropriateness of the youngster's reactions and their communicative performance. Such assessment is done through the perception of a number of relevant social cues that are interpreted in term of performance, relative to the social expectations that are associated with a given situation. This assessment allows the TARDIS system to compute both an affective reaction for the virtual recruiter and the future steps in the interaction dialogue.

The next section presents a brief overview of related work. We then present the two main components of our architecture that have been implemented in the TARDIS system: the

[2] `http://www.tardis-project.eu`

affective reasoner and the behaviour planner. The methodology used to develop our model combined a theoretical and an empirical approach. Indeed, the model is based both on the literature on social attitudes but also on the analysis of an audiovisual corpus of job interviews and on post-hoc interviews with the recruiters on their expressed attitudes during the job interview.

RELATED WORKS AND THEORETICAL BACKGROUND

Our work refers to four different domains, as it involves the computation of agent's affects in real-time, with social adaptation, in the context of training.

Attitudes: A common representation for attitudes is Argyle's Interpersonal Circumplex [4], a bi-dimensional representation with a *Liking* dimension and *Dominance* dimension. Most modalities are involved in the expression of attitude : gaze, head orientation, postures, facial expressions, gestures, head movements [5, 11, 8]. For instance, wider gestures are signs of a dominant attitude [11], while smiles are signs of friendliness [8]. Models of attitude expression for embodied conversational agents have been proposed in the recent years. Ballin *et al.* propose a model that adapts posture, gesture and gaze behaviour depending on the agents' social relations [6]. Cafaro *et al.* study how users perceive attitudes and personality in the first seconds of their encounter with an agent depending on the agent's behaviour [10]. Ravenet *et al.* propose a model for expressing an attitude as the same time as a communicative intention [29].

Social training background: The idea of using virtual characters for training has gained much attention in the recent years: an early example is the pedagogical agent Steve [19], which was limited to demonstrating skills to students and answering questions. Since then, virtual characters have also been used for training social skills such as public speaking training [12] or job interview training [16]. The MACH system [16] is a system for job interview training. However, while the recruiter used in MACH can mimic behaviour and display back-channels, it does not reason on the user's affects and does not adapt to them.

Agents with real-time reaction to affects: Several agent models have proposed to process users' behaviours to infer affects in real-time. Acosta and Ward proposed a spoken dialogue system capable of inferring the user's emotion from vocal features and to adapt its response according to this emotion [1], which leads to a better perceived *rapport*. Prendinger and Ishizuka presented the Empathic Companion [28], which is capable to use users' physiological signals to interpret their affective state and to produce empathic feedback and reduce the stress of participants. The Semaine project [30] introduced Sensitive Artificial Listeners, *i.e.* virtual characters with different personalities that induce emotions in their interlocutors by producing emotionally coloured feedback. Audio and visual cues are used to infer users' emotions, which are then used to tailor the agent's non-verbal behaviour and next utterance. However the agents' behaviour is restricted to the face, and the agent mainly produces backchannels (*i.e.* listening behaviour). The SimSensei virtual human interviewer is designed to handle a conversation with users about psychological distress [14]. The agent can react to the user's behaviours as well as some higher level affective cues (*e.g.* arousal). However these cues only affect some of the dialogue choices of the agent and its reactive behaviour: no further adaptation is performed. We try to overcome these limitations in our own work.

Social adaptation: Buschmeier and Kopp [9] describe a model for adapting the dialogue model of an embodied conversational agent according to the common ground between a human and the agent. They analyse feedback signals from the human to determine if symbols used in the dialogue are understood or misunderstood by the human. They then propose strategies to co-construct *common ground*, that is to negotiate symbol meaning with the human so that they both share the same understanding of what the symbol means.

The relational agent Laura introduced by Bickmore [7] is a fitness coach agent designed to interact with users over a long period of time. They use the amount of times the users interact with Laura as a measure of friendliness, and as the friendliness grows, Laura uses more and more friendly behaviours (*e.g.* smiles, nods, more gestures...).

In these two examples, the agent adapts to the user to build rapport. However, the adaptation is only at the verbal level and the agent do not adapt in real time to the non-verbal behaviour of the interlocutor.

