Who's Afraid of Job Interviews? Definitely a Question for User Modelling

Kaśka Porayska-Pomsta[1], Paola Rizzo[1], Ionut Damian[2], Tobias Baur[2], Elisabeth André[2], Nicolas Sabouret[3], Hazaël Jones[4], Keith Anderson[5], and Evi Chryssafidou[1]

[1] London Knowledge Lab, Institute of Education, London WC1N 3QS, UK
{K.Porayska-Pomsta,P.Rizzo,E.Chryssafidou}@ioe.ac.uk
[2] Human Centered Multimedia, Augsburg University, 86159 Augsburg, Germany
{damian,baur,andre}@hcm-lab.de
[3] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, 91403 Orsay, France
nicolas.sabouret@limsi.fr
[4] Laboratoire d'Informatique de Paris 6, 4 Place Jussieu, 75005 Paris, France
hazael.jones@lip6.fr
[5] Tandemis Limited, 108 Blackheath Hill, London SE10 8AG, UK
keith@tandemis.co.uk

Abstract. We define job interviews as a domain of interaction that can be modelled automatically in a serious game for job interview skills training. We present four types of studies: (1) field-based human-to-human job interviews, (2) field-based computer-mediated human-to-human interviews, (3) lab-based wizard of oz studies, (4) field-based human-to-agent studies. Together, these highlight pertinent questions for the user modelling field as it expands its scope to applications for social inclusion. The results of the studies show that the interviewees suppress their emotional behaviours and although our system recognises automatically a subset of those behaviours, the modelling of complex mental states in real-world contexts poses a challenge for the state-of-the-art user modelling technologies. This calls for the need to re-examine both the approach to the implementation of the models and/or of their usage for the target contexts.

1 Introduction

As a domain of interaction, job interviews rely crucially on the participants' mutual modelling of each other's behaviours and mental states. The ultimate goal of a job interview is for the interviewer to ascertain the fit of the candidate to a particular job and, ideally, for the candidate to assess a given company as a possible workplace [1]. Job interviews are often a game of bluff, where personas are adopted by the interactants and where it is normal, even expected, that the display of participants' real emotions may be suppressed [2]. This presents substantial challenges for real-time user modelling: the subtle nature of the behaviours manifested by the interviewees in such contexts makes them difficult

to detect as well as to interpret in terms of more complex mental states. The interpretation of the observable behaviours in terms of the mental states, such as *stress*, *boredom* or *hesitation* is important as those states may be indicative of a person's ability to cope with the demands of a given job. The primary challenge, as we see it, is in obtaining a reliable measure of the users' affective states during interactions that could inform the design of our model and/or against which the model could be evaluated. This challenge is well known in the field [3,4].

In this paper we present four studies, which have iteratively informed the implementation of the user modelling tools in the TARDIS project.[1] TARDIS implements a serious game for job interview skills coaching for young unemployed people, aged 18-25. The game is motivated by a growing need for technology-enhanced approaches to helping young people gain skills needed to secure jobs, both because of the marked youth unemployment and the expense associated with traditional methods, such as mock job interviews enactments.

The TARDIS user modelling tools, as well as the serious game more generally, have been described in [5] and [6]. Presently, we discuss some key issues, highlighted through the studies, that relate to finding a balance between the need to detect and interpret target users' subtle behaviours in ecologically valid contexts and the still limited capabilities of the state-of-the-art social cues detection technologies. Our work demonstrates that striving for ecological validity of our models, while highly desirable, further exacerbates the challenges of finding reliable measures of the phenomena of interest.

2 Related Work

Nonverbal behaviours are key in job interviews. For example, [7] found a relationship between audio-visual cues of the candidates and the interview outcomes. [8] studied how the success of simulated job interviews can be predicted from conversational engagement, vocal mirroring, speech activity, and prosodic emphasis. Other researchers have focused on the relationship between interviewers' decision making and the perceived personality of the candidate (measured along the dominance, equivalence and submissiveness dimensions) and the related behaviours [9]. [10] found a negative correlation between the interviewees' performance (interview scores) and trait anxiety, while [11] found a link between high state anxiety and information acquisition and retention, suggesting that anxiety may interfere with the applicant's acquisition and processing of the information presented to them by the recruiters and thus, with their performance. This implies that anxiety regulation is fundamental to candidates' performance in interviews.

