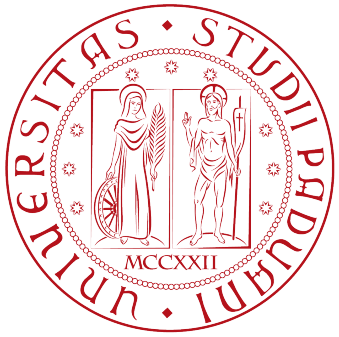


UNIVERSITY OF PADUA

DOCTORAL THESIS

Non-verbal cues of engagement during video interviews: Third-party assessment, construction and validation of a training and automatic detection

---

Author:
Mattia Furlan

Supervisor:
Anna Spagnolli

Co-supervisor:
Gian Antonio Susto

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

in

Brain, Mind and Computer Science

Department of General Psychology

Cycle XXXIV

To Lalla, my beloved aunt

To my grandparents, Teresa and Nicolino

Everything I do will always be in your memory

Contents

ABSTRACT

The use of videoconferencing platforms, through which groups of people can communicate at a distance, has increased in recent years and accelerated after the Covid-19 pandemic forcing the use of remote interactions. These interactions include work-related interactions and certainly also job interviews. This type of recruitment had already started before the outbreak of the pandemic but is certainly expected to become more and more popular in the world of work in the future. To this end, it is essential to consider what are the important components to keep in mind when conducting such an interview. It is also essential to understand how to improve and adapt to the new interview tools. In order to provide a useful tool for both candidates and recruiters, the following project was built. The aim is to identify and make explicit the non-verbal cues that characterise engagement in an interaction, in order to use them to construct training that improves the skills of both the candidate and the recruiter involved in job interviews. My project consisted of the following studies: A preliminary study thought to identify recurrent movement patterns of parties involved in interaction from the point of view of Physical Mutual Engagement (PME) was conducted. In this study, participants rated the engagement between parties involved in work-related interactions. A content analysis was conducted on the answers given by the participants to the open-ended questions, in order to identify the behaviours cues of the PME. Therefore 57 engagement cues were found, divided into 9 Behaviours and associated with 8 Meanings. A second study was carried out to validate the non-verbal engagement cues identified in the previous study. To do this, training was constructed and administered to 20 participants involved in a job interview. In order to check the effectiveness of this training, the participants' behaviour was annotated and analysed. The behaviours of Gaze, Nodding and Smiling were

identified. It was also found that the training was effective in increasing Looking into the camera and decreasing Looking away. It was also effective in increasing Nodding but not Smiling. Then, a third study was performed to verify whether the parameters found in the first study could effectively improve PME. The videos collected during the second study have been evaluated by an independent commission in order to determine whether the training had an effect on the participants' engagement. A comparison was then made between pre and post-interval videos evaluations. It was found that the training constructed during this project was indeed effective in increasing perceived engagement. Subsequently, a fourth study was carried out in order to check whether the behavioural cues annotated during the second study correlated with the engagement rates assessed by the external evaluators during the previous study. To this end, the average engagement scores found in the third study were compared with the four behavioural cues annotated in the second study. A correlation was found between Gaze behaviour and engagement scores. Higher engagement scores corresponded to higher frequencies of Looks into the camera and lower frequencies of Looks away. However, no correlation was found between Nods and engagement scores, but a slight correlation was found between Smiles and engagement scores. Finally, a fifth study was conducted in order to build a model capable of extracting and predicting the nonverbal cues found and tested in our previous studies, using state-of-the-art machine learning algorithms. The annotated frames of the videos were used to train and test the model using a network for facial recognition. Considerations on the better approach to use to predict the considered behaviours are therefore reported.

Acknowledgements

Before we start, I would like to say a few words of thanks. In my opinion, there is no such thing as individual work, so this, like any other, was a collective effort which, without the help and support of the people around me, would not have been possible.

First of all, I would like to thank my supervisor, Anna. Without her, this work would not have been possible. Thank you for your continuous support and your always constructive and precise advice.

I would also like to thank my fellow PhD students, with whom I have shared these years. First above all Giulia, my "VR twin sister". Thank you for the support, help and friendship that you never stop proving to me.

I would also like to thank my family and friends, who have always been with me and I hope forever. Thank you for accompanying me in whatever I do and for supporting my every decision.

Last but not least, I would like to thank my girlfriend Eliana, who has accompanied me through the most difficult period of this journey. Thank you for your love and for being there for me in really difficult times.

1. Introduction

The use of videoconferencing platforms, through which groups of people can communicate synchronously using audio and video, has increased in the past years and accelerated after the Covid-19 pandemic forcing the recourse to remote transactions in business, education, and recreation (Marhefka et al., 2020; Singh & Soumya, 2020). The most downloaded and used platforms worldwide are, to date, Google Meet, Zoom, Skype, Microsoft Teams, Cisco WebEx Teams and GoToMeeting (Henry & Shellenbarger, 2020; Singh & Soumya, 2020). These platforms, nowadays, allow the interaction and collaboration of a considerable number of people remotely, ranging from a minimum of 50 people (Skype) to a maximum of 500,000 (Microsoft Meets) (Henry & Shellenbarger, 2020). Among these platforms, the most popular and used is Zoom, with more than 200 million users connected every day (Singh & Soumya, 2020). Nevertheless, there is still limited evidence in the literature on the effectiveness of virtual interviews. The enormous potential of these means became even more evident during the still ongoing pandemic. The remote working situation that many companies are still adopting has given us the opportunity to ponder on this medium and to evaluate how it works when it completely replaces the traditional interview process (Chandratre & Soman, 2020).

Compared to face-to-face, videoconferencing systems certainly have different characteristics. For example, it has been remarked that visual cues in online platforms differ from what is habitual in face-to-face interaction, in the sense that there is a greater amount of non-verbal inputs (i.e., 'nonverbal overload', Bailenson, 2021). The author argues that the size of the speakers' face in the speaker view on Zoom resembles what in face-to-face would be considered an intimate distance; the sustained eye gaze experienced in videoconferences would be uncommon at a close interpersonal distance in face-to-face

meetings, where eye gaze is preferably diverted; and attention in real meetings is on the speaker, while in videoconferencing systems all participants can be constantly inspected and visible. But what do we make of the visual cues so abundantly available on a videoconference platform? And how do we use these cues? About the effect on relations, studies show that nonverbal cues guide the formation of speaker's impressions based on the interlocutor's pre-existing categories and beliefs (Todorov & Oh, 2021), affect the perceived credibility of the speaker (Burgoon et al., 1990) and suggest the social group to which the speaker belongs (Bjornsdottir & Rule, 2017). Furthermore, nonverbal cues are consequential to the speakers' wellbeing. In fact, it seems that the combination of continuous face exposure for videoconferencing and of unrealistic aesthetic canons for faces, now embedded into some AI-based camera filters (e.g., quoves.com or faceplusplus.com) have led to anxiety about the appearance of some facial features such as one's nose, forehead, glabellar wrinkles, overall skin texture and the submental/neck region (i.e., Zoom dysmorphia, Cristel et al., 2020; Rice et al., 2020) and an increase in request of aesthetic surgery (https://www.washingtonpost.com/road-to-recovery/plastic-surgery-cosmetic-covid-zoom/2020/12/07/6283e6d2-35a2-11eb-b59c-adb7153d10c2_story.html).

Although they work performantly, in that there is no necessary connection between the personality of the speaker and the nonverbal cues displayed in a video interview, current semiautomated corporate recruiting software include nonverbal cues, such as facial expressions and intonation to select the candidates' profiles (e.g., HireVue, Retorio). It is doubtful that the speakers using a videoconferencing system are aware of the effect of the nonverbal cues they seem to convey. Indeed, social psychology has repeatedly shown that people overestimate the other's ability to detect their inner feelings (i.e., illusion of transparency, Gilovich et al., 1998). The reason for this lack of awareness is that these

cues are polysemic: the same facial expression can be taken to mean very different things based on the context in which it is used (Hassin et al., 2013). Since the environment in which the parties are placed during a meeting is asymmetric, it is very likely that misunderstanding arise. The current request to be transparent about which aspects of the performance are evaluated (e.g., Illinois's Artificial Intelligence Video Interview Act effective since January 2020 and the Artificial Intelligence act in the EU), although motivated by the need to obtain truly informed consent, resonates also with the need for avoiding artefacts during interviews. The risk is to introduce a bias in the evaluation, which favours candidates who are more familiar with the new videoconferencing tools and who can therefore communicate better, look more empathic and committed, regardless of their actual skills. If all candidates are alerted about the nonverbal cues expressing this attitude, then they will be levelled up in terms of any superficial source of the impression they make.

In order to provide a useful tool for both candidates and recruiters, the following project was constructed. The aim is to identify and make explicit the non-verbal cues that characterise engagement in an interaction in order to use them to construct training that improves the skills of both the interviewee and the recruiter engaged in job interviews. My study will consist of three steps: (a) identifying which nonverbal cues are relied upon by a third party assessing a person's engagement in a video-recorded interview (preliminary study), b) assessing whether a brief training about those cues would improve a speaker's performance during a video-interview (II-III-IV studies) and c) automatically extract these cues using a novel computing system (V study). The perspective in all three cases is of a third-party not involved in the interaction, to separate the evaluation from content exchanged in the conversation and then simulate the situation of a person judging a video once it has been recorded. The training will be brief, to simulate the kind of

information that derives from a simple debriefing or disclosure of the way in which the performance was judged. I will focus on one dimension of the speaker's performance, i.e., their perceived engagement in the interaction, because it can be detected by nonverbal cues, and applies to any kind of interaction regardless of its purpose and content.

Before describing the studies, I will summarize the state-of-the-art in non-verbal cues of engagement, nonverbal cues in traditional face-to-face interviews and nonverbal cues in video interviews.

2. Background

2.1. Nonverbal cues of engagement

Engagement in a conversation, also referred to as conversational involvement, consists of "the degree to which participants in a communicative exchange are cognitively and behaviourally engaged in the topic, relationship, and/or situation" (Coker & Burgoon, 1987). Non-verbal displays of engagement have been identified with the qualitative approach of conversation analysis and consist of the orientation of the upper part of the body (Mehrabian, 1968a, 1968b). Eye gaze also shows where the interlocutor is oriented to or who the speaker is addressing (Richmond et al., 1987). Experimental studies also identified similar cues. Mehrabian (1968) collected several of these cues under the unifying label of "immediacy" (Mehrabian, 1968b), as did other authors afterwards (Andersen, 1979; Andersen, Andersen, & Jensen, 1979; McGinley, LeFevre, & McGinley, 1975; Richmond et al., 1987; Szafir & Mutlu, 2012). Nonverbal face-related cues conveying immediacy were cues such as eye contact/gaze and facial expressions. *Eye contact/gaze* was measured as the duration of each episode of mutual gazing (Mehrabian, 1968b) or as the ratio between looking at the interlocutors and somewhere else (Richmond et al., 1987; Szafir & Mutlu, 2012), while facial expressions consist of *smiling* and the spontaneous reaction of smile generates in the interlocutor (Richmond et al., 1987). Body-related cues of immediacy include *touching* the interlocutors (Mehrabian, 1968b; Richmond et al., 1987), *leaning forward* toward the interlocutors (measured as the displacement of the shoulders with respect to the pelvis line) (Mehrabian, 1968b) and the *distance* from the interlocutor (Mehrabian, 1968b; Mehrabian & Friar, 1969), further on referred to as *proximity* (Richmond et al., 1987;

Szafir & Mutlu, 2012), the *body orientation* (measured as the rotation of the body with respect to the vertical axes) (Mehrabian, 1968a), keeping a *relaxed* posture (Andersen, 1979; McGinley et al., 1975; Mehrabian, 1968b; Richmond et al., 1987; Szafir & Mutlu, 2012) and the use of *gestures*, sheerly defined as the amount of gestures used while talking (Richmond et al., 1987; Szafir & Mutlu, 2012), which seems to lead to greater liking and cooperation (Andersen et al., 1979; Mehrabian, 1971).

The set of cues conveying immediacy has subsequently been rearranged with reference to the broader notion of involvement or engagement between interlocutors, as "the degree to which participants in a communicative exchange are cognitively and behaviourally engaged in the topic, relationship, and/or situation" (Coker & Burgoon, 1987). Coker & Burgoon consider immediacy to be only one component of involvement, and the set of the nonverbal cues originally comprised within that all-encompassing umbrella concept is reorganised, along with others, into a richer set of dimensions (Coker & Burgoon, 1987). In addition to immediacy, involvement or engagement then include: *altercentrism* expressed nonverbally by the interlocutors' kinesics, proxemics, vocal warmth and consisting of displaying orientation towards the other, attentiveness, interested and adaptiveness; *expressiveness*, which is the degree of animation, dynamism and energy used in the conversation as manifested by facial expressiveness, vocal expressiveness, appropriate loudness/pitch and relaxed laughter; conversational *management*, consisting of smooth turn-taking as manifested in less silence, less time spent before responding to the interlocutor, a coordinated physical behaviour and an overall coordinated speech between interlocutors. Finally, involvement is nonverbally displayed as the lack of cues displaying *social anxiety*, such as self-manipulations of hands and arms, fidgeting, manipulation of objects, high pitch and tense posture and trunk orientation. Of all cues previously considered by the studies of immediacy, Coker and Burgoon seem not to

include the Touching dimension as essential to the definition of Conversation involvement (Coker & Burgoon, 1987; Richmond et al., 1987).

Overall, this literature identifies a set of behavioural, bodily and paralinguistic characteristics of a speaker expressing their level of engagement in the interaction with other speakers. Among these cues, I will focus in this work on the cues of engagement that can be detected regardless of the content of the speech exchanged and that are mainly visual and spatial. The reason why the verbal context has not been considered is that I would like the results of this work to be applied to as many settings as possible. Moreover, job interviews can use different verbal contents and forms depending on the work field considered. For this reason, I would like to construct an instrument that disregards the verbal content of interaction to identify whether the person is involved in the interaction or not. Likewise, paraverbal indices were excluded from this project. The reason for this is that the two roles (interviewer and interviewee) will have very different verbal tenses and the script of the interviewers will be predetermined. Furthermore, paraverbal cues would be difficult to detect and subsequently implement in the following project. Their identification, in fact, was carried out by a group of non-experts in nonverbal language, so as not to be conditioned by pre-existing theories. In the training that is planned to be carried out later, moreover, to include a part of training on the paraverbal would have been difficult to implement and would have risked taking away the authenticity of the interaction itself. I would also like the training to be usable by everyone, experts and non-experts, recruiters and candidates.

The goal of the study is to obtain a list of nonverbal cues that proves to successfully display engagement in a videoconferencing setting. The plan for this study is therefore as follows:

1) Nonverbal cues that express engagement during dyadic interactions will be identified. To do this, a bottom-up approach will be followed, so that cues are extracted without considering the verbal content of the interviews.

2) Using the identified cues, training will be constructed and administered to participants to improve their behaviour during job interviews. The effectiveness of this training on a behavioural level will be evaluated.

3) The effectiveness of the training in increasing perceived engagement will be evaluated.

4) Finally, engagement cues found and tested in previous studies will be automatically extracted using a machine learning system to provide the prototype of the core component of an online recruitment tool.

2.2. Nonverbal cues in interviews

Nonverbal cues, which are considered important components in selection interviews, have been studied for decades in the field of job recruitment. A study by Imada and Hakel (1976) found that eye contact, gestures, smiling, shorter interpersonal distances, attentive posture and more direct body orientation of a job applicant are perceived as indicative of a warmer and more enthusiastic person (Imada & Hakel, 1976).

A study by Imada and Hakel (1976) found that through the use of eye contact, gestures, smiling, proximity and a more direct body orientation, i.e. those which are characteristics of Immediacy, job applicants are perceived and described as warmer and more enthusiastic (Imada & Hakel, 1976). When a candidate in fact shows a nonverbal attitude of immediacy, he/she is perceived as having more desirable characteristics and is assessed as more liked, more qualified, more competent, more motivated, more successful, more satisfied if he/she had been given the position, more likely to be accepted and therefore

more likely to be recommended for a potential job (Imada & Hakel, 1976). Eye contact, fluency of speech, voice modulation and energy have also been shown to be discriminating cues for reviewing candidates for a second interview (McGovern, 1976). In a subsequent study (1977), eye contact was found to be associated with judgments regarding attentiveness, reliability, confidence, assertiveness, responsibility, initiative and also the final judgment of employability (Amalfitano & Kalt, 1977). Other colleagues showed that job applicants were rated more positively when they smiled, moved their heads and made eye contact (Forbes & Jackson, 1980; Young & Beier, 1977). Increased gestures have also been considered to be a cue leading to favourable evaluations for candidates (Edinger & Patterson, 1983). Other studies showed that behaviours such as gestures, as well as smiling and eye contact, may be related to more favourable interviewee evaluations (Imada & Hakel, 1977; Washburn & Hakel, 1973; Wexley et al., 1975). These studies on traditional face-to-face interviews were followed by studies on video-conference interviews, in order to adapt to the new means that are increasingly used in today's business world. An overview of these studies will therefore be presented in the next paragraph.