Yang et al [40] analysed the adaptation of a participant body language behaviour to the multi-modal cues of the interlocutor, under two dyadic interaction stances: friendly and conflictive. They combined Gaussian Mixture Model (GMM) based statistical mapping with Fisher kernels to automatically predict body language behaviour from the multi-modal speech and gesture cues of the interlocutor. They suggest a significant level of predictability of body language behaviour from interlocutor cues and they found that people with friendly attitudes tend to adapt more their body language to the behaviour of their interlocutors. This work clearly shows a path to social adaptation, but the mechanism is not based on reasoning on the interlocutor's performance: it is based on statistical mapping.

As a conclusion, progress has already been achieved in these different domains: from the computation of agent's affects in real-time, to social adaptation in the context of training. To our knowledge there is no system that combines all these aspects in a virtual agent.

ADAPTIVE VIRTUAL AGENT

Figure 1 presents the architecture of our adaptive agent. The context of the interaction is a one-to-one dialogue in which the virtual agent is leading the discussion, by asking questions or proposing conversation topics to the human participant (interviewee), and reacts in real-time to his/her non-verbal behaviour. The agent has no understanding of the answer's verbal content: we only focus on the non-verbal behaviour of the participant. We follow a linear three-steps architecture. First, a performance index (PI) of interviewee's

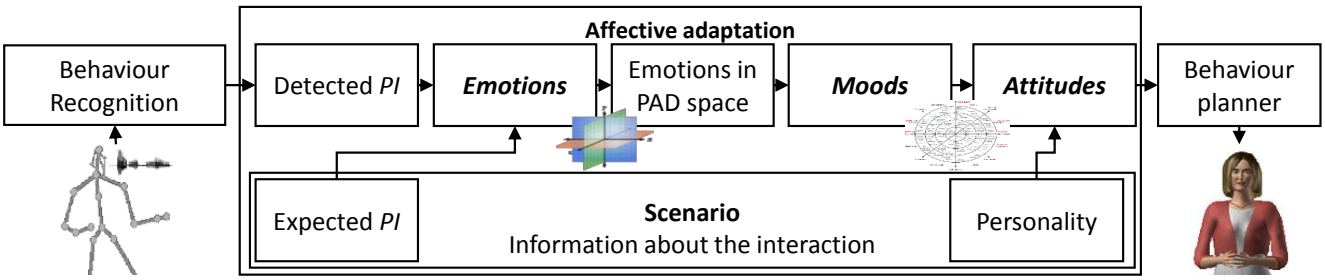


Figure 1: Overview of our virtual agent architecture based on affects adaptation to user's behaviour

non-verbal responses is computed using the Social Signal Interpretation (SSI) software [37], based on both body language and speech's non-verbal content. Second, the performance index is passed to the affective module which computes an emotion and adapts the attitude of the virtual agent toward the interviewee. Last, the behaviour planner selects the appropriate social signals to express the emotion and the attitude of the agent (*e.g.* eyebrow raise, smile, speed of gesture...). The rendering of this behaviour is based on the Greta/Semaine platform [27].

In this architecture, the agent is capable to react to the participant's behaviour almost in real time. When it speaks, the agent continuously expresses its affective state (emotions and social attitude) through its non-verbal behaviours through head movements, facial expressions and body movements. When the participant is answering, it simply uses back-channel behaviour (head tilting, smiling, *etc*) that do not depend on the affective state. Meanwhile, a new PI is computed for the next discussion turn. Concretely, a PI is computed everytime the human participant stops speaking, which generally occurs when it finishes its answer to the agent's utterance. Based on this PI, the affective response (emotion) and change in the attitude are computed and they will be expressed during the following utterance by the agent.

In this paper, we present the details of the affective adaptation and behaviour planning, based on this PI.