Less is known about interviewee's other mental states that may be relevant to achieving success in an interview. Crucially, most of the substantial evidence that links the specific social cues with candidates' traits or states has been conducted in the laboratory settings with university students. While this research is of practical importance to us, a key difference between it and the context of TARDIS is that we aim to define the characteristics of a population which is at

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

risk of marginalisation, with our technology being designed for use in real-world contexts of youth organisations across Europe. TARDIS' focus, therefore, leads to a need to (a) verify and define further the states and social cues that are pertinent to the contexts of its intended use and (b) identify, implement and test the social cue detection tools that are affordable, robust and least intrusive.

Using signal processing techniques to detect behavioural patterns is not a new idea, e.g. [12]. However, to date, most research focused on a reduced number of modalities to infer user states, such as speech [13] or facial expressions [14]. Relatively little attention has been paid to gestures or postures [15,16]. Furthermore, most work on signal processing is intended for offline analysis, rather than real-time interactive applications. For example, in Batrinca et al.'s [17] system for practicing public speaking, behaviour analysis happens post-hoc and offline, with their system not being able to react to the user's behaviour in real-time.

There are, of course, exceptions, one of which is the MACH job interview simulation system [18], which is able to detect a limited number of social cues in real time, including smiles, audio features and speech. In contrast, our system recognises a much broader range of social cues, including bodily cues, such as expressivity features, gestures and postures, physiological features and eye gaze, although it does not engage in speech recognition [5].

In the remainder of this paper we present the four studies aimed to define job interviews as a domain of interaction, specifically focusing on the evaluation of social cues and mental states for use during interactions in real-world contexts.

3 Manual Annotations of Mock Interviews by Experts

To identify the social cues and hidden mental states displayed by youngsters during mock job interviews, we conducted a study with ten youngsters and five practitioners at a youth association in France. The study's procedure involved one-on-one mock job interviews, all of which were video recorded, followed by semi-structured interviews with youngsters and practitioners, and post-hoc video walkthroughs with practitioners. The semi-structured interviews focused on identifying the youngsters' strengths and weaknesses during each mock interview. The walkthroughs served to identify the social cues observed by the practitioners and the hidden mental states that could be linked to those cues.

The walkthroughs were facilitated by the Elan annotation tool (Fig. 1, left), which allows simultaneous replay of videos and their annotations. During the walkthroughs, the practitioners were asked to stop the videos anytime they observed a critical incident. *Critical incident* was defined as a specific behaviour on the part of the interviewee, e.g. smile, or a set of behaviours, e.g. persistent smiling and gaze averting, that the practitioner thought crucial, in a positive or negative way, to the job interview and its outcome. This procedure allowed for the key behaviours in the given interactions to be identified within exact time frames and to be annotated additionally with the practitioners comments – these were used in further video data analysis by independent annotators.

Three interactions were annotated by the practitioners for social cues with additional comments linking them to specific hidden mental states. This resulted in

nineteen individual social cues, as shown on the x-axis of Fig. 1, right. One annotator coded the videos for social cues, using practitioners' walkthrough annotations as exemplars. A second independent annotator verified those annotations, ensuring that all observable behaviours of interest were captured. The inter-rater agreement analysis was not conducted at this point, however the two annotators met to agree the thresholds for annotating social cues including *long silence* (established as ≥ 3 seconds) and *short answer* to questions requiring elaboration (established as simple yes/no answer), as annotating these cues presented the most difficulty for the annotators. The videos were then re-annotated using these thresholds. However, full agreement could not be achieved with respect to the instances of *clear/low voice*. These presented significant problems primarily due to the low quality of the recordings which were taken in a minimally controlled environment of a real youth association, with the normal daily business of the association taking place at the same time, the outside noise often interfering with the recordings. Gaze *saccades* were also extremely difficult to establish through the video analysis: given that the recorded interactions were face-to-face between two humans, achieving an ideal angle of the camera to capture as fine grained detail as the youngsters' eye-gaze shifts proved virtually impossible. While this means that some social cues were hard to identify with confidence through the videos alone, given that TARDIS is intended for use in real youth associations, the need for a careful selection of the social cue sensors along with their set up in real-world contexts was clearly highlighted.