2.3. Nonverbal cues in video-interviews

The study of communication and social cues in videoconferences started in the last two decades of the twentieth century, mostly to check whether the lack of the cues available in face-to-face communication undermined the quality of the communication process and of the relations between the parties involved in it (Spears & Lea, 1992; Walther, 1996; Whittaker, 2002). The research then looked at the cues that were missing more than at those that were used, affected by a presumed superiority of face-to-face communication

reverberated in theories such as the rescued social cues (Sproull & Kiesler, 1991). Later research started to abandon that agenda and focused on the ways in which users orient to the cues and affordances available in the communication environment mediated by a technology (Arminen et al., 2016).

Several studies have considered the effect of nonverbal cues during interviews in videoconference platforms. Videoconference platforms have been found to be an appropriate environment to conduct qualitative interviews (Sedgwick & Spiers, 2009); they allow to convey non-verbal cues to the interlocutor, such as gaze, facial expressions, attention and some gestures (Iacono et al., 2016; Janghorban et al., 2014), while minimizing geographical barriers, transport problems and facilitating a larger sample of participants (Mirick & Wladkowski, 2019).

They have been found to be a more personal approach than telephone interviews (Irani, 2019), with ease of use and satisfaction sometimes greater than face-to-face and other interviewing media (Archibald et al., 2019).

Attempts have been made to automatically predict people's communication skills during dyadic interactions online, achieving an accuracy of 74% when considering nonverbal audio features and an accuracy of 83% when considering audio and video features together (Rasipuram et al., 2018; Rasipuram & Jayagopi, 2019). Furthermore, these non-verbal cues have been considered as central indicators of interpersonal behaviour in the field of social detection, giving the possibility to make inferences based on evaluations, performances or outcomes of the interaction itself. An example is the job interview, where the outcome of the interview can help to predict which behaviours describe a person's employability (Schmid Mast et al., 2015).

The cues that drive a recruiter to hire a candidate have been studied (Anderson & Shackleton, 1990; Nguyen et al., 2014), founding that the interviewers' impressions of

the candidates' hirability were dependent on the candidate's non-verbal behaviour and specifically their facial behaviours (Anderson & Shackleton, 1990; Naim et al., 2015) Nguyen and colleagues (2014) considered the possibility of predicting people's hirability using both visual and audio cues, finding that applicants' audio cues predicted their hirability, but visual cues did not. The authors suggested that this might be due to the fact that the raters took into account nonverbal behaviours other than those automatically extracted in this study when making their hirability assessments (Nguyen et al., 2014) The rationale for these results could also be that the evaluators relied primarily on audio rather than video features to make their assessments. Nevertheless, hirability is a concept that depends on different factors (i.e., the specific job position being applied for, the content of the interview and how the interview is conducted).

Naim and colleagues subsequently demonstrated how lexical (i.e., counting word categories as negative emotion terms or positive emotion terms), prosodic (i.e., speech rhythm and intonation), and facial (i.e., smiles and head movements such as nodding and shaking) features predict overall performance and hirability with a correlation coefficient of 0.70 and friendliness, engagement, and arousal with a correlation coefficient of 0.73 (Naim et al., 2015).

In the work environment, videoconference platforms to conduct job interviews enable the use of automated screening processes that assist the evaluation of the candidates' profiles. Various factors were analysed, including language, prosodic information, smiles and head gestures (Naim et al., 2015), head posture and eye gaze (Chen et al., 2016), proximity and frontal face events (Nguyen & Gatica-Perez, 2016), posture, gestures and eye contact (Rasipuram & Jayagopi, 2016), head nods and overall visual motion (Muralidhar et al., 2016). Basch and colleagues found lower performance scores in videoconferencing than in face-to-face interviews when considering eye contact, perceived social presence and

perceived impression management (Basch et al., 2020). It should be noted that social presence and impression management were assessed by the interviewees themselves and not by the interviewers. These results, therefore, might differ from the assessments of possible recruiters and lead to a different hiring decision. Nevertheless, in relation to eye contact, attempts have been made to improve gaze between people in therapeutic sessions on Zoom, acting on the angle of the webcam and the position of the interlocutors (Grondin et al., 2020).

Many of the cues observed by the recruiters concur to defining the engagement during the interview.

2.4. Related works

The aim of this project was to identify and make explicit the non-verbal cues that characterise engagement in an interaction, in order to use them to construct training that improves the skills of both the candidate and the recruiter involved in job interviews. I decided to work on nonverbal aspects during this type of interview, as the variability of verbal arguments is very large and depends on various factors, such as the type of interview and the job position for which the interview is being conducted. Specifically, I focused on body language, and how it can communicate a state of engagement or disengagement with the interlocutor. Given the pandemic in progress at the time of this study, it was decided to conduct interviews using Zoom, and to study this very medium, which holds great promise in the future world of work. In fact, this tool was one of the most used during 2020, the year the pandemic spread (Henry & Shellenbarger, 2020; Singh & Soumya, 2020) and people had to rely on remote means of communication to

continue working and interfacing socially without risking for their health (Marhefka et al., 2020; Singh & Soumya, 2020).

To date, many studies investigated nonverbal in employment interviews. For instance, Gifford and colleagues found that interviewers inferred the applicant's social skills from three nonverbal components: rate of gesturing, time spent talking, and formality of dress (Gifford et al., 1985). They also considered other behavioural cues, such as smiling, self-manipulation and object manipulation. However, their study did not find a correlation between these components and social skills during a job interview. Imada and Hakel showed that respondents who make more eye contact, smile, keep the body orientation towards the interviewer and less personal distance were perceived as more competent, more desirable, more motivated, and more successful. In fact, it seems that the immediacy (eye contact, smile, hand gestures, etc.) of the job applicant communicates a certain perceptual availability, which is then positively evaluated by the selecting employer (Imada & Hakel, 1976). Furthermore, Forbes and Jackson showed that more direct eye contact, more smiling and more nodding lead to greater employability (Forbes & Jackson, 1980). Also, Anderson and Shackleton found that more eye contact and more facial expressions during the job interview lead to a higher probability of being hired than those who maintain little eye contact and are less expressive (Anderson & Shackleton, 1990). However, none of the mentioned studies considered these cues as indices of mutual engagement between interviewer and candidate. Instead, our results show that facial expressions, gestures, gaze, eloquence, posture, body movements, head and hand movements, and backchannels are central in defining engagement.

More recently, attempts have been made to use computational systems to predict the hirability of job applicants. Nguyen and colleagues (2014) investigated the hirability, considering both verbal and nonverbal cues (Nguyen et al., 2014). In this study, hirability

was intended as consisting of five dimensions: the ability to communicate, persuade, work conscientiously, resist stress and the hiring decision score (based on the whole interview). Regarding the nonverbal (which is the focus of our study), authors considered head nods, overall visual motion, head region visual motion, smiling, gazing of both the applicant and the interviewer, and physical appearance of the applicant. They then found that candidates who showed more visual head motion received better ratings for hiring decisions and communication (Nguyen et al., 2014). In a subsequent study, Nguyen and Gatica-Perez attempted to predict interviewers' first impressions during job interviews using the same cues as in the above study. Participants were asked to rate two variables for general first impression (i.e. general hirability and general first impression), professional, social and communication skills and perceived personality. Results showed an automatic prediction of first impressions of up to 27% for extroversion, and up to 20% for social and communication skills (Nguyen & Gatica-Perez, 2016). Rasipuram and Jayagopi attempted to automatically assess communication skills during interface-based interviews. The cues used by the researchers are very similar to those identified in our study. These include posture, eye contact, body movements, head movements, facial expressions and gestures (Rasipuram & Jayagopi, 2016). The interview was carried out via an interface but was not conducted by a human interviewer. For our study, I decided to stick as closely as possible to what I think will be the practices in the future of most companies and workplaces and therefore used mediated communication (via Zoom). The studies mentioned above focused on the nonverbal cues of candidates in relation to their performance (i.e. hirability, professional skills, social skills, communication skills and personality) during the interview. None of them considered the engagement of the interviewee or the interviewer in the conversation and interaction. Partially, this was done by Naim and colleagues who with their experiment were able to predict with good

accuracy the excitement, engagement and friendliness rates given to the interview candidates (Naim et al., 2015). However, they only used facial features and did not identify nonverbal cues specific to engagement.

The studies mentioned so far were aimed at identifying and predicting nonverbal components that have a bearing on the recruitment decision of job applicants. However, none of those studies proposed training to improve candidates' (or recruiters') skills. An attempt in this sense was made by Muralidhar and colleagues, who created training on the basis of verbal and nonverbal behaviours during job interviews in hospitality (Muralidhar et al., 2016). Their training consisted of a feedback session in groups of three to eight students, who watched and commented on fragments of their first interview video. It started with a 20-minute presentation on nonverbal, followed by group discussion and concluded with personalised feedback written by human resources or hospitality professionals (Muralidhar et al., 2016). However, it is not clear from this work what actual recommendations were given to participants to improve their performance. Also, what nonverbal cues were used to create the training given to students to improve their skills is not specified. Finally, the effectiveness of the training was evaluated on the basis of the impressions (evaluations of the first two minutes of the videos) of a group of five students, finding a difference between the overall impressions before and after the training. Nonetheless, there was no control group, so this result may not be due specifically to the training.

2.5. Project and state of the art

A combined approach will be adopted in the present project, which will first identify the cues affecting the impression that speakers make, and then train the users to display them to see if the impression changes. I will focus on one dimension of performance that can be detected by nonverbal cues, i.e., the speakers' engagement, because it applies to any kind of interview regardless of its purpose and content.

Overall, this literature identifies a set of behavioural, bodily and paralinguistic characteristics of a speaker expressing their level of engagement in the interaction with other speakers. Among these cues, I will focus in this work on the cues of engagement that can be detected regardless of the content of the speech exchanged and that are mainly visual and spatial (Table 3). Despite the uniformity of the terminology with which nonverbal cues are identified in the literature, the studies from which they emerge are carried out with various methodologies and under different assumptions, not always consistent with the approach needed here. Even considering only the studies in which the speaker is displayed in a video or videoconference, such variety persists. First, the epistemic role of non-verbal cues varies across studies, since they are sometimes taken as an expression of the speakers' inner feeling and sometimes as performative resources for the management of impressions. A study by Imada and Hakel (1976) found that eye contact, gestures, smiling, shorter interpersonal distances, attentive posture and more direct body orientation of a job applicant are perceived as indicative of a warmer and more enthusiastic person (Imada & Hakel, 1976). Other studies showed that behaviours such as gestures, as well as smiling and eye contact, may be related to more favourable interviewee evaluations (Imada & Hakel, 1977; Washburn & Hakel, 1973; Wexley, Fugita, & Malone, 1975). Sometimes the categories are taken for granted and sometimes

they are themselves the object of investigation. For example, Nguyen and colleagues (2014) used cues known from the literature to predict the hirability of interview candidates. To annotate these cues, they used an audio-visual material recorder during job interviews. They found that the audio features of the applicants and the visual cues of the professionals were predictive of hirability. The audio features they studied were *speaking activities* like speaking time, speaking turns, pauses and short utterances, and *prosody* elements like the energy, the perceived fundamental frequency and the voiced rate. In terms of visual features, they considered *head nods*, extracting the number of nods and the total time of nodding, *overall visual motion*, as an indication of expressiveness, *head region visual motion*, *smiling, gazing* and *physical appearance* (i.e., how attractive the person was perceived to be) (Nguyen et al., 2014).

Finally, the nonverbal cues are usually presented during a hearable speech; so the contribution of verbal and nonverbal cues to the speaker's impression is hard to sort out. Indeed, speech and gestures converge to define the meaning of the speaker's utterance, so that being the speech available, it will be hard to ignore in the formation of impressions on the speaker. Rasipuram and Jayagopi (2016) collected videos of 106 interviews and had them annotated by three naive evaluators. The participants assessed how adequate the interviewees were with regard to the dimensions of speech activity, prosody and nonverbal cues such as posture, eye contact, gestures and head movements. These videos were presented with both audio and video, thus not ensuring that ratings regarding nonverbal behaviours were not influenced by the content of the conversation (Rasipuram & Jayagopi, 2016). Nguyen and colleagues (2014) collected 62 videos of interviews of applicants for a marketing job. They then had two students assess the hireability of the candidates and attempted to predict hireability scores from verbal and non-verbal features automatically extracted from the videos. The results show that combining verbal and non-

verbal cues together decreases the likelihood of predicting hireability scores compared to verbal cues alone (Nguyen et al., 2014). This could be due to the fact that assessors considered only verbal content for their ratings or were more influenced by it than nonverbal content.

3. Project structure

1. First study: Identification of nonverbal cues of engagement

The goal of this study was to identify recurrent movement patterns of parties involved in interaction from the point of view of physical mutual engagement. By Physical Mutual Engagement (PME) I mean a state of a reciprocal compelling interest in which people seem to have a constructive and positive interaction, not only a state of workflow. Participants watched four muted videos and answered engagement-related questions about the videos. A content analysis was conducted on the answers given by the participants to the open-ended questions, in order to identify the behaviours cues of the PME.

2. Second study: cues validation

The aim of this study was to validate the non-verbal engagement cues identified in the previous study. To do this, training was constructed and administered to 20 participants involved in a job interview. The training was administered to half of the participants. The other half were in a control condition or only watched the video of the first round of questions. In order to check the effectiveness of this training, the participants' behaviour was annotated and compared by experimental condition.

3. Third study: Engagement during video-interviews validation

The goal of this study was to verify whether the parameters found in the previous study could effectively improve physical mutual engagement (PME). The videos collected during the second study have been evaluated by an independent commission in order to determine whether the training had an effect on the participants' engagement. Participants were shown both videos of one of the trainees (first and second round of questions) and asked to rate the level of engagement of the trainees for each video. A comparison was then made between pre and post-interval videos evaluations.

4. Fourth study: Engagement and behavioural cues correlation

The aim of this study was to check whether the behavioural cues annotated during the second study correlated with the engagement values assessed by the external evaluators during the third study. To this end, the average engagement scores found in the third study were compared with the four behavioural cues annotated in the second study.

5. Fifth study: automatic detection of behavioural cues

The aim of this study was to build a model capable of extracting and predicting the nonverbal cues found and tested in our previous studies, using state-of-the-art machine learning algorithms. In order to test which system could predict the behaviours identified in the first study, the annotation of the second study has been considered. The annotated frames of the videos were used to train and test the model using a network for facial recognition. Predictions were then compared with the annotated labels.

General conclusions regarding the work carried out are therefore reported. Implications and future developments are then discussed.

4. Preliminary study: Identification of nonverbal cues of engagement

4.1. Objective

On the background of the categories established so far as relevant to identify nonverbal cues, I wanted to add preliminarily an exploration of the nonverbal cues that are of specific interest to our purposes. I am interested in nonverbal cues of engagement considered by a third party when watching speakers in a video interview. I, therefore, collected a small set of movie segments showing an interview (healthcare or job interviews) and asked participants to assess the engagement of the parties involved in the interaction as well as the bases of such assessment.

4.2. Materials

The material used in this preliminary study consists of a segment of videos depicting a dyadic interview in a work setting (healthcare consultations and job interviews) and clearly showing the body movements of both parties involved. The characters in the videos were a professional and a client/applicant. Eight (8) videos were retrieved online (YouTube) and edited to eliminate pauses, mute the audio, and generate shorter videos (max 1.30 min): four (4) healthcare consultations and four (4) job interviews. These videos were either extracts from films, TV series or videos uploaded by YouTube users.

Table 1: Video content

	Video content
Video 1	Therapeutic consultation
Video 2	Medical consultation
Video 3	Therapeutic consultation
Video 4	Medical consultation
Video 5	Job interview
Video 6	Job interview
Video 7	Job interview
Video 8	Job interview

A preliminary pilot study highlighted the need to shorten the experiment to keep participants focused until the end. Participants were therefore presented with only 4 of the 8 videos, one therapeutic consultation, one medical consultation and two job interviews (videos 1-3-5-7 and videos 2-4-6-8), these groups of videos were administered randomly to participants (Table 1).

An ad hoc survey made of both open-ended and closed-ended questions was constructed using SurveyMonkey, investigating participants' opinions about the interactions, evaluations about the interlocutors' physical engagement and evaluations about the professionals' attitudes. All the questions were asked for each video:

I. Indicate the party looking more engaged in the interaction. The answer was provided on a 10-point scale, where the two interlocutors were at the extremes. A score of -5 attributed a greater engagement to interlocutor A, while a score of +5 attributed a greater engagement to interlocutor B; 0 indicated a perception of an

equal engagement of the two interactants. This was a warm-up question to prepare the next, open-ended questions of actual interest to us.

II. Explain what in the parties' physical behaviour (and not in the content of the dialogue) account for the opinion expressed in item 1. This was an open-ended question.

III. Say which of the two [parties] should change something (and what) of their behaviour to show more involvement? This was an open-ended question.

4.3. Participants

Participants asked to judge the videos were university students not enrolled in a psychology course (for potential non-verbal coding prior experiences) and with good or correctable eyesight that did not prevent them from seeing the videos properly. Thirty 30 students were recruited, 15 females and 15 males, aged between 19 and 28 years.

4.4. Procedure

Participants filled out the informed consent, then they were asked to watch four muted videos and then answer some questions about the videos just seen. Each participant was assigned a code in order to guarantee anonymity and thus the free expression of his or her opinion on the behaviour observed in the videos. After that, they were seated on a chair in front of a pc screen with a keyboard and a mouse. Then, they were presented with the muted videos (randomized) and, after each video, the questions. The whole procedure took around 30 minutes.