Social Signal Interpretation

The performance interpretation consists in evaluating the interviewee's body language and speech during his/her answer to the virtual agent's utterance. We define $PI_d \in [0,1]$ the detected Performance Index computed by the SSI system [37] as

$$PI_d = \sum_{i=1}^{N} \left(Param^i_{score} \times Param^i_{weight} \right) \qquad (1)$$

where $Param^i$ is one the $(N = 5)$ following parameters: speech duration, speech delay (hesitating before answering), speech rate, speech volume and pitch variation. $Param^i_{weight}$ is the weight of each parameter fixed to $\frac{1}{N}$. The $Param^i_{score}$ is defined such that

$$Param^i_{score} = \left(\frac{Param^i_{detected} - Param^i_{reference}}{Param^i_{detected} + Param^i_{reference}} \right) /2 \qquad (2)$$

where $Param^i_{detected}$ is the detected value using SSI while $Param^i_{reference}$ is the average value computed from a recorded corpus in preliminary study. The gender-dependency of the last three parameters was taken into account. Elements from the body language (head movement, crossing hand, etc.) could be used for the computation of the performance index. The details of this work is described in [3].

Affective adaptation

For each utterance by the virtual agent, we define an expected performance, $PI_e \in [0,1]$, whose value is based on the expert's anticipation of the participant's reaction. For instance, in the context of job interview that encompass our study, a difficult question about the person's weaknesses would be associated to a rather low PI expectation for the population we work with (inexperienced youngsters).

The detected performance index PI_d is compared to the expected performance index PI_e received from the scenario module.

Since we want to build a different reaction for good and low performances, we separate the state space in two. Let PI^H represent the good performance ($PI \geq 0.5$), which will lead to positive affects for our virtual recruiter, and PI^L the low performance ($PI < 0.5$), which will lead to negative reactions from the recruiter:

$$PI^H = \begin{cases} 2PI - 1 & if\ PI \geq 0.5 \\ 0 & otherwise \end{cases} \qquad (3)$$

$$PI^L = \begin{cases} 1 - 2PI & if\ PI < 0.5 \\ 0 & otherwise \end{cases} \qquad (4)$$

We compute the virtual agent's emotions by comparing youngster detected and expected affects (i.e. PI_d and PI_e respectively), following the OCC theory [26]. Joy is the occurrence of a desirable event: we simply assumed that youngster's detected positive affects (PI^H_d) increase the joy of the agent. The intensity of joy felt by the agent is defined as:

$$E_f\,(Joy) = PI^H_d \qquad (5)$$

Similarly, the distress is raised by the occurrence of an undesirable event, *i.e.* low performance in our case:

$$E_f\left(Distress\right) = PI_d^L \tag{6}$$

Following the same approach, we define the intensity of the hope and fear using the expected performance index:

$$E_f\left(Hope\right) = PI_e^H \tag{7}$$

$$E_f\left(Fear\right) = PI_e^L \tag{8}$$

Note that when the expected performance decreases, the fear increases and the hope decreases.

Disappointment is activated when the expected performance is positive and higher than the detected one: the youngster did not behave as good as the virtual agent expected:

$$E_f\left(Disappointment\right) = \max\left(0, PI_e^H - PI_d^H\right) \tag{9}$$

Similarly, we compute admiration (positive expectation with higher detected performance), relief (negative expectation with higher detected performance) and anger (negative expectation and even worse performance):

$$E_f\left(Admiration\right) = \max\left(0, PI_d^H - PI_e^H\right) \tag{10}$$

$$E_f\left(Relief\right) = \max\left(0, PI_e^L - PI_d^L\right) \tag{11}$$

$$E_f\left(Anger\right) = \max\left(0, PI_d^L - PI_e^L\right) \tag{12}$$

Using these emotions, we compute the mood of the agent, which represents the long-term affective state, from which we derive the attitude. The details of the computation for the mood is given in [20] and relies on the ALMA model [15] and the Mehrabian's theory [25]. The outcome is a set of 7 categories of mood, with values in $[0, 1]$: friendly, aggressive, dominant, supportive, inattentive, attentive and gossip. This is combined with the agent's personality to compute the attitude, following the work by [31] and [38]. For example, an agent with a non-aggressive personality may still show an aggressive attitude if its mood becomes very hostile. The exact combination, based on a logical-OR with fuzzy rules and transformation of categories into continuous values in Isbister's interpersonal circumplex [17], is given in [20]. In short, we have n attitude values ($val\left(a\right)$) positioned in the interpersonal circumplex and we compute two values in friendly ($Axis_{Friendly}$) and dominance ($Axis_{Dominance}$) axis:

$$Friendly = \frac{1}{n}\sum\left(val\left(a_i\right) \times Axis_{Friendly}\left(a_i\right)\right) \tag{13}$$

$$Dominance = \frac{1}{n}\sum\left(val\left(a_i\right) \times Axis_{Dominance}\left(a_i\right)\right) \tag{14}$$

The level of dominance and friendliness represent the global attitude of the agent toward the interlocutor. This attitude values are stored and their variations will serve as inputs, together with the emotions and mood, to the behaviour planner for the selection of non-verbal behaviour.