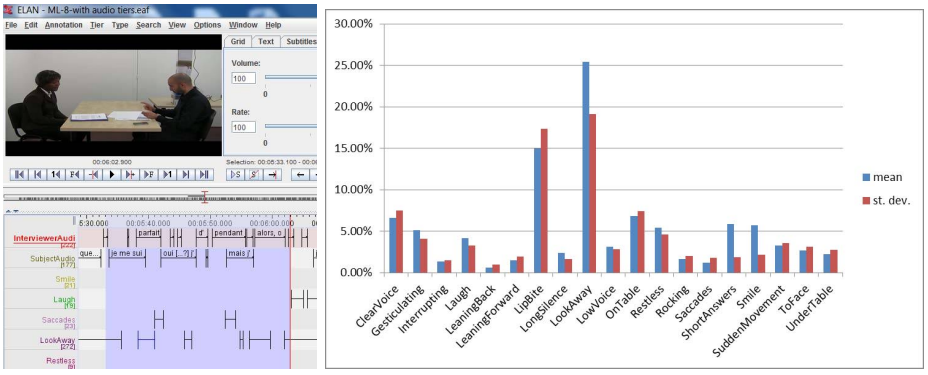

Fig. 1. Human-to-human mock interview with social cues manually annotated (left) and percentages of frequencies of social cues across participants (right)

Eight complex mental states have been identified during walkthroughs, including: (i) stressed, (ii) embarrassed, (iii) hesitant, (iv) ill-at-ease, (v) bored, (vi) focused, (vii) relieved and (viii) relaxed. These mental states have been associated by the practitioners with specific social cues in the videos annotated. For example, observable behaviours such as *looking away*, *laughter* and *hand-to-mouth*, have been associated with youngsters' *embarrassment*, whereas *restless*

hands – with *stress*. The mental states annotations, along with the practitioners' comments provided the basis for further manual annotations of the videos. Two independent annotators coded the video data for the eight mental states. Unfortunately, with Cohen's Kappa below 0.2, the inter-rater agreement was significantly below the level necessary to provide a reliable measure of youngsters' affects that could (a) be generalisable to other youngsters and (b) could serve as a reliable measure against which to evaluate the TARDIS user model directly [3]. Amongst the eight states identified, the greatest source of difficulties amongst the annotators related to the difference between *embarrassed* and *ill-at-ease*, which one annotator found virtually impossible to distinguish. On the other hand, *stress* seemed so ubiquitous that it became at times difficult for the annotators to differentiate it from the other states.

The difficulties in finding a good agreement between mental states annotations are not altogether surprising given that other researchers have reported similar set-backs when trying to establish some ground truth for eliciting emotion recognition models [3]. One typical culprit is the use of labels which are a liability owing to the imprecision of language, with the meaning of a label being typically constrained by context and linguistic repertoire of the labeller [4] - in our case the practitioners. The fact that several practitioners came up with the same labels for the youngsters mental states may be an artefact of their working and training together, which may have resulted in their labelling habits being aligned. Another potential reason for the imprecision of labels may be the fact that they have been provided in French and then translated into English, leaving further scope for linguistic imprecision.

However, the most compelling explanation seems to lie in the great variability in the behaviours manifested by the youngsters (in Fig.1(right), many standard deviations of frequency and duration of the social cues identified are higher than the mean occurrences of those cues), which makes any standardisation of the mental states labelling very difficult. Furthermore, the individual differences between the youngsters' behaviours may represent persistent behavioural traits rather than being dependent on the context of the interaction. For example, one youngster looked away from the interviewer over 250 times in one interview, compared to another two who have only done so 60-100 times, and two others who have never been observed to look away. Similarly, all youngsters seem to lip bite to some extent (10-15 times each), but one youngster did it 80 times in the course of a 20 minute interaction. This suggests that although the practitioners were able to name some of the youngsters' mental states, they did it relative to their individual behavioural habits. The individual differences between the youngsters also suggest that the nineteen social cues may not help us to uniquely identify the specific mental sates without recourse to some qualifying information such as the interviewer's questions, some of which, e.g. questions related to the candidate's weaknesses, may be generally more difficult than other.