4.5. Results

The data collected describe the perceived engagement of the characters in the interview with their interlocutor in the video clips, as well as the cues leading to that perception. The evaluation of engagement of the two parties in the movies, elicited by the first item, was not of interest to us; the first question was meant to warm up the participants, forcing them to rate the characters in the movie segments in terms of engagement and therefore to focus on the cues enabling such assessment. The data analysed in this study were thence the answers to the open questions, which were exported from the survey platform in which they were collected (SurveyMonkey) and then imported as a text file in ATLAS.ti (version 8.4.4) to be manually coded. The answers given by the participants to the open questions underwent a two-stage thematic analysis combining an inductive and a deductive process (Braun et al., 2017; Corbin et al., 2008), supported by the software. Each answer was treated as a text unit, for a total of 240 text units.

In the initial inductive phase of the thematic analysis, each answer was analysed to find any mention to nonverbal cues of engagement as well as to any specific dimension of engagement referred to in the answer. For instance, in the answer "*he welcomes the interlocutor B in a positive way, shakes his hand, invites him to sit down and moreover he is the one who starts and maintains the conversation*" one nonverbal cue was the invitation gesture, and the specific dimension of engagement was the welcoming. In this way, two sets of categories were progressively created, one organizing the cues mentioned by the participants, and the other organizing the meaning of the cues, all in the participant's words. Eventually, 57 cues were found subsequently grouped into nine overarching groups: eloquence, facial expressions, gestures, gaze, posture, body movements, head movements, hand movements and backchannel (Table 3). In a

subsequent, deductive phase of the thematic analysis, each unit of text was coded using eight meaning categories:

- **Affiliation/disaffiliation:** the speaker displayed a (dis)approving attitude, e.g. “Interlocutor B should not look away while the other is talking to him and should not shake his head as if in mockery when answering a question.” ("Interlocutore B dovrebbe non guardare altrove mentre l'altro gli parla e non scuotere la testa mentre come per presa in giro quando risponde ad una domanda).
- **Welcoming:** the speaker appeared to welcome the interlocutor, e.g. "he welcomes the interlocutor B in a positive way, shakes his hand, invites him to sit down and moreover he is the one who starts and maintains the conversation (accoglie l'interlocutore B in modo positivo, gli stringe la mano, lo invita ad accomodarsi e inoltre è lui a iniziare e mantenere la conversazione)."
- **Interest:** the speaker seemed focused on the interlocutor, e.g. "Both looked at each other and both seemed interested in listening to each other (Entrambi si guardavano ed entrambi sembravano interessati ad ascoltarsi)."
- **Emotion**: the speaker showed signs of emotional involvement, e.g. "The interlocutor B by his facial expressions, gaze and movements was able to communicate emotional involvment, to look concerned (L'interlocutore B con le espressioni facciali, gli sguardi ed il modo in cui si è mosso ha fatto trasparire un coinvolgimento emotivo, mostrandosi preoccupato)."
- **Respect:** the speaker gives signs of respect or disrespect, e.g. "Interlocutor A should stop eating the gum (L'interlocutore A dovrebbe smettere di mangiare la gomma…)."
- **Asymmetry:** the speaker seemed to look down on the interlocutor, e.g. "B could have looked at A in the eyes, perhaps sitting at the same height instead of standing

or looking out of the window (B forse avrebbe potuto mettersi ,fin da subito, nelle condizioni di guardare A negli occhi, magari sedendosi alla stessa altezza invece che rimanere in piedi (o a guardare fuori dalla finestra)."

- **To appear warm/cold:** warm or cold attitude towards the other interlocutor, e.g. "Interlocutor B hardly ever looks at interlocutor A, nor smiles at her, showing a rather cold and detached attitude (L'interlocutore B non guarda quasi mai l'interlocutore A, ne le sorride mostrando un atteggiamento alquanto freddo e distaccato)."

- **To pay/not to pay attention:** paying attention or showing attention to the interlocutor, e.g., "A looks around and does other things, like flipping through and reading what's in front of him/her" (A si guarda intorno e fa altre cose, come sfogliare e leggere quello che ha davanti a sé)."

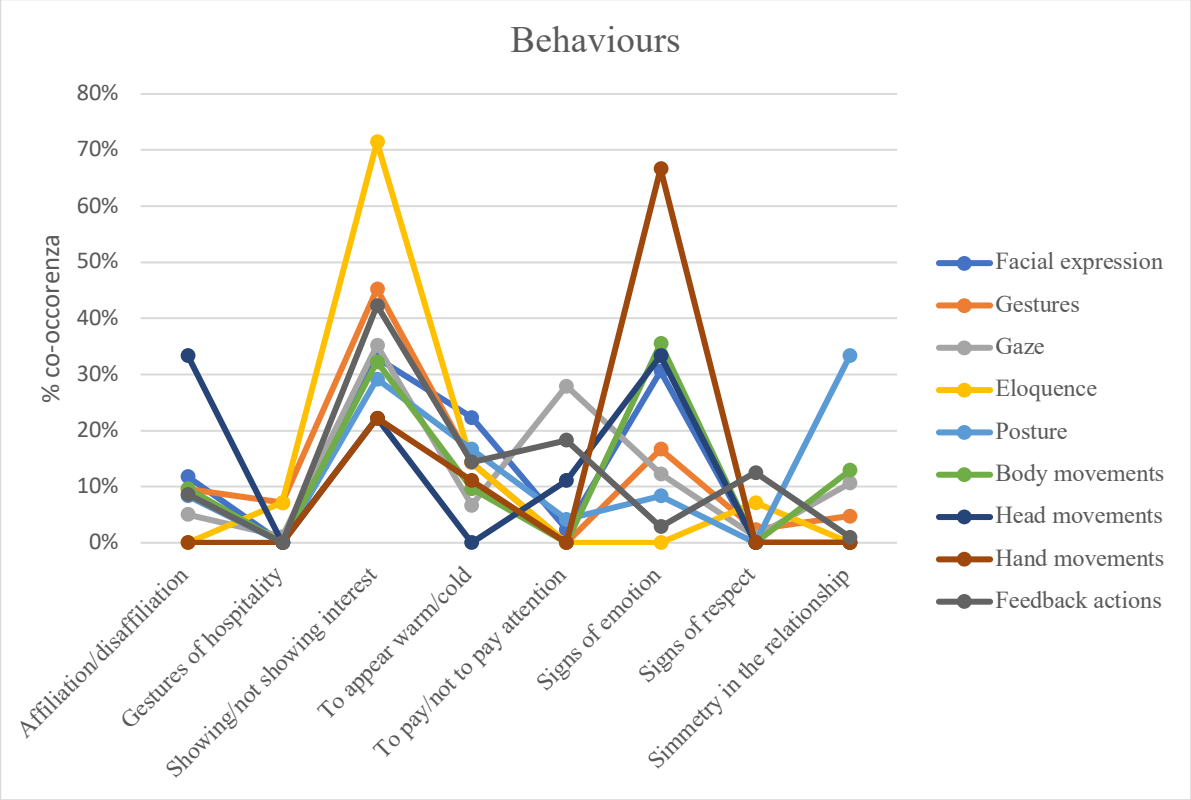
This coding was done independently by the author and by a coder who did not participate in the prior bottom-up phase. This latter coder, a student doing her internship in our lab, was trained to use the codes on 30% of the quotes (100 quotes), and then categorized the remaining quotes independently, for a total of 331. The agreement between the two judges' encodings was 58,31%. Their coding was then compared to solve the disagreements and have the full set of text units coded. The two judges confronted each other on each unit of text that they had coded differently and, after a brief discussion, reached an agreement on which of the two codes to keep and which to reject. Breaking down the dimensions of engagement referred to in the answers provided a parameter to select which nonverbal cues to use in the main study. I was looking for nonverbal cues that mapped into as many dimensions as possible so that training a person in using these cues would affect the highest possible number of dimensions of engagement with the minimum effort. I then calculated the distribution of each nonverbal cue across the different subdimensions of engagement (Table 2).

Table 2: Behaviours composition

	Affiliation/ disaffiliation	**Gestures of hospitality**	**Showing/not showing interest**	**To appear warm/cold**	**To pay/not to pay attention**	**Signs of emotion**	**Signs of respect**	**Symmetry in the relationship**
Facial expression	12%	0%	33%	22%	2%	31%	0%	0%
Gestures	10%	7%	45%	14%	0%	17%	2%	5%
Gaze	5%	1%	35%	7%	28%	12%	1%	11%
Eloquence	0%	7%	71%	14%	0%	0%	7%	0%
Posture	8%	0%	29%	17%	4%	8%	0%	33%
Body movements	10%	0%	32%	10%	0%	35%	0%	13%
Head movements	33%	0%	22%	0%	11%	33%	0%	0%
Hand movements	0%	0%	22%	11%	0%	67%	0%	0%
Backchannels	9%	0%	42%	14%	18%	3%	13%	1%

After this analysis, it appears that the more transversal groups of nonverbal cues of engagement were Facial expressions, Backchannels and Gaze (Figure 1). Within those categories, I decided to implement into the subsequent training those marked with an asterisk in Table 3; because they are visible in video interviews since they refer to the face and upper part of the torso and are easy to implement in a small training such as the one I wanted to carry out in the next study.

Figure 1: Behaviours composition



In addition, facial expressions such as smiling and laughing have been found to be essential in defining involvement and immediacy (Coker & Burgoon, 1987; Richmond et al., 1987), mutual gaze and eye contact have been found to increase engagement and immediacy (Richmond et al., 1987; Szafir & Mutlu, 2012), nodding was found to be positively correlated with immediacy in the teaching environment (Richmond et al., 1987; Szafir & Mutlu, 2012), and smooth speech co-ordination and turn management were considered to be important factors of both conversational involvement and immediacy (Coker & Burgoon, 1987; Richmond et al., 1987).

Table 3: Nonverbal cues displaying (dis)engagement according to the survey participants. An asterisk marks those selected for the training in the main study.

Facial expressions	
Eyelift	Closed eyes
Eyebrow lift	Half-closed eyes
Facial expressions	Winking
Smile*	Crying
	Laughter*
Gestures	
Broad gestures	Generic gesture
Kick out (dismissive) gesture	Invitation to sit down
Shut up gesture	Waving
	Way of handing out tissues
Gaze	
Avoiding eye contact	Looking into the eyes*
Gaze	Looking towards the exit
Looking elsewhere	Looking over your glasses
Looking down	Searching for eye contact*
Looking towards the interlocutor*	
Eloquence	
To talk more	To talk little
To talk a lot	To talk fast
Posture	
Similar position of interlocutors	Relaxed posture
Posture	To sit

Composed posture	Body orientation towards the other
Curved posture	Arms position
Rigid posture	To stand
Body movements	
Body movements	Proxemics
Rigid movements	Torso twist
Head movements	
Bowed head	Head movements
	Head shaking
Hands' movements	
Steady hands	Hand movements
Backchannels	
To write down with care	Letting the interlocutor speak*
To nod*	Chewing a gum
To do other things*	Answering the phone
To interrupt interlocutor*	Taking off the hat
	To leave

4.6. Discussion

The thematic analysis carried out on the answers given by participants revealed 57 nonverbal cues expressing engagement between two people in dyadic interaction with each other. These 57 cues were grouped into 9 main behaviours: eloquence, facial expressions, gestures, gaze, posture, body movements, head movements, hand movements and backchannel. By Eloquence, I referred to the nonverbal categories of speech that were identified by the participants in the experiment, such as speed of speech or time spent speaking. Speech characteristics have often been considered in the literature

as components of engagement between interlocutors. These components of language are often referred to as *prosody* and have also been found to be essential in the field of hirability (Nguyen et al., 2014; Rasipuram & Jayagopi, 2019). Furthermore, by facial expressions, I referred to all those movements of the facial area that the participants identified. Facial expressions, like smiles, laughers and general expressiveness have been found to be essential components of immediacy and conversational involvement (Coker & Burgoon, 1987; Richmond et al., 1987). By gestures, I identified those behaviours which the participants themselves defined as "gestures" and which were expressed through body movements. Gestures, often defined as the amount of gestures used while speaking (Richmond et al., 1987; Szafir & Mutlu, 2012), seems to lead to greater sympathy and cooperation (Andersen et al., 1979; Mehrabian, 1971). By Gaze, I defined all those eye movements that were pointed out by the participants. Gaze was also found to be an essential component of engagement during a conversation, with people being more engaged when a gaze was held between interlocutors and eye contact was held further apart and less involved when the gaze strayed elsewhere (Mehrabian, 1968b; Richmond et al., 1987; Szafir & Mutlu, 2012). Posture was defined as the way in which the interlocutors held their bodies. Therefore, a similar stance of the interlocutors, sitting or standing, body orientation and posture state ("relaxed" or "tense"). A relaxed posture, in fact, has been found to be connected with teacher immediacy (Andersen, 1979; Richmond et al., 1987) and to be indicative of one interlocutor's attitude towards the other, as immediacy seems to be directly related to positive attitude for moderately/very relaxed postures, but negatively related to positive attitude for less tense postures (McGinley et al., 1975; Mehrabian, 1968b, 1968a; Szafir & Mutlu, 2012). Under Body movement, I grouped all those characteristics of the interlocutors' body movements in space described by the participants in the experiment. Thus proximity is the way of moving the body

("rigid" or "flexible") and the torsion of the trunk. Body movements have been identified in the literature as touching the interlocutor (Mehrabian, 1968b; Richmond et al., 1987), leaning forward (Mehrabian, 1968b), distance from the interlocutor (Mehrabian, 1968b; Mehrabian & Friar, 1969), also referred to as *proximity* (Richmond et al., 1987; Szafir & Mutlu, 2012), and body orientation (Mehrabian, 1968a). With regard to head and hand movements, I grouped together all those text units in which the participants referred to these two components of nonverbal language to justify the involvement or non-involvement of the interlocutor. In the literature, head and hand movements have been found to be central in defining interlocutors' attitudes. As far as the head is concerned, its orientation (towards the interlocutor or elsewhere) but also its position (bowed or upright) has been found to be an essential indicator for attitude definition. Furthermore, the position of the hands and arms were found to be good discriminators between postural attitudes (Mehrabian, 1969). As far as backchannel is concerned, I grouped under this category all those text units that included nonverbal feedback given to one's interlocutor regarding one's state of involvement in the interaction (Moran et al., 2015). This category included nodding movements, taking notes or doing other things (distraction) and whether or not to interrupt the speaker. In the literature, nodding has long been considered an indicator of immediacy (Andersen, 1979; Richmond et al., 1987), generally considered to be engagement feedback provided by the listener to the speaker (Nakano & Ishii, 2010). In addition, conversational management, characterized by a coordinated speech between interlocutors and smooth turn-taking, has also been deemed as a central element of conversational involvement (Coker & Burgoon, 1987; Richmond et al., 1987).

The same text units labelled with the 57 cues were then labelled with 8 meanings by two independent evaluators: affiliation/disaffiliation, gestures of hospitality, showing/not showing interest, to appear warm/cold, to pay/not to pay attention, signs of emotion, signs

of respect, symmetry in the relationship. These 8 meanings were extracted from the participants' answers using a ground-based methodology. The aim of this grouping was to check which behaviours were associated with the greatest number of meanings. It was found that the three most cross-sectional behaviours were Facial Expressions, Gaze and Backchannels.

5. Second study: cues validation

5.1. Objective

The aim of this study was to validate the non-verbal cues of engagement identified in the previous study. To do this, a training session was constructed and administered to participants involved in a (simulated) job interview. The training was based on the cues identified in the first study and was aimed at improving the behaviour of the participants who would receive this training. A structured interview was then designed and administered to 20 participants. These interviews were carried out together with a short training session, administered to half of the participants. In order to verify the effectiveness of this training, the behaviours of the participants were annotated and compared by experimental condition.

5.2. Materials

5.2.1. Video collection

Participants recruited online were sent a survey, administered via Google Forms, containing the informed consent and some biographical data questions. They were then given different information sheets according to whether they would be 'Interviewee' or 'Interviewer'. When the trainee was the Interviewee, the instructions were as follows: *"This study concerns the effectiveness of training to improve certain behaviours during videoconferencing interviews. You will be asked to take part in a two-way interview on Zoom. Your role will be as interviewee: you will have to prepare two short topics that you will then present to the interviewer. You will also have the opportunity to reflect on*

the progress of the interview with an expert." When the trainee was the Interviewer, the instructions will be as follows: *"This study concerns the effectiveness of training to improve certain behaviours during videoconferencing interviews. You will be asked to take part in a two-way interview on Zoom and your role will be as interviewer. You will have to listen and ask a few questions to a person who is presenting a short topic. You will also have the opportunity to reflect on the progress of the interview with an expert."* The biographical data requested included gender, age and spoken language. They were also asked whether they had used Zoom or other video conferencing systems recently. Subsequently, documents were sent to the participants containing the questions they would be asked if they were Interviewees and the questions they would have to ask if they were Interviewers. When the trainee was the Interviewee, the instructions were as follows: *We ask you to prepare answers to the following questions for the call that will take place via Zoom:*

A) Describe a work experience or non-academic activity in which you participated (e.g., a competition) and what you think it gave you.

B) Tell and describe a teacher or colleague who has had an influence on you and how.