Behaviour Planner

The *Sequential Behaviour Planner* module is in charge of planning the virtual agent's behaviour. It receives two inputs. The first input is the next utterance to be said by the virtual agent in the interview, annotated with communicative intentions (*e.g. ask* a question, *propose* a conversation topic). Communicative intentions are expressed by non-verbal behaviour, and the planning algorithm makes sure that appropriate signals are chosen to express the input intentions. For instance, the *deny* intention can be expressed through a head shake or/and a finger-wagging gesture.

The second input is set of emotion values and attitude variations computed by the affective adaptation model presented the Section . In this paper, we do not present the emotion expression mechanism, which is based on previous work by the Greta team [27]: we focus on the attitude expression, which is the novelty of our agent.

As shown in section , non-verbal signals can convey attitudes and some models for attitude expression have already been proposed [6, 10, 29], but they only look at signals independently. Yet, the meaning of a signal can vary depending on the signals that precede or follow it, as showed [21, 39]. This is the reason why we propose, in our model, to consider the sequence of signals rather than independent attitude expressions.

To choose an appropriate sequence of signals to express an attitude variation, our algorithm relies on a dataset of signals sequences frequently observed before this type of attitude variation in the context of the interaction. Thus, our model is domain dependant and relies on a annotated corpus of interviews. In this Section, we first present the methodology for extracting non-verbal signals sequences expressing attitude variations from a corpus. We present the data obtained in our job interview corpus that serve as the behavioural source for our virtual recruiter. Last, we explain how these sequences are used to drive the virtual agent's behaviour according to an input attitude variation, while expressing the input sentences' communicative intentions.

Job interview corpus

The first step in building our attitude expression model was to collect a corpus of relevant interactions. Job interviews between human resources practitioners and youngsters performed in a job center were recorded. The non-verbal signals of the recruiters, their attitudes and the turn taking (*i.e.* speaking or listening) were then annotated manually on 3 videos, for a total of slightly more than 50 minutes of data. For the non-verbal signals annotation, we adapted the MUMIN multi-modal coding scheme [2] to our task. We considered the following modalities: gestures (*e.g.* adaptors, deictics), hands rest positions (*e.g.* arms crossed), postures (*e.g.* lean back), head movements (*e.g.* nods) and directions (*e.g.* head upwards), gaze and facial expressions (*e.g.* smiles).

The annotation of the recruiters' interpersonal attitudes was performed using the continuous annotation tool GTrace [13]. We adapted the software for the interpersonal attitude dimensions of dominance and friendliness. The speech was filtered

to be unintelligible, as we wanted to focus on the perception of non-verbal signals 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 at a time to reduce the cognitive load of the task. With this process, we collected two to three continuous annotation traces per attitude dimension per video.