A further data analysis (based on one coder's annotations of mental states) seems to confirm the weak discriminative power of the cues identified. Specifically, given that social cues can occur either in isolation or in combination with

other cues, we decomposed the social cue data into all groups (defined as any overlap between 2 or more social cues) that occurred across all participants. We then assigned probabilities to each grouping of cues to represent the likelihood that it implies an emotional state. This was done by measuring the duration of each social cue grouping (CG), and the duration of its intersection D_t with the presence of an emotional state (ES), using the following simple formula:

$$P(ES|CG) = \frac{D_t(CG)}{D_t(ES)} \tag{1}$$

Despite there being many groups of cues that were found uniquely or very strongly to imply the presence of a single emotional state, there were many groupings that co-occurred rarely with an emotional state. For example, the combination of *leaning forward* while *looking away* was found to lead to a high probability of *stress* ($P = 0.83$), based on its total occurrence across all annotations of 7.6 seconds and its total co-occurrence with the *stressed* state of 6.3 seconds. However, *leaning back* and *speaking clearly* was found to imply *boredom* with a probability of only $P = 0.01$. Apart from a large number of individual cues (groupings) that correlate weakly with many mental states, many of the groupings occurred only once across all participants, raising a question of the extent to which many of the correspondences are generalisable to other participants and suggesting the need to reduce the number of cues and possibly the mental states modelled to only the key ones. However, the selection of the cues whose detection should be abandoned needs to be done in tandem with the investigations of what social cues are feasible to detect automatically in the job interview contexts.

4 Computer-Mediated Interaction

To ascertain the feasibility of detecting the different social cues during interaction, we conducted a further study with six youngsters and two practitioners in the UK. The study's procedure mirrored that of the study described in Section 3. However, in order to facilitate the use of the automatic detection tools as well as an approximation of the future human-agent interaction, the mock interviews were mediated through a video link, headphones and microphones. The youngster and the practitioner were situated in opposite corners of the same room, back to back (Fig. 4). This arrangement together with the isolating earphones allowed the participants to see and hear each other only through the media link. In addition, a Microsoft Kinect depth sensor was positioned over the monitor facing the youngster. This allowed us to record the participants' audio, video and skeleton tracking data. As well as informing the social cue detection framework in TARDIS, this set-up allowed us to assess the ease and the credibility of a job interview experience delivered via a computer screen and microphone.

The recording of the user's social signals was handled by TARDIS's social cue recognition [19] component which uses the Social Signal Interpretation framework [20]. The system enabled playback of the recorded data and thus, the testing of the behaviour recognisers in an online context even after the studies.

Fig. 2. Computer-mediated interview

Upon analysis of the data, we observed clear indications of (a) what social cues we can feasibly detect during interactions between youngsters and TARDIS and (b) which of these social cues may be the most robust and informative.

To this end, we refined the list of the 19 social cues identified in the previous study: vocal social cues such as *clear/ low voice* proved to be difficult to recognise due to the heterogeneity of the speakers and the physical environments in which the studies took place. Both of these cues rely on audio intensity analysis [17] – a speaker and hardware dependent feature that is highly susceptible to noise (e.g. coughs or voice clearing). Here, cues involving pitch variation, proved more robust. *Gesticulation*, *restlessness* and *sudden movements*, while correctly recognised by our automatic recognisers real-time and online, had to be joined together due to insufficient accuracy in skeleton tracking. We encountered no issues for turn taking cues such as *interrupting*, *short answers* or *long silences*, as these mainly relied on the user's voice activity compared to the practitioner's.

While recent advances in the domain of signal processing show that automatic recognition of *laughter* is feasible, this is usually the case for highly expressive forms of laughter [21]. In contrast, our analysis revealed subtle types of laughter, which proved not to be distinctive enough. Similarly, *lip biting*, *rocking* and *saccades* also turned out to be too subtle for our sensing equipment. To perceive these social cues, we would require more accurate sensors, such as an eye tracker or body worn motion tracking devices (see Section 5), which, apart from being quite expensive, may be too intrusive for some users in our target population.

Data processing also revealed that the recognition of gestures and postures (*lean front*, *lean back*, *hand to face* and *look away*) and smiles was possible using the FUBI [22] and SHORE [23] frameworks respectively. Finally, some social cues (*hands on/under table*) had to be eliminated due to the table-less setting of the study, chosen to ensure correct skeleton tracking using the Microsoft Kinect.