We ask you to be as thorough as possible in answering the questions. During the call, you will be able to keep notes with you (in paper or digital form) to be used to answer the above questions verbally. When the trainee was the Interviewer, the instructions were as follows: *During the Zoom call, you will have the role of Interviewer. We therefore ask you to ask the interviewee the following questions:*

A) Describe a work experience or non-academic activity in which you have participated (e.g., a competition) and what you think it gave you.

B) Tell and describe a teacher or colleague who has had an influence on you and how.

We ask you to ask the first question in the first phase of the Zoom call, the second during the second phase.

We also ask you to ask the following questions to the interviewee during his/her presentation:

Questions A:

- How long did this experience last?

- Is it an experience you would repeat?

- Would you recommend this type of experience to others?

Questions B:

- Do you think this person was a role model? If so, why?

- Do you think this person's influence has changed the way you do things?

- Do you think this type of behaviour is rare?

You can keep these questions on a piece of paper or digital file for use during the Zoom call.

The order of questions A and B, and their subsequent specification questions, was randomised. Half of the participants, therefore, received question A as their first question and the other half received question B as first. This was done to prevent the order of the questions from having an effect on the participant's behaviour and therefore it would have not been possible to establish whether any improvement was due to the training or to the questions themselves.

Participants who underwent the training (50%) were shown a list of behaviours, before watching the video of their interview, and were asked to identify these behaviours if they occurred during the first round of questions. These behaviours were: smile, look at the interlocutor, nod, interrupt the interlocutor, distraction, signs of fatigue and do other

things. The experimenter then gave the participant some suggestions to improve the following interview:

- Move your webcam in front of you
- Move the zoom window as close as possible to your webcam
- Directly look at the webcam as you had your interlocutor in front of you
- Enlarge the zoom window if this guarantees you more eye contact
- Look as little as possible at your notes
- Smile
- Nod

5.2.2. Video evaluation

The materials used were the forty (40) videos collected, the eighty (80) annotations made on these videos, consisting of the frames annotated for each of the four behaviours, and the data from the sample of participants collected during the interview.

5.3. Participants

For the video collection of the study, the trainee's sample was made of 20 students recruited online, who wanted to improve their ability to use the Zoom platform (9 M – 11 F). The 50% (10) of the participants were in the Training condition, 25% (5) in Video-rewatch and the remaining 25% (5) in the No training condition (Table 4).

Table 4: Video collection sample

	Male	Female	Total
Training	6	4	10
Video Rewatch	2	3	5
No Training	1	4	5
Total	9	11	20

5.4. Procedure

5.4.1. Training Design

Based on the previous study, three groups of transversal behavioural cues expressing engagement were selected: Facial expressions, Backchannels and Gaze (Figure 1). Within these categories, I decided to implement those marked with an asterisk in Table 3 in the following training; because they are visible in the video interviews as they refer to the face and upper torso and are easy to implement in a small training like the one I wanted to carry out in this study. In addition, as expressed in section 4.5, facial expressions such as smiling and laughing were found essential to define engagement and immediacy, mutual gaze and eye contact were found to increase engagement and immediacy and nodding was found positively correlated with immediacy in the teaching environment. The selected cues were associated with suggestions to be given to the participants. With regard to Gaze, the suggestions were "Move the webcam in front of you", "Move the zoom window as close as possible to your webcam", "Look directly at the webcam as if you had your interlocutor in front of you" and "If needed, enlarge the zoom window of your interlocutor". Regarding Facial Expressions, the advice was "Smile more, in

accordance with what your interlocutor is saying". Regarding Backchannels, the advice was "Nod more, according to what your interlocutor is saying" and "Look as little as possible at your notes". In addition, the list of behavioural cues was prepared to be shown to the participants who would carry out the training. In fact, these participants, after the first interview, would be shown the list of cues and asked to check - while watching the video recording - which of those actions they did or did not do during the interview: Smile, Watch interlocutor, Nod, Interrupt interlocutor, Do other things, Distracting.

In addition, to test the validity of our training, two more experimental conditions were added. In total, therefore, the participants would be divided into three experimental conditions. Training (T), in which participants would watch their first interview and receive advice from the experimenter. Video Re-watch (VRW), in which participants would watch their first interview without receiving any advice. And finally, No Training (NT), in which participants would neither review the interview nor receive advice.

5.4.2. Video collection

To test the in-the-wild validity of the identified cues of Mutual Engagement, training was constructed and administered via Zoom to participants involved in a videoconference interview. Participants were randomly assigned to one of three experimental conditions: Training (T), Video Re-watch (VRW) and No Training (NT). As previously discussed, of the twenty (20) interview participants, 50% (10) received our training between two rounds of questions, 25% (5) watched the video of their first round and then carried out the second one, and the remaining 25% (5) went on to the second round without any intermediate activity (Table 4). Furthermore, participants were randomly assigned the role of Interviewer (IR) or interviewee (IE). The video collection has been divided into

three phases: first round, interval, second round. In each interview, there was a participant, recruited online, and the confederate, collaborator of the principal investigator but not informed of the purpose of the video collection. When the participant was assigned the role of Interviewer (IR), the confederate played the role of Interviewee (IE); when the participant was assigned the role of Interviewee (IE), the confederate played the role of Interviewer (IR). The training (or video re-watch or no training) has always been addressed to the recruited participants, who in condition IR-T, IR-VRW and IR-NT was the Interviewer and in condition IE-T, IE-VRW and IE-NT was the Interviewee. The experiment followed a 3x2 design, with the activity performed during the interval and the role of the participant as independent variables (Table 5).

Table 5: Experimental conditions

Participant's role	Training	Video re-watching	No training
Interviewer	IR-T	IR-VRW	IR-NT
Interviewee	IE-T	IE-VRW	IE-NT

Conditions:

IR-T

IR-VRW

IR-NT

IE-T

IE-VRW

IE-NT

The Zoom sessions were video-recorded so that the videos could be used both as part of the procedure in the Training and Video Re-watch groups and also for further annotation.

The Interviewer (IR) asked the Interviewee (IE) two questions A and B (of which the answers had been prepared in advance). The questions asked to the Interviewee (IE) were also randomized, in order to avoid the effect of the questions themselves on the spontaneous reactions of the participants. During the first round, the confederate/participant was asked to make a five-minute presentation answering one of the following questions about him/her and his/her vision of work/study he/her received in advance, using **notes**:

A) Describe a work experience or non-academic activity in which you participated (e.g. a competition) and what you think it gave you.
B) Tell and describe a teacher or colleague who has had an influence on you and how.

In the interval phase, 50% of the participant was administered training on non-verbal engagement cues, explaining only that this was training to improve their interactions on Zoom, using the list of behaviours and suggestions in the Materials section (4.2.1.), while re-watching the video of his/her performance. Another 25% of the participants re-watched him/her-self without any comment from the experimenter. In a third (control) condition, the participant was shown nothing and proceeded to the second round.

The training focused on three nonverbal behaviours, found to be relevant in our previous study: Facial expressions, Gaze and Backchannels. The participant was instructed:

- Smile accordingly to your interlocutor and what he or she is saying (Facial expression)
- Move your webcam in front of you and look directly into it; Move the interlocutor's window as close as possible to your webcam (Gaze)
- Nod to let your interlocutor know that you are listening and understanding what they are saying

- Look as little as possible at your notes (Backchannel)

During the second round, the Interviewee (IE) was asked to present the answer given to the second question to the Interviewer (IR) using notes he/her already has.

5.4.3. Video evaluation

To test whether our training was effective in increasing the frequency of behaviours considered crucial for Mutual Engagement, the video recordings of the interviews were analysed. In total, 40 (forty) videos were collected, two for each participant, consisting of the first and second round of questions. Of the behaviours used for the training in the video collection (smile, look at the interlocutor, nod, interrupt the interlocutor, distraction, signs of fatigue and do other things), Look at the camera, Look away, Nod and Smile were selected. This is because it was considered possible to annotate these behaviours by frame and identify them with certainty. These behaviours were then annotated frame by frame on the 40 videos collected. Look at the camera was annotated whenever the participant looked in the direction of the webcam (i.e., when looking in the direction of the annotator), Look away was annotated whenever the participant looked away. These two behaviours, therefore, were mutual; participants throughout the entire interview looked either towards the camera or away. In addition, Look away was annotated whenever the participant made a nodding gesture with the head, i.e. from top to bottom. Finally, Smile was annotated whenever the participant smiled, i.e. when the corners of the mouth arched upwards.

In order to test whether the training had therefore been successful in increasing the highlighted behaviours, a comparison within the four behaviours (i.e., Look at the camera, Look away, Nod and Smile) and between the three experimental conditions (i.e.,

Training, Video Re-watch, No Training) was carried out. The total frequency of each of the four behaviours was calculated and weighted by the total number of frames in the video (this was done to prevent the length of the video from affecting the frequency of the behaviours). Statistical analyses were then carried out to compare the frequencies of the four behaviours under the three experimental conditions.

In this study, therefore, it was expected that:

- The frequency of Looks into the camera would have been significantly higher for people who had performed the training when compared to the Video Re-watch and No training conditions. Consistently, I hypothesized that Looks away would have been significantly lower for participants who had undergone our training when compared to the Video Re-watch and No training conditions (Hypothesis 1).
- In addition, I expected the differentials in the frequency of Smiles between the first and second interviews to be significantly higher in the Training condition than in the Video Re-watch or No Training conditions (Hypothesis 2).
- Furthermore, it was expected that the Nodding behaviour would have been significantly more frequent for people who had done our Training compared to the other two conditions (Hypothesis 3).
- Also, I did not expect to find statistically different frequencies of gaze behaviours between those who had played the role of Interviewer and those who had played the role of Interviewee (Hypothesis 4).
- Finally, I expected to find statistically higher frequencies of nod and smile behaviours between those who had played the role of Interviewer and those who had played the role of Interviewee (Hypothesis 5).

5.5. Results

To test whether the training had an effect in increasing Mutual Engagement related behaviours (i.e., Look at the camera, Look away, Nod and Smile), the frequencies of the four cues for each video were calculated, expressed as the total number of frames in which this behaviour occurred, divided by the total number of frames in the video ($BehavFreq = \frac{TotBehavFrames}{TotFrames}$). This was done to avoid the length of the video having an effect on the relevance of the behaviour itself. Behaviours' frequencies were expressed as percentages and the differentials between the second round of questions' behaviours and the first round of questions' behaviours were calculated ($\Delta Behav = BehavFreqSecond\% - BehavFreqFirst\%$). Statistical analyses were therefore carried out on the differentials to see whether there had been any variation between the first and second rounds of questions and, if so, under which conditions and which behaviours.

5.5.1. Training vs. Video re-watch vs. No Training

5.5.1.1. Look into the camera

To test hypothesis 1, the differential between the frequencies of looks into the camera of the first round of questions and the second was considered as the dependent variable ($\Delta LookCamera$). Condition 1 (Training, Video Re-watch, No Training) was considered as an independent variable. Kruskal-Wallis analysis of variance was then used. As sown in Table 6, analyses resulted in a significant difference between the three conditions Training, Video Re-watch and No training (Chi-square = 12.094, p = .002**, df = 2).

Table 6: Look camera - Condition 1

Comparison	chi-squared	df	p-value
$\Delta LookCamera \sim$ Condition 1	12.094	2	0.002365**

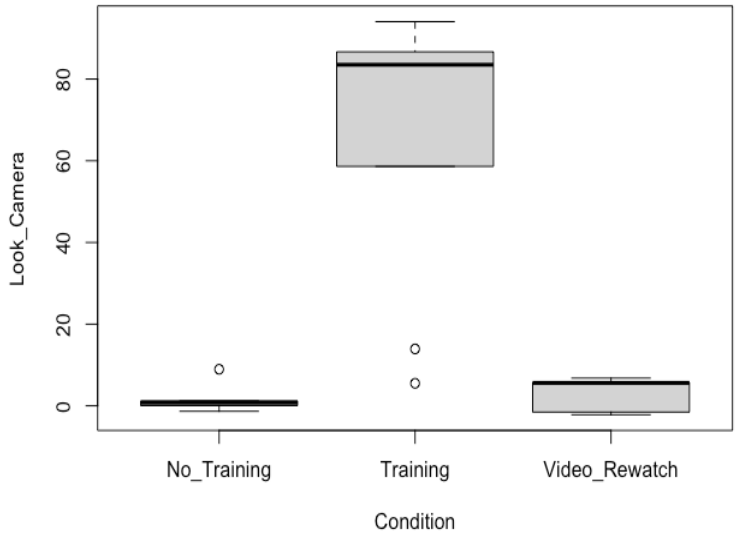
Dunn's posthoc test was then used to compare the differentials in the three pairwise conditions and check which were significantly different from each other. Before proceeding, the p-values were adjusted using the Benjamini-Hochberg method.

Table 7: Look camera - Condition 1

Comparison	Z	P.unadj	P.adj
No training - Training	-2.87004231	0.004104169	0.012312506*
No_Training - Video_Rewatch	-0.05345225	0.957371576	0.957371576
Training - Video_Rewatch	2.80832097	0.004980056	0.007470085**

As sown in Table 7, analyses resulted in a significant difference between No training and Training conditions (p = .012*) and Training and Video Re-watch conditions (p = .007**). Thus, suggesting the effectiveness of the training in increasing or decreasing looking into the camera behaviours over other treatments. To check the direction of these significances, the data were plotted. The direction was positive, so it appears that the training significantly increased the behaviours of looking into the camera and hypothesis 1 was partially confirmed (Figure 2).

Figure 2: Look camera - Condition 1



5.5.1.2. Look away

To test hypothesis 1, the differential between the frequencies of looks away of the first round of questions and the second was considered as the dependent variable ($\Delta LookAway$). Condition 1 (Training, Video Re-watch, No Training) was considered as an independent variable. Kruskal-Wallis analysis of variance was then used. As sown in Table 8, analyses resulted in a significant difference between the three conditions Training, Video Re-watch and No training (Chi-square = 12.094, p = .002**, df = 2).

Table 8: Look away - Condition 1

Comparison	chi-squared	df	p-value

$\Delta LookAway \sim$ Condition 1	12.094	2	0.002365**

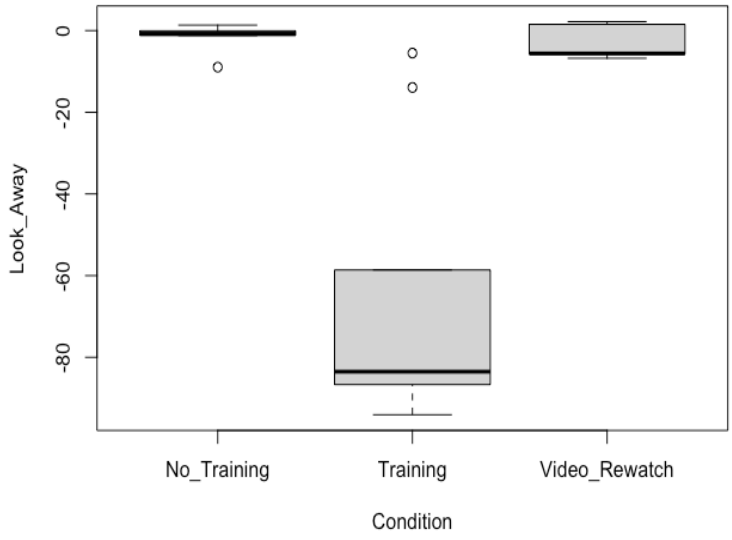
Dunn's posthoc test was then used to compare the differentials in the three pairwise conditions and check which were significantly different from each other. Before proceeding, the p-values were adjusted using the Benjamini-Hochberg method.

Table 9: Look away - Condition 1

Comparison	Z	P.unadj	P.adj
No training - Training	2.87004231	0.004104169	0.012312506*
No_Training - Video_Rewatch	0.05345225	0.957371576	0.957371576
Training - Video_Rewatch	-2.80832097	0.004980056	0.007470085**

As sown in Table 9, analyses resulted in a significant difference between No training and Training conditions (p = .012*) and Training and Video Re-watch conditions (p = .007**). Thus, suggesting the effectiveness of the training in increasing or decreasing looking away behaviours over other treatments. To check the direction of these significances, the data were plotted. The direction was negative, so it appears that the training significantly decreased the behaviours of looking away and hypothesis 1 was fully confirmed (Figure 3).

Figure 3: Look away - Condition 1



5.5.1.3. Nod

To test hypothesis 2, the differential between the frequencies of nods of the first round of questions and the second was considered as the dependent variable (ΔNod). Condition 1 (Training, Video Re-watch, No Training) was considered as an independent variable. Kruskal-Wallis analysis of variance was then used. As sown in Table 10, analyses resulted in a significant difference between the three conditions Training, Video Re-watch and No training (Chi-square = 9.4886, p = .008**, df = 2).