Extraction of sequences from the corpus

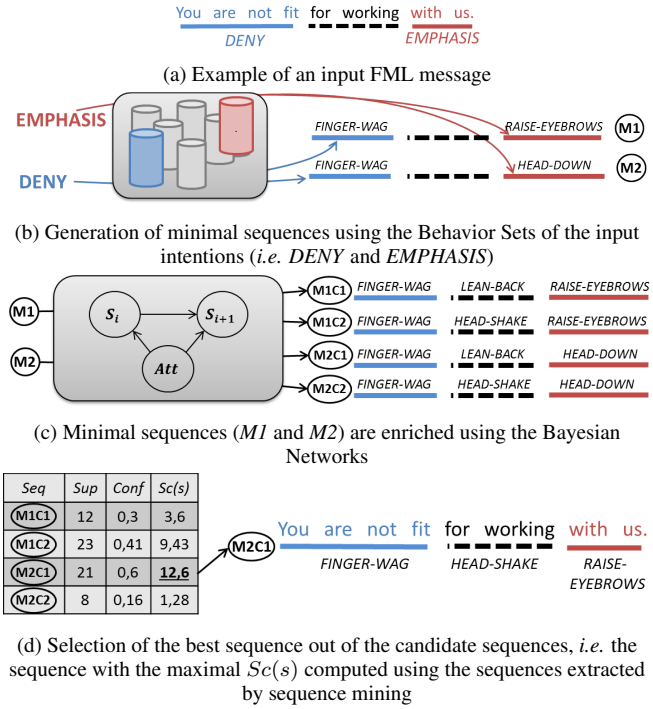
In order to extract significant sequences of non-verbal signals conveying variations in interpersonal attitudes from our corpus, we begin by parsing and pre-processing the non-verbal signals annotations to filter the annotation modalities to investigate and to convert them into a single long sequence of non-verbal signals using their starting times. Then, the attitude annotation files are analysed to find the timestamps where the annotated attitudes vary. The non-verbal signals streams are then segmented using the attitude variations timestamps. We obtain 245 non-verbal behaviour segments preceding dominance variations and 272 preceding friendliness variations, which are then split further depending on the type of attitude variation (we define 8 attitude variation types for: rise or fall, large or small variation, dominance or friendliness dimension). For instance, we obtain a dataset of 79 segments leading to a large drop in friendliness, and a dataset of 45 segments leading to a large increase in friendliness.

The resulting segments are given as input datasets to the Generalized Sequence Pattern (GSP) frequent sequence mining algorithm described in [32]. The GSP algorithm requires a parameter Sup_{min}, *i.e.* the minimum number of times a sequence happens in the dataset to be considered frequent. This algorithm follows an iterative approach: it begins by retrieving all the individual items (*i.e.* in our case, non-verbal signals) that happen at least Sup_{min} times in the dataset. These items can be considered as sequences of length 1: the next step of the algorithm consists in trying to extend these sequences by appending another item to them and checking if the resulting sequence occurs at least Sup_{min} times in the dataset. This step is then repeated until no new sequences are created. We run the GSP algorithm on each dataset and obtain frequent sequences for each kind of attitude variation. We also compute the *confidence* value for each of the extracted frequent sequences: confidence represents the ratio between the number of occurrences of a sequence before a particular attitude variation and the total number of occurrences of that sequence in the data [33].

Sequences selection for attitude expression

The *Sequential Behaviour Planner* module's role is to turn communicative intentions into signals while expressing the attitude variation. We built this module using the frequent sequences extracted with the method described in the previous section. Figure 2 is a graphical representation of the generation algorithm.

The planning algorithm starts by generating a set of *minimal non-verbal signals sequences* that express the communicative intentions (2b). This is done using a lexicon, a file that defines the sets of signals (called *Behaviour Sets*) that can be used to



(a) Example of an input FML message

(b) Generation of minimal sequences using the Behavior Sets of the input intentions (*i.e. DENY* and *EMPHASIS*)

(c) Minimal sequences (*M1* and *M2*) are enriched using the Bayesian Networks

(d) Selection of the best sequence out of the candidate sequences, *i.e.* the sequence with the maximal $Sc(s)$ computed using the sequences extracted by sequence mining

Figure 2: Graphical description of the sequence generation algorithm

express communicative intentions [23]. *These minimal sequences* are computed to make sure that the communicative intentions are expressed by the adequate signals: for instance, this is how we choose gestures that are aligned with the virtual recruiter's speech, both in timing and in form [24]. The full set of *minimal sequences* is obtained by computing all the possible combinations of signals of the different intentions' behaviour sets.

Then, the algorithm creates *candidate sequences* from the *minimal sequences* by enriching them with additional non-verbal signals (2c). We designed a Bayesian Network (BN) to model the relationship between pairs of adjacent signals and with the different attitude variations, and trained it on our corpus. The BN is represented in Figure 2c, where S_i and S_{i+1} are adjacent signals in a sequence, and Att is the type of attitude variation. Taking a *minimal sequence* as an input, the algorithm looks for the available time intervals in the input FML message. It then tries to insert every kind of non-verbal signal considered in the network into this interval and computes the probability of the resulting sequence: if that probability exceeds a threshold α, the sequence is considered as a viable *candidate sequence* and carries over to the next step. The α parameter was determined by hand for achieving a compromise between computing time and number of *candidate sequences*.