5 Wizard of Oz Experiment

The WOZ experiment aimed to (1) identify a combination of sensors that can enhance the recognition of youngsters' behaviours during simulated interviews to enable inferences about the users' internal states, and (2) ascertain any impact of specific types of interview question, i.e. those that might be considered difficult or aggressive, on participant's nonverbal behaviours.

The study involved three participants, who were seated in an armless chair in front of a 40" display with a Microsoft Kinect depth camera situated on top. They wore a headset, eye tracking glasses, a motion tracking glove and SC/BVP sensors on their fingers (Fig.3(a)). From the user's point of view, s/he interacted with a virtual recruiter (VR), which was, in fact, controlled by a human interviewer seated in another room (Fig.3(b)).

All sensors performed flawlessly during the interaction and the recorded data gave us a large amount of information regarding the participants' non-verbal behaviours. In particular, the skin conductance values showed the impact of the interview questions on the user, with the questions, e.g. *'What are your weaknesses?'* or harsh statements, e.g. *'I don't think you are right for this job'* correlating with higher SC values (Fig.4(b)). This suggests a possible relationship between certain types of interview questions and the candidate's emotional states, even though the interviewer posing these questions was a synthetic character.

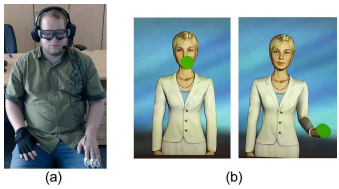

Fig. 3. Participant wearing the study apparatus (a) and images showing a user's point of view (b) including gaze information (green point) captured using the eye tracking glasses.

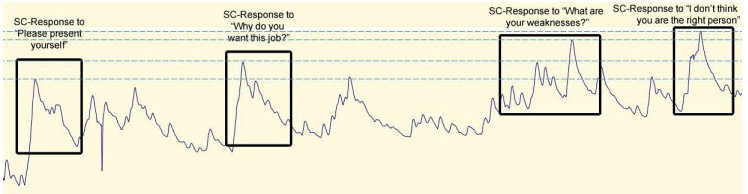

Fig. 4. Skin conductance data of one user. Highlighted areas represent user's SC response to various utterances. The blue dotted lines mark the peaks of each highlighted area.

The gaze cues clearly mark the regions of interest during the interaction. All users focused heavily on the face, in particular the mouth area of the virtual agent, followed by its torso and then, by its hands. The gaze only dropped to the hands when the agent performed a gesture as illustrated in Fig.3(b).

The study also revealed that even with the more challenging scenario, the users still performed very little in terms of physical movements. While this may have also been an effect of the sensing devices worn by the users, the observation is in line with the previous studies reported in this paper.

Additionally, even though the eye tracking data yielded some interesting trends, the eye tracking glasses' high intrusion level combined with their incompatibility with prescription glasses make them ill-fitted for large scale field studies. Given this, we decided to limit the number of sensors for future studies to the following three least intrusive sensors: depth camera, microphone and SC sensor.

6 Self-reports during Human-Agent Interaction

Building on the results of the WOZ experiment, we piloted the use of a pop-up questionnaire with seven French youngsters. The pop-up questionnaire aimed to elicit self-reports from the youngsters about their anxiety levels during their interaction with the TARDIS VR. The youngsters were asked to score their anxiety level on a 1 (not at all anxious) to 5 (extremely anxious) scale. A similar approach has been adopted in [3] to obtain emotional self-reports during the interaction with a tutoring system.

In total 124 scores were obtained against thirty interviewer questions. The questions were asked by two types of VR: (i) an understanding VR, which had a gentle manner and (ii) a demanding VR, which was more aggressive. The 124 scores were grouped according to those two conditions, resulting in 70 scores for the "demanding" and 54 for the "understanding" questions. Owing possibly to the small sample, the statistical analysis did not reveal any significant effects either with respect to the differences in the anxiety means between the two conditions (t-test: $t(122) = 0.71$, one tail $p = 0.23$), or between anxiety vs. questions asked under the three categories: (i) skills needed for the job, (ii) knowledge about the job, and (iii) salary level (ANOVA comparison: $F = 0.11$, $p = 0.89$). Nevertheless, the results shown in Fig.5(left) suggest a possible trend towards youngsters exhibiting trait anxiety, which would be in line with some of the studies reported in Section 2. The results, shown in Fig.5(right) also seem to suggest that some types of interviewer's questions may lead to greater anxiety than others: for example "Elaboration_Jobskills_Understanding" that groups questions about the skills needed for the job in the "understanding" mode, and