Table 10: Nod - Condition 1

Comparison	chi-squared	df	p-value
ΔNod ~ Condition 1	9.4886	2	0.008701**

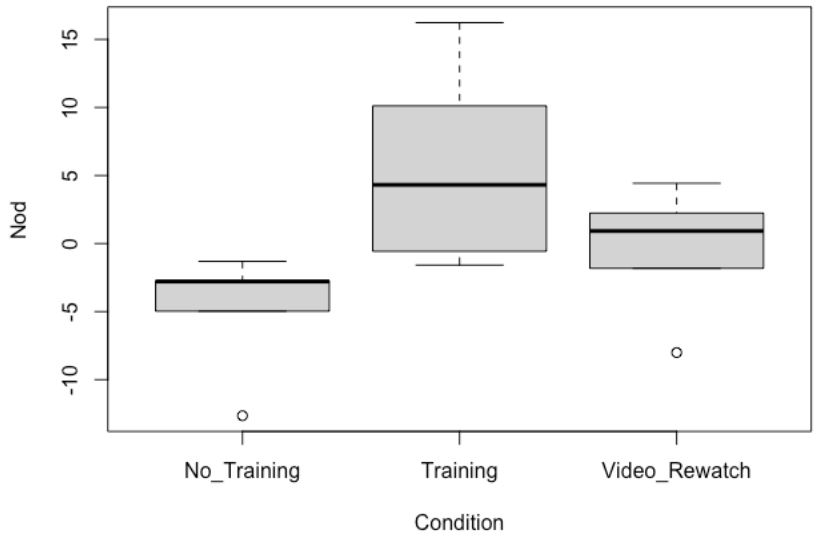
Dunn's posthoc test was then used to compare the differentials in the three pairwise conditions and check which were significantly different from each other. Before proceeding, the p-values were adjusted using the Benjamini-Hochberg method.

Table 11: Nod - Condition 1

Comparison	Z	P.unadj	P.adj
No training - Training	-3.055206	0.002249057	0.006747172**
No_Training - Video_Rewatch	-1.443211	0.148961124	0.223441687
Training - Video_Rewatch	1.388730	0.164914823	0.164914823

As sown in Table 11, analyses resulted in a significant difference between No training and Training conditions (p = .006**). No significant differences were found between Training and Video Re-watch conditions (p = .165) and between No training and Video Re-watch conditions (p = .223). Thus, suggesting the effectiveness of the training in increasing or decreasing nods behaviours over No training condition. To check the direction of these significances, the data were plotted. The direction was positive, so it appears that the training significantly increased the behaviours of nod I compared to No training condition and hypothesis 2 was partially confirmed (Figure 4).

Figure 4: Nod - Condition 1



5.5.1.4. Smile

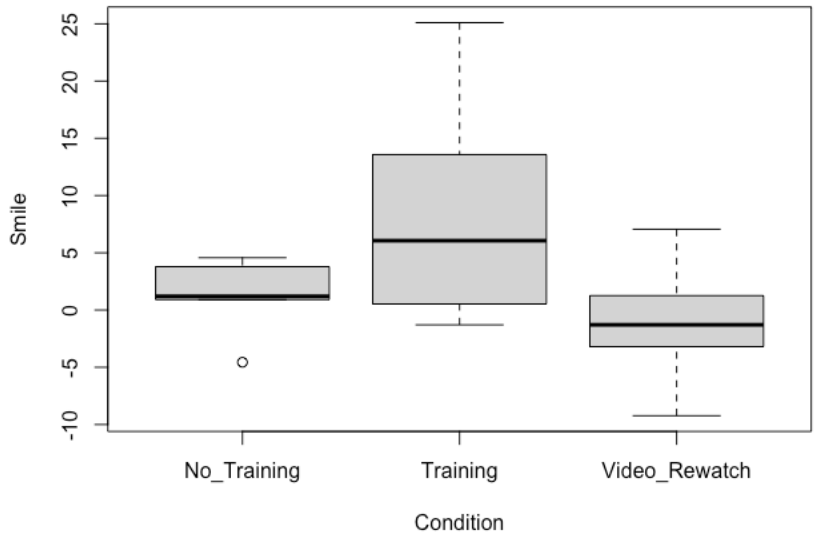
To test hypothesis 3, the differential between the frequencies of smiles of the first round of questions and the second was considered as the dependent variable ($\Delta Smile$). Condition 1 (Training, Video Re-watch, No Training) was considered as an independent variable. Kruskal-Wallis analysis of variance was then used. As sown in Table 12, analyses resulted in no significant difference between the three conditions Training, Video Re-watch and No training (Chi-square = 4.2739, p = .118, df = 2).

Table 12: Smile - Condition 1

Comparison	chi-squared	df	p-value
$\Delta Smile \sim$ Condition 1	4.2739	2	0.118

Therefore, in contrast to what was expected, it seems that there was no significant increment or decrement in smiling behaviours due to the training provided. So, it appears that the training didn't significantly increase or decrease the behaviours of smiling and hypothesis 3 was disconfirmed (Figure 5).

Figure 5: Smile - Condition 1



5.5.2. Interviewer vs. Interviewee

5.5.2.1. Look into the camera

To test hypothesis 4 looks into the camera of the first round of questions and the second were considered as the dependent variable (LookCamera). Condition 2 (Interviewer vs. Interviewee) was considered as an independent variable. Kruskal-Wallis analysis of variance was then used. As sown in Table 13, analyses resulted in no significant difference between the two conditions Interviewer and Interviewee (Chi-square = 0.18732, p = .665, df = 1). Therefore, as expected, it seems that there was no significant

difference in looking into the camera behaviours between the Interviewers and the Interviewee groups and hypothesis 4 was confirmed (Table 13).

Table 13: Look camera - Condition 2

Comparison	chi-squared	df	p-value
LookCamera ~ Condition 2	0.18732	1	0.6652

5.5.2.2. Look away

To test hypothesis 4, looks away of the first round of questions and the second was considered as the dependent variable (LookAway). Condition 2 (Interviewer vs. Interviewee) was considered as an independent variable. Kruskal-Wallis analysis of variance was then used. As sown in Table 14, analyses resulted in no significant difference between the two conditions Interviewer and Interviewee (Chi-square = 0.18732, p = .665, df = 1). Therefore, as expected, it seems that there was no significant difference in looking away behaviours between the Interviewers and the Interviewee groups and hypothesis 4 was confirmed (Table 14).

Table 14: Look away - Condition 2

Comparison	chi-squared	df	p-value
LookAway ~ Condition 2	0.18732	1	0.6652

5.5.2.3. Nod

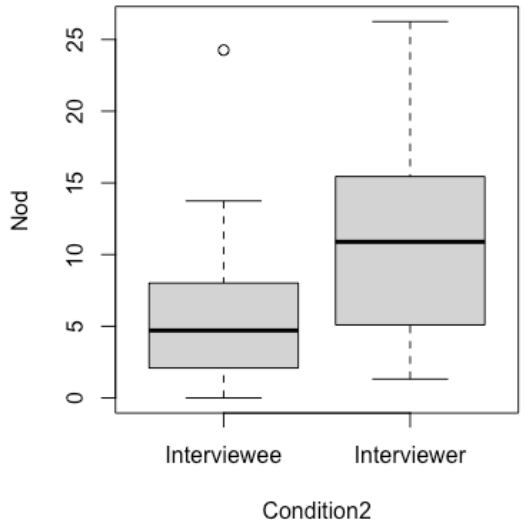
To test hypothesis 5, nods of the first round of questions and the second were considered as the dependent variable (Nod). Condition 2 (Interviewer vs. Interviewee) was considered as an independent variable. Kruskal-Wallis analysis of variance was then used. As sown in Table 15, analyses resulted in a significant difference between the two conditions Interviewer and Interviewee (Chi-square = 4.5679, p = .033*, df = 1).

Table 15: Nod ~ Condition 2

Comparison	chi-squared	df	p-value
Nod ~ Condition 2	4.5679	1	0.03258*

To check the direction of this significance, the data were plotted. The direction was positive for the condition Interviewer, so it appears that the frequencies of nods were significantly higher for the interviewers and hypothesis 5 was confirmed (Figure 6).

Figure 6: Nod ~ Condition 2



5.5.2.4. Smile

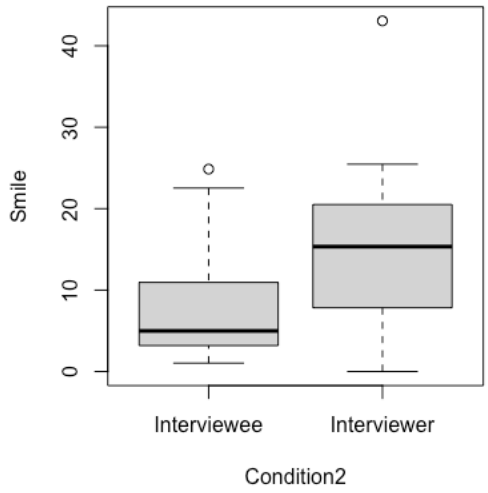
To test hypothesis 5, smiles of the first round of questions and the second were considered as the dependent variable (Nod). Condition 2 (Interviewer vs. Interviewee) was considered as an independent variable. Kruskal-Wallis analysis of variance was then used. As sown in Table 16, analyses resulted in a significant difference between the two conditions Interviewer and Interviewee (Chi-square = 7.3171, p = .007**, df = 1).

Table 16: Smile ~ Condition 2

Comparison	chi-squared	df	p-value
Differential ~ Condition 1	7.3171	1	0.00683**

To check the direction of this significance, the data were plotted. The direction was positive for the condition Interviewer, so it appears that the frequencies of smiles were significantly higher for the interviewers and hypothesis 5 was confirmed (Figure 7).

Figure 7: Smile ~ Condition 2



5.6. Discussion

In this study, therefore, training based on a previous study was constructed, administered and evaluated to improve behaviours usually linked to engagement during interviews. The behaviours focus of this training were Look into the camera, Look away, Nodding and Smiling. To do this, differentials were calculated between the frequency of behaviours during the second round of questions and those expressed during the first round of questions ($\Delta Behav = BehavFreqSecond\% - BehavFreqFirst\%$). This is because it was expected that each participant would have shown a different degree of behaviours and therefore the training would have a different effect on each of them.

Therefore, a statistically significant difference was found with respect to the frequencies of gaze behaviour in the three experimental conditions. Look into the camera appears to have increased consistently in the training condition during the second round of questions (NT-T: p = .012*, T-VRW: p = .007**; Table 7), whereas Look away appears to have decreased in the same way in the training condition (NT-T: p = .012*, T-VRW: p = .007**; Table 9). This, therefore, seems to suggest the effectiveness of the proposed training in increasing looking into the direction of the camera behaviours and decreasing looking away behaviours. As seen above, these behaviours have often been linked to the concept of engagement during a conversation, with people being more engaged when looking towards their interlocutor and less engaged when looking away (Mehrabian, 1968b; Richmond et al., 1987; Szafir & Mutlu, 2012). Nevertheless, in the case of a videoconferencing conversation, looking at the interlocutor can be equated with looking at the camera, as far as the perception of the interlocutor is concerned, and looking in another direction, even if directed towards some part of the screen, can be perceived as looking elsewhere (Grondin et al., 2020).

In addition, it was found that the proposed training was effective in increasing nodding behaviour, compared to the condition in which this training was not provided (NT-T: p = .006**). The same cannot be said, however, with respect to the condition in which the participants were shown the video of their first round of questions (p = .165). This could be due to the fact that just seeing oneself during an interview may suggest the need to increase nodding behaviour. These results therefore partially confirmed hypothesis 2. As seen in the previous study, this behaviour was found to be central to the definition of both immediacy (Andersen, 1979; Richmond et al., 1987) and engagement (Nakano & Ishii, 2010). Improving this aspect, therefore, is expected to mean improving engagement between the two interlocutors as well.

With regard to smiles, however, no increase or decrease due to the experimental condition was found. This seems to disconfirm hypothesis 3. As far as I know, this result could be due to the fact that the training was not effective in increasing this behaviour or that the improvement was not so evident as to be statistically significant.

Furthermore, as hypothesised (hypothesis 4), the gaze of a person engaged in a job video interview does not seem to be influenced by their role. In fact, no relevant difference was found between the gazes of Interviewers and Interviewees (Look into the camera: $p = .665$, Look away: $p = .665$).

With regard to nodding and smiling behaviour, on the contrary, it seems that these are influenced by the role. Apparently, it is the Interviewers who nodded and smiled the most during the video interview (Nod: $p = .033$*, Smile: $p = .007$**) and hypothesis 5 was confirmed. This result, in my opinion, can be explained by the fact that the interviewer in this type of interview is often more silent than the interviewee and therefore has to give much more feedback to the interlocutor to let him/her know that he/she is listening and therefore nods more. Also, the fact that the interviewee speaks more I think may give the interviewee less chance to smile but also give the interviewer more chance to react with smiles to what the interlocutor is saying.

At this point, I asked myself whether the cues identified in the preliminary study and then implemented in the training administrated via Zoom actually had an effect on perceived engagement, i.e. on the interrelational dimension I wanted to improve. To this end, the following study was carried out.

6. Third study: Engagement during video-interviews evaluation

6.1. Objective

The aim of this study was to investigate whether the training administered in the second study, in addition to increasing behaviours indicative of engagement, increased perceived engagement itself. To this end, an independent commission of evaluators was recruited online and asked to assess the videos collected during the interviews in the previous study.

6.2. Materials

6.2.1. Video preparation

The videos collected so far were kept in their entirety except in cases where there were additional questions from the participant. In this case, they were cut in such a way as to include only the agreed questions and answers. The videos were also edited to have the participant's face as the main, central image and the confederate in a box at the top left of the video. This decision was made to make it easier for the evaluators to focus on the trainees. The duration of each video ranged from 3:45 minutes up to 8:18 minutes.

6.2.2. Online Survey

Participants were shown both videos of an interview participant in random order and for each video they were asked to indicate their level of agreement/disagreement (from Strongly disagree to Strongly agree) with the following statements:

- I think the person was enjoying the conversation.
- I think the person was satisfied with the conversation.
- I think the person was involved in the conversation.
- I think the person was paying attention to his/her interlocutor.
- I think the person seemed well disposed.
- I think the person's manner was engaging.

6.3. Participants

For this experiment, a total of 197 responses were collected. 54 of these responses were eliminated because the responses were repetitive and were therefore suspected to be given by random chance. A further 13 responses were randomly eliminated to balance the sample by gender. Thus, the final sample consisted of 130 (67 M – 63 F), English native speakers, recruited online using the Prolific platform.

Table 17: Video evaluation sample

	Male	Female	Total
Training	22	20	42
Video Rewatch	25	24	49
No Training	20	19	39
Total	67	63	130

6.4. Procedure

The collected videos were included in a questionnaire created through the Qualtrics platform and administered through Prolific. Participants were shown both videos of one of the interviews' participants (first and second round of questions), in random order, and asked to rate the level of engagement of the participants for each video. The commission had to assess the couple's mutual engagement, through a general assessment of the progress of the interview and some evaluations of the specific performance of the participant (the person shown large in the videos). Each of the participants, therefore, saw both interviews of only one interview participant and therefore assessed (unknowingly) the interviews as part of one experimental condition.

In this study, therefore, it was expected that:

- The differential of engagement between the second round of questions and the first one would have been significantly higher for people who had performed our training when compared to the Video Re-watch and No training conditions (Hypothesis 1).
- Engagement scores would not have been significantly different according to the role of participants, i.e., Interviewer and Interviewee condition (Hypothesis 2).
- Engagement scores would not have been significantly different according to the gender of respondents or protagonists or the combination of the gender of respondents and protagonists. (Hypothesis 3).

6.5. Results

6.5.1. Training vs. Video rewatch vs. No training

To test hypothesis 1, the differential between the engagement scores of the first round of questions and the second was considered as the dependent variable ($\Delta Engagement$). Condition 1 (Training, Video Re-watch, No Training) was considered as an independent variable. Kruskal-Wallis analysis of variance was then used. As sown in Table 18, analyses resulted in a significant difference between the three conditions Training, Video Re-watch and No training (Chi-square = 12.536, p = .002**, df = 2).

Table 12: Engagement - Condition 1

Comparison	chi-squared	df	p-value
$\Delta Engagement$ ~ Condition 1	12.536	2	0.001896**

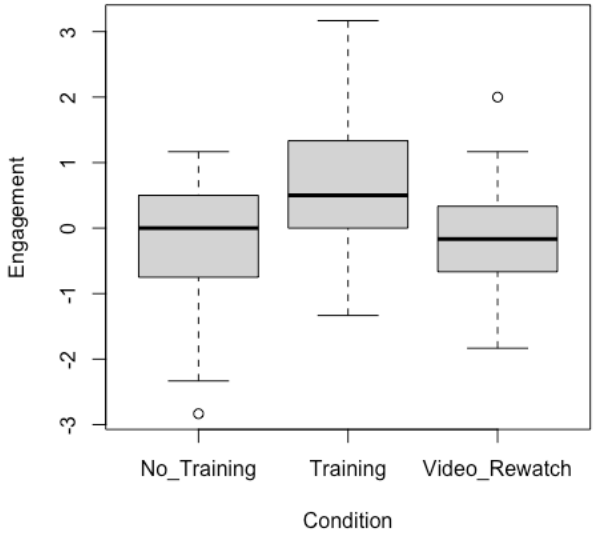
Dunn's posthoc test was then used to compare the differentials in the three pairwise conditions and check which were significantly different from each other. Before proceeding, the p-values were adjusted using the Benjamini-Hochberg method.

Table 13:Engagement - Condition 1

Comparison	Z	P.unadj	P.adj
No training - Training	-2.8582608	0.004259702	0.006389553**
No_Training - Video_Rewatch	0.2264709	0.820835203	0.820835203
Training - Video_Rewatch	3.2537742	0.001138828	0.003416483**

As sown in Table 19, analyses resulted in a significant difference between No training and Training conditions (p = .006**) and Training and Video Re-watch conditions (p = .003**). Thus, suggesting the effectiveness of the training in increasing or decreasing engagement between interlocutors over other conditions. To check the direction of these

Figure 8: Engagement – Condition 1



significances, the data were plotted. The direction was positive, so it appears that the training significantly increased engagement and hypothesis 1 was fully confirmed (Figure 8).