Once the set of *candidate sequences* has been computed, the selected sequence is obtained using a majority-voting method derived from [18]. For every candidate sequence s, we find the k sub-sequences of signals sub_i contained in s, that have

the highest *confidence* scores, regardless of the type of attribute variation it was extracted in. Every subsequence votes for the type of attitude variation it was extracted in, and the sequence s is then classified to express the type of attitude variation that got the most votes. Additionally, we compute a score $Sc(s)$: if the sequence s was not extracted in the sequence mining process, we define $Sc(s) = \sum_{i=1}^{k} \lambda_{sub_i} * Sc(sub_i)$, with $\lambda_{sub_i} = 1$ if sub_i was extracted for the chosen attitude variation, 0 otherwise. If s was extracted in the sequence mining process, we simply define $Sc(s) = Conf(s) * Sup(s)$. Finally we pick the sequence s in this set that has the highest score Sc (2d). The chosen sequence is then expressed in the BML mark-up language [36]. This BML message is then interpreted by our virtual agent platform's *Behaviour Realizer*, which creates an animation for the virtual recruiter.

JOB INTERVIEW SIMULATION

Our adaptive virtual agent could be used in several applications. The context of our development is job interview simulation. We have developed an adaptive virtual recruiter based on the architecture presented in the previous section. This recruiter is capable of perceiving the participant's behaviour, using a Kinect camera and a high quality microphone. It simulates emotions and attitudes dynamics and expresses them.

Different configurations could be used for the adaptive virtual agent. It corresponds to a different set of values for the personality, the recruiter's questions and the corresponding expectations: one agent is considered as more demanding, while the other is supposed to have a more understanding personality.

The scenario of the job-interview simulation as well as the expected performance index (PI_e) for each utterance are defined based on the job interview corpus. We have proposed a series of 16 utterances (questions, explanations, etc) that a recruiter asks to the interviewee in our job interview scenario. This series has been validated by domain experts (practitioners from job center). For each utterance, experts have anticipated possible non-verbal reactions from the interviewee, based on the interaction topic. For instance, a difficult question about the applicant's weaknesses should trigger expressions of embarrassment from an inexperienced applicant. We have defined the values for the expected performance index based on this expertise (for instance, the question about the applicant's weaknesses will be associated with a low PI_e).

In a real-world application, the developed adaptive agent is in use at the job-centers (in France and UK) in order to help youngsters improving their social skills to get a job. The practitioners at job-centers can rate the interview performance during TARDIS. The practitioners then use their ratings, along with any observation notes that they make while observing the youngsters interacting with TARDIS? virtual recruiter, to structure their feedback to the youngsters during a debriefing sessions. Following the youngster-TARDIS interaction, the youngster and the practitioner review together one or more recorded interaction with the TARDIS virtual recruiter. The recorded virtual recruiter?s questions and the youngster?s responses can be displayed. The practitioner can identify the most or least successful responses of the youngster and comment on them with reference to the posture, the tone of voice, the smiles and the content of the response. Similarly, the youngster can identify and comment himself/herself on positive and negative aspects of his/her own performance.

As preliminary experiment, we collect data of youngsters in interaction with the virtual recruiter in order to evaluate subjectively the performance of the agent. Preliminary results show the interest of practitioner on our research which is helpful in term of given feedback to youngsters and time saving as well as for the youngsters who find the adaptive virtual agent credible and useful to train their self before passing a real job interview.

CONCLUSION AND FUTURE WORK

With the aim to aid young job seekers in acquiring social skills, we developed a socially adaptive virtual recruiter who is able to analyse the behaviour of the human participant, to update a cognitive model with social constructs (*e.g.* attitude, relationship) depending on the behaviour of their interlocutor, and to show coherent social attitude expression. This approach is in evaluation step in the context of job-interview simulation.