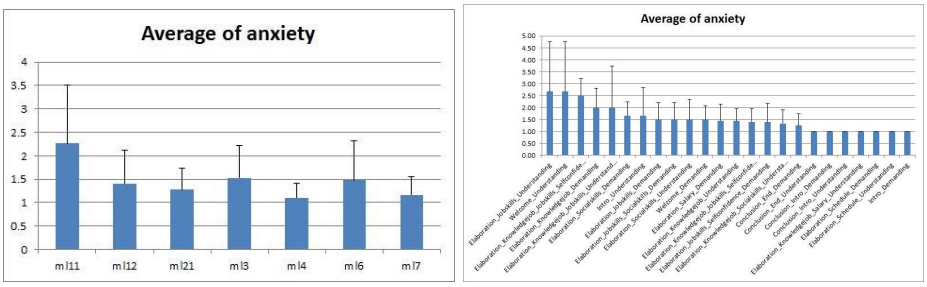

Fig. 5. Self-assessed anxiety means and standard deviations per participant (left) and self-assessed anxiety means and standard deviations per question(right)

"Welcome_Understanding", that groups questions for welcoming the participant in the "understanding" mode show higher anxiety than all the other questions analysed. Unfortunately, the small data sample and the high standard deviations for the several questions, prevent us from drawing definitive conclusions, which means that the results reported can only serve as the basis for further hypothesis generation.

7 Conclusion

In this paper we presented four formative studies which define job interviews as a domain of interaction. Each study contributed knowledge needed for the implementation of the TARDIS' real-time user model in this domain: (1) what social cues and mental states are relevant, (2) what is feasible to detect with non-intrusive technology, (3) what aspects of the interaction cause (detectable) nonverbal behaviours in users, and (4) how to evaluate anxiety.

Although, the studies presented do not offer definitive answers, they do demonstrate the magnitude of the challenge of building adaptive complex systems for real-world use, which, as TARDIS, are based on user modelling, while also having some grounding in the real world. One lesson learnt is that the use of non-intrusive sensors, coupled with the field conditions, and the peculiar nature of this interaction domain where emotional displays seem to be suppressed, lead to a reduced set of detectable cues. To address this requires a careful balancing of what is relevant to model with what is feasible to detect. Our studies suggest that focusing on key social cues, such as voice that can be reliably detected through the sensing technologies, coupled with a focus on state anxiety may be the way forward in this domain. The studies also point to a need for TARDIS to allow for an online initial training phase during which individual users' baseline of social cues can be established to allow for a tailored parameter adjustment based on the frequency of a given users' cues. This points to a continuous model, instead of a category-based one, in which users' behaviours are compared to their typical baseline and where peak behaviours that are likely indicators of corresponding peak internal reactions are identified. A complimentary approach, currently piloted in TARDIS and whose use is motivated directly by the studies reported, involves open user modelling, where the models generated online are displayed to the users who can accept or correct them according to their self-perception. This allows to both validate TARDIS' user models and to foster self-awareness in the youngsters - a pre-requisite job interview skill. Our next studies will assess youngsters' performance in human-human interviews before and after using TARDIS, in a bid to evaluate our modelling tools indirectly within TARDIS.

Acknowledgments. This work was partly funded by the European Commission (TARDIS project FP7-ICT2011-7-288578). The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC, and the EC is not responsible for any use that might be made of data appearing therein.