6.5.2. Interviewer vs. Interviewee

To test hypothesis 2, all the engagement scores (i.e., the first round of questions and the second) were considered as the dependent variable (Engagement). Condition 2 (Interviewer vs. Interviewee) was considered as an independent variable. Kruskal-Wallis analysis of variance was then used. As sown in Table 19, analyses resulted in no

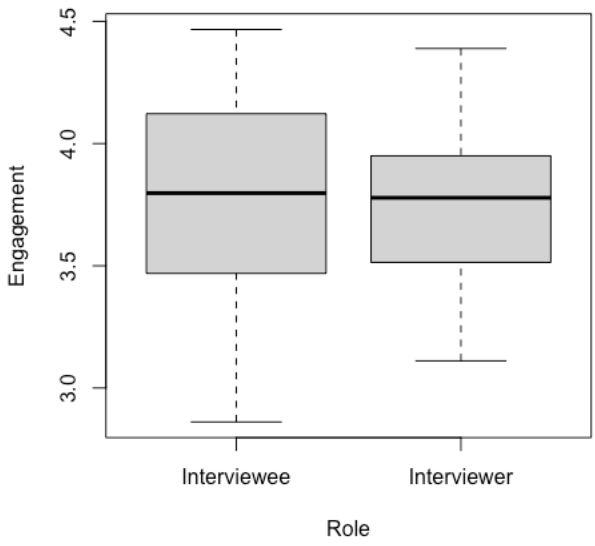
significant difference between the conditions Interviewer and Interviewee (Chi-square = 0.096805, p = .756, df = 1).

Table 19: Engagement - Condition 2

Comparison	chi-squared	df	p-value
Engagement ~ Condition 2	0.096805	1	0.7557

Therefore, it appears that the role played by participants did not significantly impact perceived engagement and hypothesis 2 was fully confirmed (Figure 9).

Figure 9: Engagement – Condition 2



6.5.3.Gender

6.5.3.1. Evaluators' gender

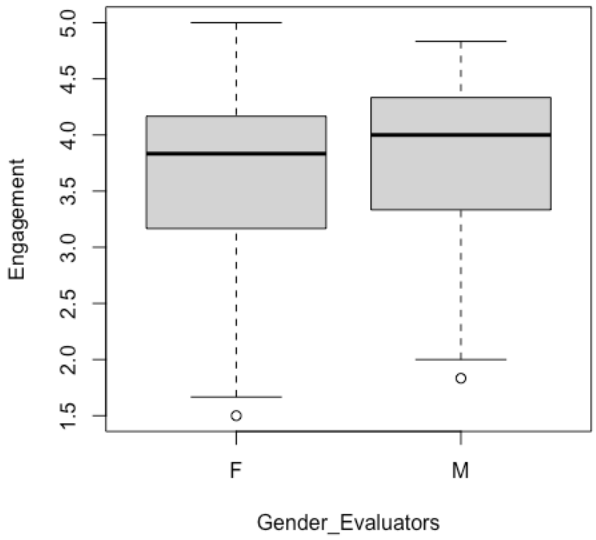
In order to test hypothesis 3, all the engagement scores (i.e., the first round of questions and the second) were considered as the dependent variable (Engagement). Evaluators' gender (Male vs. Female) was considered as an independent variable. Kruskal-Wallis analysis of variance was then used. As sown in Table 20, analyses resulted in a slightly significant difference between the conditions Male and Female (Chi-square = 3.9278, p = .048*, df = 1).

Table 20: Engagement – Evaluators' gender

Comparison	chi-squared	df	p-value
Engagement~ EvaluatorsGender	3.9278	1	0.0475*

Thus, suggesting the effectiveness of the evaluators' gender on the evaluations themselves. To check the direction of this significance, the data were plotted. The direction was positive for the Male condition, so it seems that males gave slightly higher ratings to the engagement of the protagonists and hypothesis 3 was partially disconfirmed (Figure 10).

Figure 10: Engagement – Evaluators' gender



6.5.3.2. Protagonists' gender

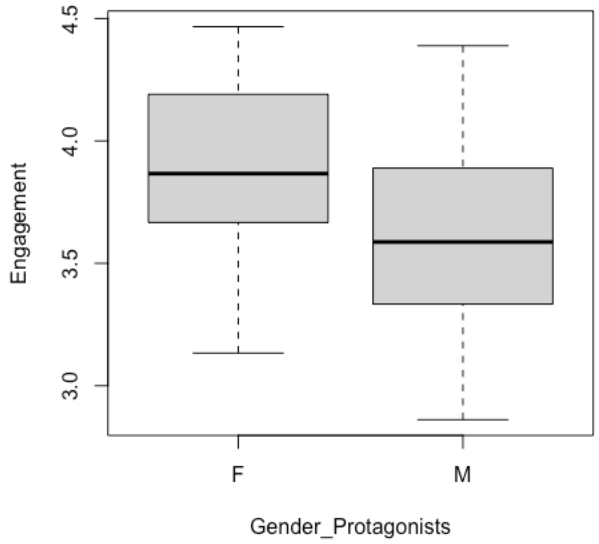
In order to test hypothesis 3, all the engagement scores (i.e., the first round of questions and the second) were considered as the dependent variable (Engagement). Protagonists' gender (Male vs. Female) was considered as an independent variable. Kruskal-Wallis analysis of variance was then used. As sown in Table 21, analyses resulted in a slightly significant difference between the conditions Male and Female (Chi-square = 3.9401, p = .047*, df = 1).

Table 21: Engagement - Protagonists' gender

Comparison	chi-squared	df	p-value
Engagementt ~ ProtagonistsGender	3.9401	1	0.04715*

Thus, suggesting an effect of the protagonists' gender on the evaluations given by the evaluators. To check the direction of this significance, the data were plotted. The direction was positive for the Female condition, so it seems that females were rated as slightly more engaged than males and hypothesis 3 was partially disconfirmed (Figure 11).

Figure 11: Engagement – Protagonists' gender



6.5.3.3. Gender combination

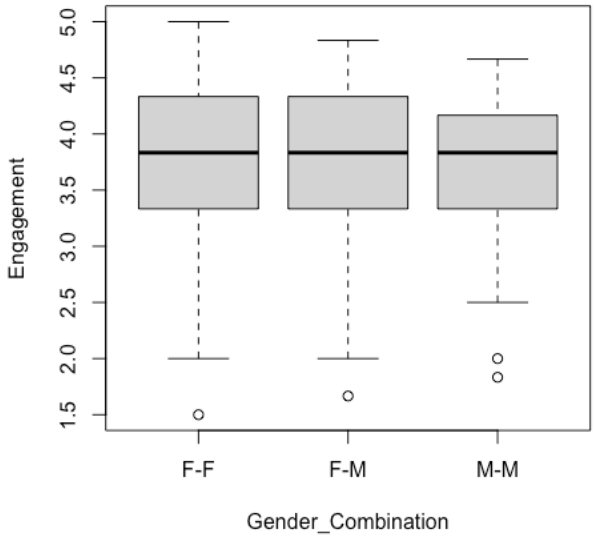
In order to test hypothesis 3, all the engagement scores (i.e., the first round of questions and the second) were considered as the dependent variable (Engagement). The combination of evaluators and protagonists' gender (Female-Female vs. Female-Male vs. Male-Male) was considered as an independent variable. Kruskal-Wallis analysis of variance was then used. As sown in Table 22, analyses resulted in a slightly significant difference between the conditions Male and Female (Chi-square = 1.7534, p = .4162, df = 2).

Table 22: Engagement – Gender combination

Comparison	chi-squared	df	p-value
Engagement ~ GenderCombination	1.7534	2	0.4162

Therefore, it appears that the gender combination of evaluators and protagonists did not significantly impact engagement evaluations and hypothesis 3 was partially confirmed (Figure 12).

Figure 12: Engagement – Gender combination



6.6. Discussion

In this study, the interviews collected in the previous experiment were evaluated by an independent commission of 130 participants. This evaluation was related to the levels of

engagement of interview participants perceived by people external to the interview itself. This evaluation was central to understanding whether the constructed training, in addition to increasing the frequency of the supposedly increased behaviour (effectiveness), was also able to increase the engagement between the two interlocutors (efficacy). To do this, differentials were calculated between the engagement scores given by the evaluators to the second round of questions and those given to the video representing the first round of questions ($\Delta Engagement$). This is because it was expected that each participant would start from a different level of engagement skills and therefore the training would have a different effect on each of them.

Therefore, a statistically significant difference was found with respect to engagement in the three experimental conditions (NT-T: p = .006**, T-VRW: p = .003**, Table 19). Also, it appears that this difference was positive for what concerns the Training condition (Figure 8). It therefore appears that the training constructed and administered in the previous study was successful in increasing engagement between interactants. These results therefore fully confirmed hypothesis 1.

Furthermore, it seems that the role played by the participants did not have an effect on the perceived engagement by the external evaluators, as expected (p = .756). These results therefore fully confirmed hypothesis 2.

With regard to gender, however, there was a slight difference in engagement scores due to both the gender of the external commission and the gender of the participants. It seems that males tended to give higher engagement ratings and that females were rated as more involved in the interview (commission: p = .048*, participants: p = .047*). However, when taken together, this effect on engagement ratings seems to fade (p = .4162). Hypothesis 3 is therefore partially confirmed and partially disconfirmed. These

differences could be due to the size of the sample itself. In one case it was 130 evaluators, in the other 20 participants.

Given the observed effectiveness of administered training in increasing engagement, and given the increase in some behaviours over others in the training condition, the question arose as to whether behavioural cues could explain the engagement scores collected. For this purpose, the following study was designed.

7. Fourth study: Engagement and Behavioural cues correlation

7.1. Objective

The aim of this study was to check whether the behavioural cues annotated during the second part of the second study could explain the engagement values rated by the 130 external evaluators during the third study. To this end, the mean engagement scores found in the previous study were compared with the four behavioural cues annotated in the second study (i.e., Look into the camera, Look away, Nod, Smile).

7.2. Materials

The materials used were those of studies 2 and 3 and the results obtained from them. With regard to annotated behaviours, forty (40) videos were collected (duration between 3:45 minutes up to 8:18 minutes) and eighty (80) annotations were made on these videos. For each protagonist, the frequency of each of the four behaviours was considered. With regard to engagement evaluations, 130 evaluations were collected during the second study. From these evaluations, the mean score given to each participant was calculated.

7.3. Procedure

The mean engagement scores given by the 130 evaluators of the third study were calculated for each of the protagonists of the video interview from the second study. In addition, the frequencies of each of the four behavioural cues annotated on these videos (Look into the camera, Look Away, Nod and Smile) were considered.

In this study, therefore, it was expected that:

- The frequency of Looks into the camera would have correlated with previously assessed engagement scores. A positive correlation was expected, i.e., the higher the engagement the more frequent the looks into the camera (Hypothesis 1).

- Consistently, I hypothesized that the frequency of Looks away would have correlated with previously assessed engagement scores. A negative correlation was expected, i.e., the higher the engagement the less frequent the looks away (Hypothesis 2).
- In addition, I expected that the frequency of Nods and Smiles would have correlated with previously assessed engagement scores. A positive correlation was expected, i.e., the higher the engagement the more frequent the nods and smiles (Hypothesis 3).

7.4. Results

To carry out these analyses, the mean engagement scores were calculated for each video interview evaluated (tot. 40 videos). In addition, the frequencies of each of the four behavioural cues annotated on these videos (Look into the camera, Look Away, Nod and Smile) were calculated and expressed as the total number of frames in which the behaviour occurred, divided by the total number of frames in the video ($BehavFreq = \frac{TotBehavFrames}{TotFrames}$). This was done to avoid the length of the video having an effect on the relevance of the behaviour itself. Behaviours' frequencies were then expressed as percentages.

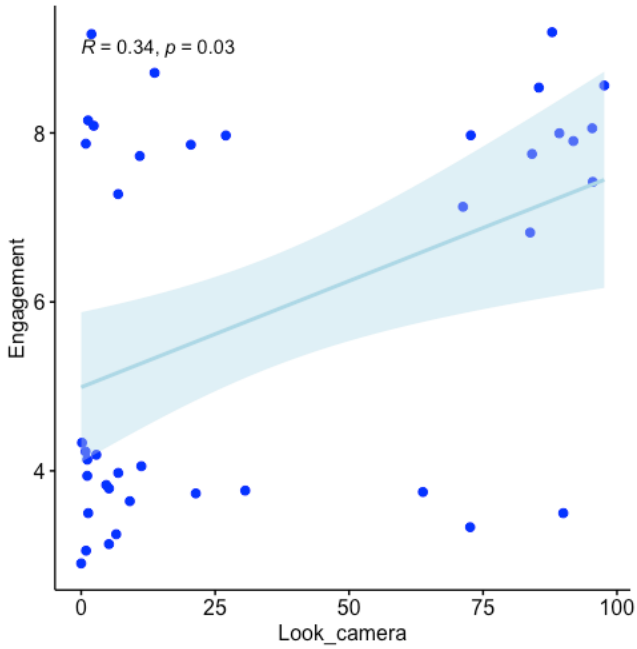
7.4.1.1. Look into the camera

To test hypothesis 1, the correlation between engagement scores and the frequency of Look into the camera was investigated using Spearman's rank correlation coefficient. As shown in Table 23, analyses resulted in a significant correlation between engagement scores and the behaviour Look into the camera (S = 6992.8, p = .0297*, R = .34). Also, the correlation resulted to be positive thus indicating that the higher the engagement the more frequent the looks into the camera were (R = .34). Hypothesis 1 was therefore confirmed (Figure 13).

Table 23: Engagement - LookCamera

Comparison	S	rho	p-value
Engagement ~ LookCamera	6992.8	0.3440124	0.02974*

Figure 13: Engagement - LookCamera



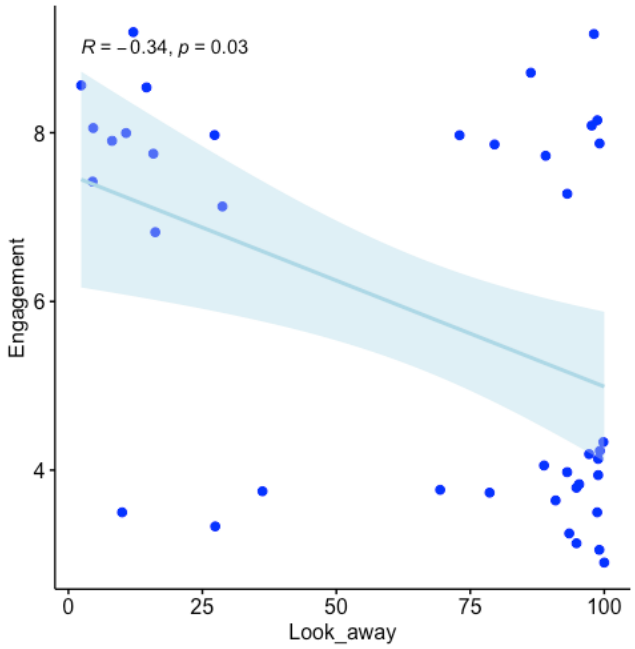
7.4.1.2. Look away

In order to test hypothesis 2, the correlation between engagement scores and the frequency of Look away was investigated using Spearman's rank correlation coefficient. As shown in Table 24, analyses resulted in a significant correlation between engagement scores and the behaviour Look away (S = 6992.8, p = .0297*, R = - .34). Also, the correlation resulted to be negative thus indicating that the higher the engagement the less frequent the looks away (R = - .34). Hypothesis 2 was therefore confirmed (Figure 14).

Table 24: Engagement - LookAway

Comparison	S	rho	p-value
Engagement ~ LookCamera	14327	-0.3440124	0.02974*

Figure 14: Engagement - LookAway



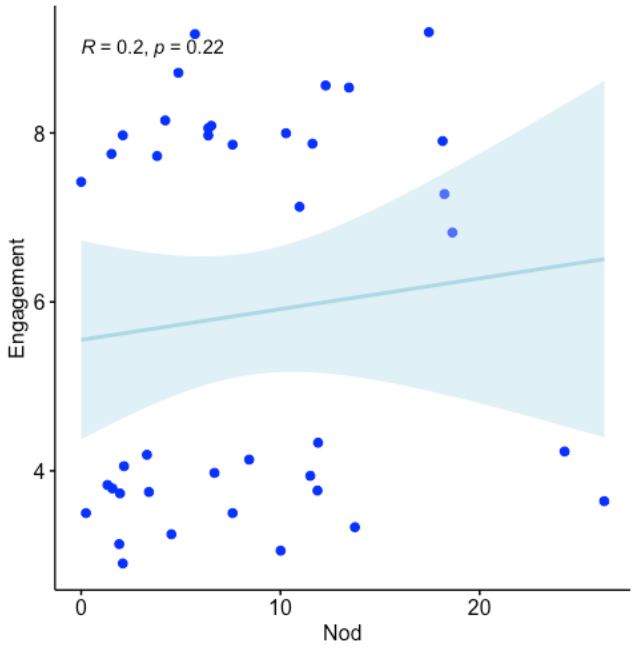
7.4.1.3. Nod

In order to test hypothesis 3, the correlation between engagement scores and the frequency of Nods was investigated using Spearman's rank correlation coefficient. As shown in Table 25, analyses resulted in no significant correlation between engagement scores and the behaviour Nodding (S = 8549.1, p = .221, R = .20). Hypothesis 3 was therefore disconfirmed (Figure 15).