We plan to extend our model that relies on the computation of the performance index. We intend to consider additional signals to compute this performance index such as body language (head movements, crossing hands, postures, ...), facial expressions (smile, surprise, ...). It could also be enriched with a speech recognition for contextual information. Yet, this performance index is currently seen as an exact value, while it is computed from social cues detected with perception software that can produce errors. The next version of our adaptive agent should rely on an extended model with uncertainty values.

ACKNOWLEDGMENT

This research has received funding from the European Union Information Society and Media Seventh Framework Programme FP7-ICT-2011-7 under grant agreement 288578.

REFERENCES

1. Acosta, J. C., and Ward, N. G. Achieving rapport with turn-by-turn, user-responsive emotional coloring. *Speech Communication 53*, 9-10 (Nov. 2011), 1137–1148.

2. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation 41*, 3-4 (2007), 273–287.

3. Anderson, K., Andr, E., Baur, T., Bernardini, S., Chollet, M., Chryssafidou, E., Damian, I., Ennis, C., Egges, A., Gebhard, P., Jones, H., Ochs, M., Pelachaud, C., Porayska-Pomsta, K., Rizzo, P., and Sabouret, N. The tardis framework: Intelligent virtual agents for social coaching in job interviews. In *Advances in Computer*

Entertainment, D. Reidsma, H. Katayose, and A. Nijholt, Eds., vol. 8253 of *Lecture Notes in Computer Science*. Springer International Publishing, 2013, 476–491.

4. Argyle, M. *Bodily Communication*. University paperbacks. Methuen, 1988.
5. Argyle, M., and Cook, M. *Gaze and Mutual Gaze*. Cambridge University Press, 1976.
6. Ballin, D., Gillies, M., and Crabtree, B. A framework for interpersonal attitude and non-verbal communication in improvisational visual media production. In *First European Conference on Visual Media Production* (2004), 203–210.
7. Bickmore, T. W., and Picard, R. W. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI) 12*, 2 (2005), 293–327.
8. Burgoon, J. K., Buller, D. B., Hale, J. L., and Turck, M. A. Relational messages associated with nonverbal behaviors. *Human Communication Research 10*, 3 (1984), 351–378.
9. Buschmeier, H., Baumann, T., Dosch, B., Kopp, S., and Schlangen, D. Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Association for Computational Linguistics (2012), 295–303.
10. Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., Jóhannsdóttir, K. R., and Valgarsson, G. S. First impressions: users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In *Intelligent Virtual Agents*, IVA'12, Springer-Verlag (Berlin, Heidelberg, 2012), 67–80.
11. Carney, D. R., Hall, J. A., and LeBeau, L. S. Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior 29*, 2 (2005), 105–123.
12. Chollet, M., Sratou, G., Shapiro, A., Morency, L.-P., and Scherer, S. An interactive virtual audience platform for public speaking training. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, International Foundation for Autonomous Agents and Multiagent Systems (Richland, SC, 2014), 1657–1658.
13. Cowie, R., Cox, C., Martin, J.-C., Batliner, A., Heylen, D., and Karpouzis, K. Issues in data labelling. In *Emotion-Oriented Systems*, R. Cowie, C. Pelachaud, and P. Petta, Eds., Cognitive Technologies. Springer Berlin Heidelberg, 2011, 213–241.
14. DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., et al. Simsensei kiosk: a virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, International Foundation for Autonomous Agents and Multiagent Systems (2014), 1061–1068.
15. Gebhard, P. ALMA - A Layered Model of Affect. *Artificial Intelligence* (2005), 0–7.
16. Hoque, M. E., Courgeon, M., Martin, J.-C., Mutlu, B., and Picard, R. W. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, ACM (2013), 697–706.
17. Isbister, K. *Better Game Characters by Design: A Psychological Approach (The Morgan Kaufmann Series in Interactive 3D Technology)*. Morgan Kaufmann Publishers Inc., 2006.
18. Jaillet, S., Laurent, A., and Teisseire, M. Sequential patterns for text categorization. *Intelligent Data Analysis 10*, 3 (2006), 199–214.
19. Johnson, W. L., and Rickel, J. Steve: An animated pedagogical agent for procedural training in virtual environments. *ACM SIGART Bulletin 8*, 1-4 (1997), 16–21.
20. Jones, H., and Sabouret, N. TARDIS - A simulation platform with an affective virtual recruiter for job interviews. In *IDGEI (Intelligent Digital Games for Empowerment and Inclusion)* (2013).
21. Keltner, D. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology 68* (1995), 441–454.
22. Kempe, B., Pfleger, N., and Lckelt, M. Generating verbal and nonverbal utterances for virtual characters. In *Virtual Storytelling. Using Virtual Reality Technologies for Storytelling*, G. Subsol, Ed., vol. 3805 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2005, 73–76.
23. Mancini, M., and Pelachaud, C. Generating distinctive behavior for embodied conversational agents. *Journal on Multimodal User Interfaces 3*, 4 (2009), 249–261.
24. McNeill, D. *Hand and Mind: What Gestures Reveal about Thought*. Psychology/cognitive science. University of Chicago Press, 1996.
25. Mehrabian, A. Pleasure-arousal-dominance: A general framework for describing and measuring individual Differences in Temperament. *Current Psychology 14*, 4 (1996), 261.
26. Ortony, A., Clore, G. L., and Collins, A. *The Cognitive Structure of Emotions*. Cambridge University Press, July 1988.
27. Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., and De Carolis, B. Greta. a believable embodied conversational agent. In *Multimodal intelligent information presentation*. Springer, 2005, 3–25.