References

1. Posthuma, R.A., Morgeson, F.P., Campion, M.A.: Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. Personnel Psychology 55(1), 1–82 (1982)
2. Sieverding, M.: Be cool!: Emotional costs of hiding feelings in a job interview. International Journal of Selection and Assessment 17(4), 391–401 (2009)
3. Conati, C.: How to evaluate models of user affect? In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 288–300. Springer, Heidelberg (2004)
4. Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C., Baker, R.: Knowledge elicitation methods for affect modelling in education. International Journal of Artificial Intelligence in Education 22(3), 107–140 (2013)
5. Porayska-Pomsta, K., Anderson, K., Damian, I., Baur, T., André, E., Bernardini, S., Rizzo, P.: Modelling users' affect in job interviews: Technological demo. In: Carberry, S., Weibelzahl, S., Micarelli, A., Semeraro, G. (eds.) UMAP 2013. LNCS, vol. 7899, pp. 353–355. Springer, Heidelberg (2013)
6. Anderson, K., André, 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., Sabouret, N.: The TARDIS framework: Intelligent virtual agents for social coaching in job interviews. In: Reidsma, D., Katayose, H., Nijholt, A. (eds.) ACE 2013. LNCS, vol. 8253, pp. 476–491. Springer, Heidelberg (2013)
7. De Groot, T., Janaki, G.: Can nonverbal cues be used to make meaningful personality attributions in employment interviews? Journal of Business Psychology 24, 179–192 (2009)
8. Curhan, J., Pentland, A.: Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. Journal of Applied Psychology 92(3), 802–811 (2007)
9. Schmidt, N.: Social and situational determinants of interview decisions: Implications for the employment interview. Journal of Personnel Psychology 29, 79–101 (1976)
10. Ryan, A.M., Daum, D., Friedel, L.: Interviewing behavior: Effects of experience, self-efficacy, attitudes and job-search behavior. In: Annual Conference of the Society for Industrial and Organizational Psychology, San Franscisco, CA (1993)
11. Barber, A.E., Hollenbeck, J.R., Tower, S.L., Phillips, J.M.: The effects of interview focus on recruitment effectiveness: a field experiment. Journal of Applied Psychology 79, 886–896 (1994)
12. Vinciarelli, A., Pantic, M., Heylen, C., Pelachaud, C., Poggi, F., Errico, A., Schroeder, M.: Bridging the gap between social animal and unsocial machine: A survey of social signal processing. IEEE Transactions on Affective Computing. 3(1), 69–87 (2012)
13. Vogt, T., André, E., Lewis, T., R., Leibbrandt, Powers, D.: Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: IEEE International Conference on Multimedia and Expo, pp. 474–477 (2005)
14. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. 31(1), 39–58 (2009)
15. Kapoor, A., Picard, R.W.: Multimodal affect recognition in learning environments. In: Proceedings of ACM MM 2005, pp. 677–682 (2005)

16. Kleinsmith, A., Bianchi-Berthouze, N.: Form as a cue in the automatic recognition of non-acted affective body expressions. In: Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction, Amsterdam, Netherlands. Part I, pp. 155–164 (2011)
17. Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., Scherer, S.: Cicero - towards a multimodal virtual audience platform for public speaking training. In: Aylett, R., Krenn, B., Pelachaud, C., Shimodaira, H. (eds.) IVA 2013. LNCS, vol. 8108, pp. 116–128. Springer, Heidelberg (2013)
18. Hoque, M.E., Courgeon, M., Martin, J., Mutlu, B., Picard, R.W.: Mach: My automated conversation coach. In: International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2013 (2013)
19. Damian, I., Baur, T., André, E.: Investigating social cue-based interaction in digital learning games. In: Proceedings of the 8th International Conference on the Foundations of Digital Games, SASDG (2013)
20. Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., André, E.: The social signal interpretation (ssi) framework - multimodal signal processing and recognition in real-time. In: Proceedings of ACM MULTIMEDIA 2013, Barcelona (2013)
21. Niewiadomski, R., Hofmann, J., Urbain, J., Platt, T., Wagner, J., Piot, B., Cakmak, H., Pammi, S., Baur, T., Dupont, S., Geist, M., Lingenfelser, F., McKeown, G., Pietquin, O., Ruch, W.: Laugh-aware virtual agent and its impact on user amusement. In: Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS 2013, pp. 619–626. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2013)
22. Kistler, F., Endrass, B., Damian, I., Dang, C.T., André, E.: Natural interaction with culturally adaptive virtual characters. Journal on Multimodal User Interfaces 6, 39–47 (2012)
23. Küblbeck, C., Ernst, A.: Face detection and tracking in video sequences using the modifiedcensus transformation. Image Vision Comput. 24(6), 564–572 (2006)