Table 24: Engagement - Nod

Comparison	S	rho	p-value
Engagement ~ LookCamera	8549.1	0.1980203	0.2206

Figure 15: Engagement - Nod



7.4.1.4. Smile

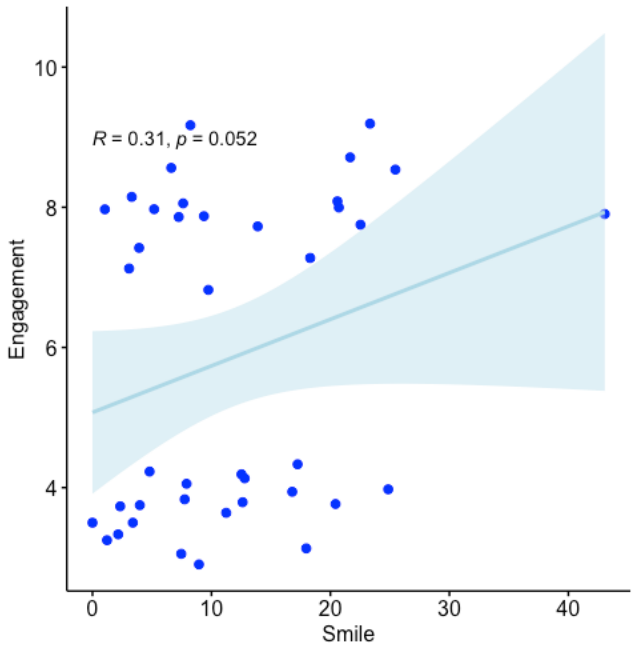
To test hypothesis 3, the correlation between engagement scores and the frequency of Smiles was investigated using Spearman's rank correlation coefficient. As shown in Table 26, analyses resulted in no significant correlation between engagement scores and the

behaviour of Smiling (S = 7366.8, p = .052, R = .31). Nevertheless, a positive tendency can be noted (Figure 16). In this respect, further investigations were carried out.

Table 24: Engagement - Nod

Comparison	S	rho	p-value
Engagement ~ LookCamera	7366.8	0.3089263	0.05243

Figure 16: Engagement - Smiles



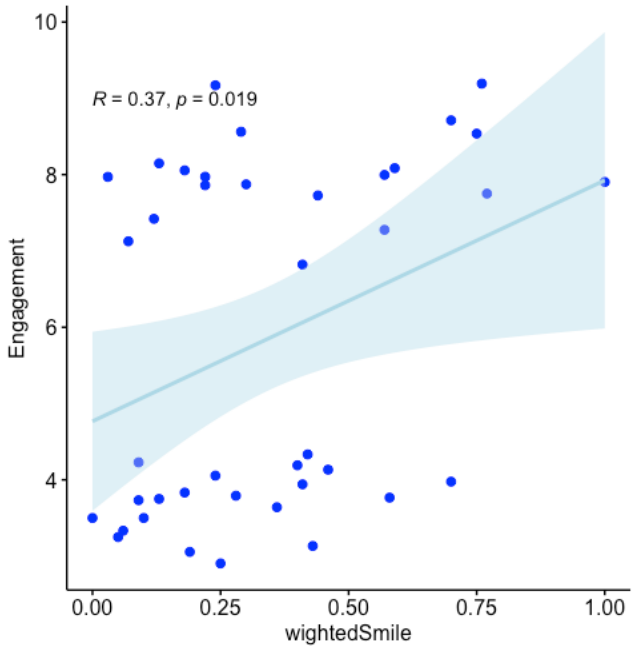
7.4.1.5. Weighted smile

The correlation between engagement scores and the frequency of Smiles, weighted on the maximum frequency of smiles shown, was investigated using Spearman's rank correlation coefficient. As shown in Table 25, analyses resulted in a significant correlation between engagement scores and the behaviour (weighted) Smiling (S = 6730.3, p = .019*, R = .37). Also, the correlation resulted to be positive thus indicating that the higher the engagement the more frequent smiles were (R = .37). Hypothesis 3 was therefore partially confirmed (Figure 17).

Table 25: Engagement - weighted Smile

Comparison	S	rho	p-value
Engagement ~ weightedSmile	6730.3	0.3686359	0.01925*

Figure 17: Engagement - weightedSmiles



7.5. Discussion

In this study, I wanted to test whether the behavioural cues annotated during the previous studies could explain the engagement values assessed by the 130 external evaluators during the third study. To this end, the engagement scores given by the external evaluators were compared with the frequency of behaviours annotated.

With regard to gaze behaviours, it was found that engagement and gaze are closely related (p = .0297*). In fact, it seems that higher engagement scores were associated with more frequent looking into the camera behaviours (R = .34). On the contrary, it seems that higher engagement scores were associated when looking away behaviours were fewer (R = - .34). Hypotheses 1 and 2 of this study were thus confirmed, endorsing the theories that gaze is related to involvement and immediacy.

With regard to nodding behaviour, however, no correlation was found between this behaviour and the engagement scores expressed ($p = .221$). This might be due to the fact that these behaviours were not taken into account by the evaluators or were not always taken into account. The reason for this could be that some of these behaviours are not easily recognisable. Sometimes during the annotation process, the videos had to be reviewed many times to annotate them thoroughly. Hypothesis 3 was therefore rejected. With regard to smiling behaviour, no strong correlation was found, but rather a trend between engagement scores expressed by the 130 evaluators and this behaviour ($p = .052$). Further analysis was therefore carried out, given this trend. The frequency of these behaviours was then weighted by the highest number of smiles found (43.06% of the total frames) so that this behaviour was considered on its own and not in relation to the other three behaviours. This subsequent analysis led to the identification of a correlation between the scores expressed by the evaluators and the smiling behaviour ($p = .019^*$), suggesting that this behaviour had been taken into account in assigning the degree of engagement to the participants.

8. Fifth study: automatic detection of behavioural cues

8.1. Objective

The aim of this study was to build a model capable of extracting and predicting the nonverbal cues found and tested during the previous studies. This, in order to be able to automatically extract cues that, as seen above, have been identified qualitatively. In particular, Gaze (looking at the camera and looking away), Smiling and Nodding were considered. Automatic extraction guarantees that this process is carried out much faster. The aim of this study, therefore, was to build a reliable model to extract these cues automatically using state-of-the-art machine learning algorithms, based on our specific dataset. This model was built on the basis of the videos collected in the second study, manually annotated by me.

8.2. Materials

The videos used for the annotations were 40 videos collected in the previous experiment of 20 participants engaged in job interviews. These videos were considered without audio, hence mutated, as they had been used for the engagement assessment carried out online in the previous study. The videos were all in mp4 format and ranged from 3:45 to 8:18 minutes, from 54 MB to 327 MB, from 5648 frames to 12'450 frames. The total number of annotated frames was 482'927. The annotations were a total of 80 files in .txt format containing, for each behaviour, the start frame, the end frame, and the corresponding label. The images were 347'861 (one per frame) 160x160 pixels black and white.

8.3. Procedure

In order to test which system could predict the behaviours identified in the first study, the 40 videos collected in the second study were initially annotated. These videos were analysed frame by frame and annotated following the cues identified in the first study and selected to create the training administered in the second study. The specific annotations were " look at the camera", "look away", "smile", and "nod". Any time the participant looked in the direction of the webcam, it was annotated as "looking into the camera." Any time the participant looked in a direction other than the camera, it was annotated as "look away". The label "smile" was associated whenever the participant smiled clearly (i.e., the corners of the mouth began to turn upward until he or she assumed a neutral expression again). The label "nod" was associated whenever the participant made assenting movements of the head, tilting it from top to bottom, from the first head movement to the last consecutive one. Initially, the videos were annotated using the 4 labels simultaneously. This, however, generated an overlap of the gaze labels and those of the other two behaviours "smile" and "nod". In fact, participants, throughout the interview, either looked directly into the camera or looked away and, at the same time, performed actions such as smiling or nodding. For this reason, it was decided to split the annotations into two separate datasets: a dataset regarding annotations on Gaze, containing two classes " look into the camera" and "look away", and a dataset regarding information on Actions, containing "smile" and "nod". Thus, two separate annotations were generated for each of the videos, for a total of 80 annotations.

During the pre-processing stage, the videos of the dataset were converted into frames. Each video was recorded at 25 fps so every frame of each video could be extracted. These frames were fed in Facenet (Schroff & Philbin, 2015), a system able to detect and crop

faces given an image. The network outputs images were 160x160 pixels and only the first prediction per input image was kept if more than one face was depicted. These images were converted in grayscale and saved to be used during the training. The total training set was split into 2 different datasets Gaze, Action. The following table is an overview of the classes, their constituent datasets and their size (Table 26).

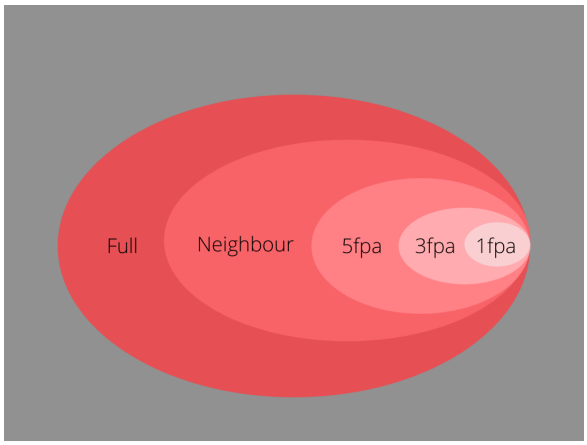
Table 26: Datasets

Dataset	Frames	Labels	
Gaze	347861	Look_camera	123869
		Look_away	223992
Action	67980	Nod	28555
		Smile	39425

In this study, different approaches were considered in order to understand which was the most efficient in predicting the behaviour of interest. The first approach (Approach 1) consisted in using the video frames of 5 random interviews as a test set and the frames of the remaining 15 interviews as a training set. The second approach (Approach 2) consisted in using the frames of only one interview (two videos) as a test set and then using the frames of the remaining 19 interviews as a training set. Finally, the third approach (Approach 3) consisted in using a reduced number of frames for each behaviour, as the most descriptive frames of a given behaviour might not occur throughout the entire annotation. In fact, videos were recorded at 25fps, so there might be multiple frames that were very similar and there might be no reason in counting them in the dataset (Figure 18):

- **1fpa** (frame per annotation): in this subset, only the middle frame of each annotation was kept;
- **3fpa**: in this subset, the middle frame of each annotation and the frame at 25% and 75% of the annotation time period were kept;
- **5fpa**: in this subset, the middle, the start, the end frame of each annotation, and the frames at 25% and 75% of the annotation time period were kept;
- **Neighbour**: this subset contains 20% of the total frames per annotation located around the middle frame;
- **Full**: consisted of the total number of frames.

Figure 18: Frame subsets



8.4. Results

The network used in all the experiments for Facial Expression Recognition was DAN (Wen et al., 2021). For Approach 1, 10 sets were therefore created. The following table shows the accuracy of the network used in predicting Gaze behaviour, from easiest to hardest to predict (Table 27). It appears that set 4 was the easiest to predict and set 10 the

most difficult. Also, example 2 seems it was difficult to predict, as it was present in all bad-performing test sets. The cells highlighted in red were known unique errors that were corrected later. The average accuracy of this network in the Gaze dataset applying Approach 1 was found to be .63.

Table 27: Approach 1 - Gaze

	test set					Accuracy	train_size	test_size	total
set4	3	7	7	16	18	0.7715	279122	68739	347861
set8	6	8	11	19	20	0.7221	259568	88293	347861
set6	4	6	14	17	18	0.7206	255290	92571	347861
set5	9	13	15	16	17	0.7096	265580	82281	347861
set1	1	4	6	9	14	0.6664	257917	89944	347861
set7	3	9	11	12	13	0.5788	267143	80718	347861
set3	1	2	5	12	17	0.5786	266328	81533	347861
set2	5	6	6	7	12	0.5499	286879	60982	347861
set9	2	7	8	13	13	0.4924	273492	74369	347861
set10	2	9	11	17	20	0.4878	259783	88078	347861
Average						0.62777	267110	80751	

The following table shows the accuracy of the network used in predicting Action behaviours, from easiest to hardest to predict (Table 28). It appears that set 3 was the easiest to predict and set 8 the most difficult. Also, examples 9, 11 and 20 seem they were

difficult to predict, as they were present in all bad-performing test sets. The cells highlighted in red were known unique errors that were corrected later. The average accuracy of this network in the Action dataset applying Approach 1 was found to be .82.

Table 28: Approach 1 - Action

	test set					**Acc**	**train_size**	**test_size**	**total**
set3	1	2	5	12	17	0.8688	52296	15684	67980
set6	4	6	14	17	18	0.8503	54036	13944	67980
set9	2	7	8	13	13	0.828	51046	16934	67980
set2	5	6	6	7	12	0.827	53236	14744	67980
set7	3	9	11	12	13	0.824	50009	17971	67980
set5	9	13	15	16	17	0.8237	48227	19753	67980
set4	3	7	7	16	18	0.8163	57284	10696	67980
set1	1	4	6	9	14	0.8093	54858	13122	67980
set10	2	9	11	17	20	0.7886	47174	20806	67980
set8	6	8	11	19	20	0.7653	44414	23566	67980
Average						0.82013	51258	16722	

For Approach 2, 20 sets were therefore created. The following table shows the accuracy of the network used in predicting Gaze behaviour, from easiest to hardest to predict (Table 29). It appears that set 16 was the easiest to predict and set 6 the most difficult. Moreover, as assumed above, example 2 was one of the hardest to predict. The average accuracy of this network in the Gaze dataset applying Approach 2 was found to be .71. Thus, there

was an increase in average accuracy of approximately 8% using Approach 2 compared to Approach 1 in the same dataset.

Table 29: Approach 2 - Gaze

	Acc	train_size	test_size	total
set16	0.9896	333210	14651	347861
set18	0.98	327221	20640	347861
set11	0.9508	329422	18439	347861
set17	0.9444	332627	15234	347861
set1	0.9121	330576	17285	347861
set20	0.894	330533	17328	347861
set19	0.8866	329803	18058	347861
set4	0.8722	328149	19712	347861
set15	0.7594	327668	20193	347861
set12	0.7426	333419	14442	347861
set14	0.685	326145	21716	347861
set10	0.6743	332389	15472	347861
set7	0.6429	330047	17814	347861
set3	0.5425	332227	15634	347861
set5	0.5404	334404	13457	347861
set8	0.5061	328662	19199	347861

set13	0.449	331620	16241	347861
set9	0.437	331899	15962	347861
set2	0.4351	326746	21115	347861
set6	0.3986	332592	15269	347861
Average	0.71213	330468	17393	

The following table shows the accuracy of the network used in predicting Action behaviour, from easiest to hardest to predict (Table 30). It appears that set 2 was the easiest to predict and set 4 the most difficult. Moreover, as assumed above, examples 9, 11 and 20 were among the hardest to predict. The average accuracy of this network in the Action dataset applying Approach 2 was found to be .85. Thus, there was an increase in average accuracy of approximately 3% using Approach 2 compared to Approach 1 in the same dataset.

Table 30: Approach 2 - Action

	Acc	**train_size**	**test_size**	**total**
set2	0.9618	64757	3223	67980
set1	0.9294	67328	652	67980
set14	0.9257	66391	1589	67980
set3	0.8939	66368	1612	67980
set12	0.8934	65804	2176	67980
set6	0.8915	64072	3908	67980

set19	0.8636	61296	6684	67980
set16	0.8609	64450	3530	67980
set8	0.8503	63323	4657	67980
set10	0.8481	65465	2515	67980
set7	0.8373	64740	3240	67980
set18	0.8332	65666	2314	67980
set17	0.831	63767	4213	67980
set13	0.8287	62166	5814	67980
set5	0.8183	62560	5420	67980
set11	0.8182	64664	3316	67980
set9	0.8169	62927	5053	67980
set15	0.8101	66837	1143	67980
set20	0.801	62979	5001	67980
set4	0.7208	66060	1920	67980
Average	0.851705	64581	3399	

Subsequently, based on the findings of Approach 2, four subsets of five examples were created based on the difficulty of the network to predict behaviours: Easy, Normal, Intermediate and Hard. A methodology applied during approach 1 was used with the generated subsets. The 5 sets of each subset were used as test sets and the remaining 15 as training sets. The results show an average accuracy of .67 for the Gaze dataset and .84

for the Action dataset (Tables 31-32). Thus, there was an increase in average accuracy of approximately 4% using Approach 2 in the difficulty subset of Gaze compared to the Gaze dataset. There also was an increase in average accuracy of approximately 2% using Approach 2 in the difficulty subset of Action compared to the Action dataset.