28. Prendinger, H., and Ishizuka, M. The empathic companion: A character-based interface that addresses users'affective states. *Applied Artificial Intelligence 19*, 3-4 (2005), 267–285.

29. Ravenet, B., Ochs, M., and Pelachaud, C. From a user-created corpus of virtual agent's non-verbal behaviour to a computational model of interpersonal attitudes. In *Intelligent Virtual Agents*, Springer-Verlag (Berlin, Heidelberg, 2013).

30. Schroder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., Ter Maat, M., McKeown, G., Pammi, S., Pantic, M., et al. Building autonomous sensitive artificial listeners. *Affective Computing, IEEE Transactions on 3*, 2 (2012), 165–183.

31. Snyder, M. The influence of individuals on situations: Implications for understanding the links between personality and social behavior. *Journal of Personality 51*, 3 (1983), 497–516.

32. Srikant, R., and Agrawal, R. Mining sequential patterns: Generalizations and performance improvements. *Advances in Database Technology 1057* (1996), 1–17.

33. Tan, P.-N., Steinbach, M., and Kumar, V. *Introduction to Data Mining (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

34. Van Deemter, K., Krenn, B., Piwek, P., Klesen, M., Schröder, M., and Baumann, S. Fully generated scripted dialogue for embodied agents. *Artificial Intelligence 172*, 10 (2008), 1219–1244.

35. Vardoulakis, L. P., Ring, L., Barry, B., Sidner, C. L., and Bickmore, T. W. Designing relational agents as long term social companions for older adults. In *Intelligent Virtual Agents - 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September, 12-14, 2012. Proceedings* (2012), 289–302.

36. Vilhjálmsson, H., Cantelmo, N., Cassell, J., E. Chafai, N., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A. N., Pelachaud, C., Ruttkay, Z., Thórisson, K. R., Welbergen, H., and Werf, R. J. The Behavior Markup Language: Recent Developments and Challenges. In *Intelligent Virtual Agents*, Springer-Verlag (Berlin, Heidelberg, 2007), 99–111.

37. Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., and André, E. The social signal interpretation (ssi) framework-multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain* (2013).

38. Wegener, D. T., Petty, R. E., and Klein, D. J. Effects of mood on high elaboration attitude change: The mediating role of likelihood judgments. *European Journal of Social Psychology 24*, 1 (1994), 25–43.

39. With, S., and Kaiser, W. S. Sequential patterning of facial actions in the production and perception of emotional expressions. *Swiss Journal of Psychology 70*, 4 (2011), 241–252.

40. Yang, Z., Metallinou, A., and Narayanan, S. Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues. *IEEE Transactions on Multimedia 16*, 6 (Oct 2014), 1766–1778.