Table 31: Approach 2 – Gaze subsets

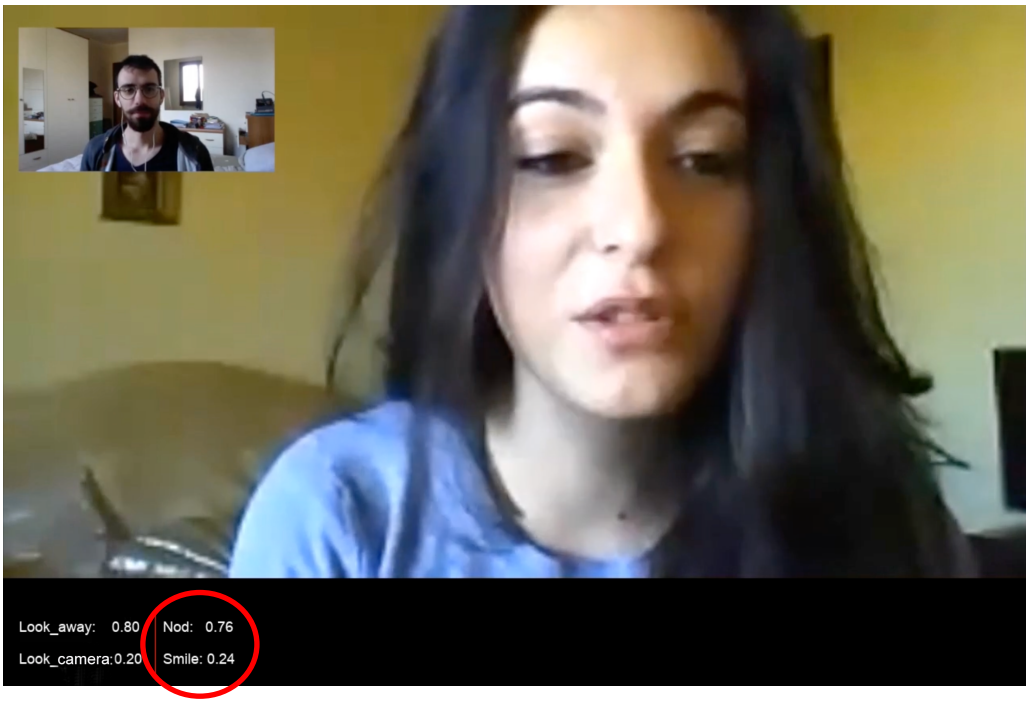
Gaze						
set	**test set**					**Acc**
easy	set16	set18	set11	set17	set1	0.8641
normal	set20	set19	set4	set15	set12	0.7838
intermediate	set14	set10	set7	set3	set5	0.5718
hard	set8	set13	set9	set2	set6	0.4796
Average						0.674825

Table 32: Approach 2 – Action subsets

Action						
set	**test set**					**Acc**
easy	set2	set1	set14	set3	set12	0.9010
normal	set6	set19	set16	set8	set10	0.8340
intermediate	set7	set18	set17	set13	set5	0.8115
hard	set11	set9	set15	set20	set4	0.8026
Average						0.8373

At this point, the research team observed that the high prediction accuracy of the Action dataset was due to misdetection. The network seemed to correctly detect Smiles but predicted Nods whenever the subject was not smiling (Figure 19).

Figure 19: Nod prediction



To overcome this issue, a third dataset with an additional class (Nods, Smiles and Neutral) was created, named Action_Neutral. For the Action_Neutral dataset, the frames selected for the Neutral class were sampled randomly given all the frames that were not annotated as Nods or Smiles and consisted of the 50% of the Action_Neutral dataset (Table 33).

Table 33: Datasets

Dataset	Frames	Labels	
Gaze	347861	Look_camera	123869
		Look_away	223992

Action	67980	Nod	28555
		Smile	39425
Action_Neutral	135066	Nod	28555
		Smile	39425
		Neutral	67086

Consequently, the three datasets (Gaze, Action and Action_Neutral) were used to train and test the system. Approach 2 was used to identify easiest to hardest examples and to outline the four difficulty subsets: easy, normal, intermediate and hard (Table 34). As stated above, the results show an average prediction accuracy for Gaze behaviour of .71 and average prediction accuracy for Action behaviour of .85. As expected, the results also show a lower average prediction accuracy in the Action_Neutral dataset (.63).

Table 34: Datasets difficulty

	gaze		action		action_neutral	
Difficulty Level		Acc		Acc		Acc
easy	set16	0.9896	set2	0.9618	set1	0.8173
	set18	0.98	set1	0.9294	set14	0.737
	set11	0.9508	set14	0.9257	set10	0.7338
	set17	0.9444	set3	0.8939	set18	0.6676
	set1	0.9121	set12	0.8934	set9	0.6478
normal	set20	0.894	set6	0.8915	set11	0.6397
	set19	0.8866	set19	0.8636	set2	0.6369
	set4	0.8722	set16	0.8609	set3	0.6275

	set15	0.7594	set8	0.8503	set6	0.6121
	set12	0.7426	set10	0.8481	set7	0.6119
intermediate	set14	0.685	set7	0.8373	set4	0.6065
	set10	0.6743	set18	0.8332	set15	0.6036
	set7	0.6429	set17	0.831	set13	0.6025
	set3	0.5425	set13	0.8287	set17	0.5968
	set5	0.5404	set5	0.8183	set16	0.587
hard	set8	0.5061	set11	0.8182	set5	0.5859
	set13	0.449	set9	0.8169	set20	0.5693
	set9	0.437	set15	0.8101	set19	0.5686
	set2	0.4351	set20	0.801	set8	0.5631
	set6	0.3986	set4	0.7208	set12	0.4942
	Average	0.71213	Average	0.851705	Average	0.625455

Approach 3 was then finally applied to all datasets, considering the four subsets of difficulty. Therefore, the final form of the dataset consists of 4 difficulty sets for each dataset and 5 frame selection splits (Full, Neighbour, 5fpa, 3fpa, 1fpa).

Table 45: Gaze prediction accuracy

Gaze					
set	Accuracy				
	full	1fpa	3fpa	5fpa	neighbour
easy	0.8641	0.6979	0.6875	0.6573	0.8820

normal	0.7838	0.6491	0.6296	0.5664	0.8042
intermediate	0.5718	0.6533	0.6899	0.6028	0.6152
hard	0.4796	0.6591	0.6268	0.6091	0.4928
Average	0.6748	0.6649	0.6585	0.6089	0.6986

Table 36: Action prediction accuracy

Action					
set	Accuracy				
	full	1fpa	3fpa	5fpa	neighbour
easy	0.9010	0.9453	0.9087	0.8778	0.9169
normal	0.8340	0.8711	0.8750	0.8185	0.8887
intermediate	0.8115	0.8242	0.8175	0.7299	0.8789
hard	0.8026	0.7969	0.7862	0.7324	0.8694
Average	0.8373	0.8594	0.8469	0.7897	0.8885

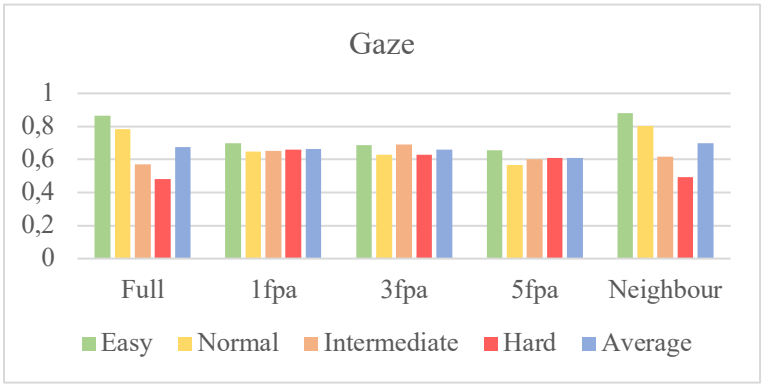
Table 37: Action_Neutral prediction accuracy

Action_Neutral					
set	Accuracy				
	full	1fpa	3fpa	5fpa	neighbour
easy	0.6576	0.5938	0.6203	0.5739	0.6593
normal	0.6078	0.5718	0.5451	0.5105	0.5414

intermediate	0.5981	0.5539	0.5818	0.5383	0.5957
hard	0.5722	0.5520	0.5304	0.5056	0.5566
Average	0.6089	0.5679	0.5694	0.5321	0.5883

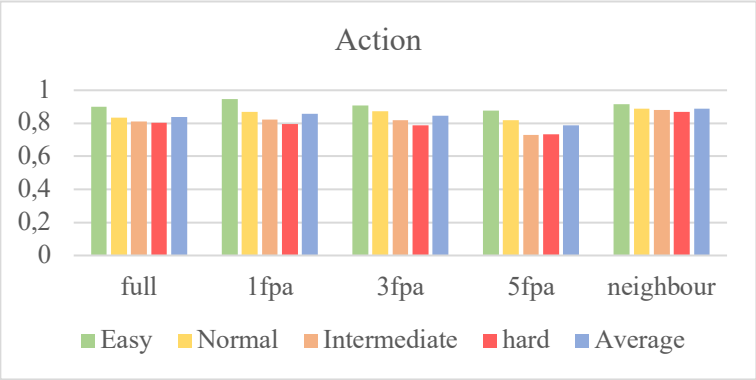
With regard to Gaze, the prediction accuracy ranged between .66 and .86 in the Easy subset and between .48 and .66 in the Hard subset (Table 35). The subsets that provided the most accuracy were Full and Neighbour, with an average accuracy of .67 and 0.70, so around 70% (Figure 20).

Figure 20: Gaze



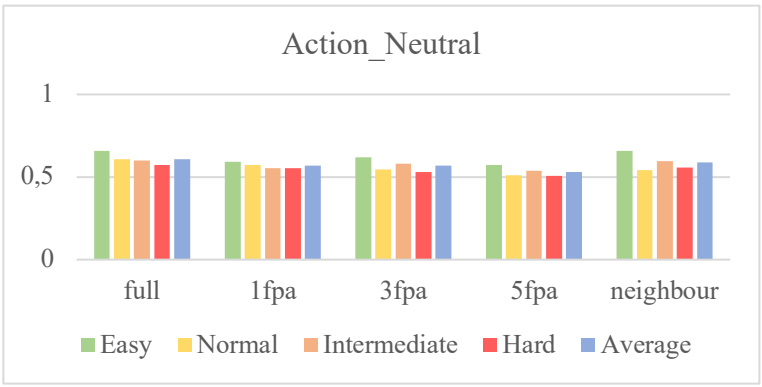
Regarding Action, the prediction accuracy ranged between .88 and .94 in the Easy subset and accuracy between .73 and .87 in the Hard subset (Table 36). The subsets that provided the most accuracy were 1fpa and Neighbour, with an average accuracy of .86 and .89, so around 85-90% (Figure 21).

Figure 21: Action



As for Action_Neutral, the prediction accuracy ranged between .57 and 0.66 in the Easy subset and accuracy between .51 and .57 in the Hard subset (Table 37). The subsets that provided the most accuracy were Full and Neighbour, with an average accuracy of .61 and .59, so around 60% (Figure 22).

Figure 22: Action_Neutral



8.5. Discussion

The aim of this study was to build a model capable of extracting and predicting the Gaze, Nodding and Smiling behaviours found and tested in previous studies. This, in order to be able to automatically extract these cues and thus consider a future application of our behaviour coding system during remote job interviews. The results show that the best way of splitting up the three datasets (Gaze, Action and Action_Neutral) were approaches 2 and 3. In fact, the datasets were divided into four subsets according to the difficulty of predicting the behaviour (Easy, Normal, Intermediate and Hard) and into five frame selection splits (Full, Neighbour, 5fpa, 3fpa, 1fpa). Once the datasets have been divided in this way, a methodology similar to the one used during the first approach was used (5 sets as test set and 15 for training set). As seen above, the system was able to predict engagement-related behaviours with good accuracy. In fact, as far as the Gaze is concerned, the system predicted it with an average accuracy of around 70%. Action behaviours, on the other hand, were initially predicted with an accuracy of about 90%. It was realised, however, that this result was fallacious. The system detected nods whenever the person was not smiling (Figure 19). Adding a class called Neutral, therefore, resulted in an accuracy of 61%. Certainly lower but much more accurate than the previous one. Moreover, the results show that the frame splits that work best for this system are Full and Neighbour. Thus showing that these behaviours are more easily detectable during the whole duration of the behaviour or in the 20% of frames that are in the middle of the annotation. It is therefore not sufficient to select single frames at different points in the annotation to detect these behaviours with greater accuracy. In conclusion, the system presented proved to be quite reliable in predicting behaviour during Zoom work interviews. A future application, therefore, could be that of an online behaviour-

recognition-system, i.e. while the interaction is taking place. This automatic annotation can be used to provide feedback to both recruiters and candidates to improve their performance. It could also be a useful tool for assessing the candidate's performance both for the candidate himself, who could then reflect on his performance and for the recruiter, who could take the automatically extracted information into account for his final decision. It should be stressed that this would be a tool and not the final assessor himself and that, as such, it would provide outputs that would have to be weighed and contextualised. The recruiter using such a tool could take it into account as he/she does with other candidate screening tools. The advantage that this tool could provide is the speed and immediacy of one of the possible assessment criteria that would otherwise require hours and hours of time.

9. General discussion

In our first study, 57 body language cues (grouped in 9 Behaviors) indicating engagement or disengagement during a dyadic interaction were identified. Facial expressions, Backchannels and Gaze have been found to be central in the composition of engagement during this kind of conversation. Of these 3 Behaviours, 9 cues observable via the Zoom platform were selected. These cues were associated with suggestions to be given to the participants in a training administered during the second study. The participants were therefore divided into three groups: those who received the training (Training), those who watched the video of their first round of questions without any expert suggestion (Video re-watch) and those who did not receive the training and did not watch the video (No training). The results obtained show a significant difference between the three conditions as far as Gaze is concerned. It seems that the training increased the behaviour of Looking into the camera and decreased the behaviour of Looking away (Figures 2-3). Also with regard to Nodding behaviour, the training proved to be effective (Figure 4). However, in contrast to what was expected, the training does not seem to have had an effect on Smile behaviour (Figure 5). This result could be due to the actual ineffectiveness of the training in improving smiling behaviour, or it could be due to the role played by the participants. It seems that the participants who played the role of Interviewer nodded and smiled more (Figures 6-7). It seems plausible that the Interviewers, by speaking less, needed to provide their interlocutor with feedback such as nodding and smiling to let him/her know they were listening. The role, on the other hand, seems not to have had an effect on perceived engagement. Interviewers and Interviewees, in fact, were not rated as significantly more or less engaged by the independent commission (Figure 9), suggesting that our training was effective in both types of roles. This, therefore, suggests the possibility of applying

our training to both job interview candidates and recruiters, providing the possibility of an improvement in engagement and therefore communication with the interlocutor.

In line with the results of the second study, it was also found that the training actually had an effect in increasing engagement in the second round of interviews compared to the first (Figure 8). This show the effectiveness of the training in improving the overall engagement of the trainees who participated in the Zoom interview. Above all, it is shown that the expert's advice, based on nonverbal cues, has an effect in improving trainees' behaviour. In fact, the trainees not only reviewed their first interview but also received the expert's suggestions, in contrast to the Video Rewatch condition, in which the interviewees only reviewed their first interview.

I then verified whether gender had an effect on the engagement rates given by the evaluation committee. The effect of the gender of the respondents to the questionnaire (evaluators), the gender of the protagonists of our Zoom interview and the combination of the gender of the evaluators with that of the protagonists they evaluated was therefore analysed. A slightly significant effect was found with regard to the gender of the evaluators, showing marginally higher engagement ratings given by males (Figure 10). Nevertheless, female protagonists seemed to be evaluated marginally as more engaged than the male protagonists (Figure 11). This significance, however, disappears when the combination of evaluators and protagonists' genders are considered (Figure 12), suggesting the effectiveness of the proposed training regardless of biases related to the gender of the protagonist/trainee.

In addition, the correlation between the engagement scores given by the evaluators and the frequency of the annotated behaviour was assessed. It was found that these two dimensions were generally correlated. Gazes were found to be correlated with engagement scores, showing that more engagement corresponds to more frequent Looks

in the camera and less frequent Looks away. However, Nodding appeared to be uncorrelated with engagement in this work. This may be due to the difficulty for external evaluators to use this cue as an element to assess engagement. It is a behaviour that often lasts a few fractions of a second. The frequency of Smiles, on the other hand, appears to correlate with engagement scores when weighted on the higher frequency of smiles in this study. This could be due to the scarcity of Smile annotations compared to Gaze annotations (present for the entire duration of each video).

The Behaviours I focused on, therefore, appear to be central in building engagement, during job-related interviews: Facial Expressions, Gaze and Backchannels. Furthermore, in literature the body cues this project focused on seem to be correlated with more extroverted personalities when considering the big five personality theory (Neff et al., 2010). The position of the head and trunk are considered to be visually valid indicators of status and attitude. Leaning towards the interlocutor and not turning away communicates a positive attitude and seems to be correlated with the extroversion trait while leaning back or turning away communicates a more negative attitude (Lippa, 1998). Moreover, it seems that extroverted personalities tend to amplify their personal space by moving their upper body forward. In contrast, introverts maintain a more vertical orientation (Frank, 2007). Finally, it seems that extroverts maintain more eye contact with the interlocutor, as well as an orientation of body, shoulders and legs turned towards their addressee (Mehrabian, 1969). Thus, it appears that the cues in this study that have been shown to be central to defining the concept of engagement also correlate with extroverted personality in the literature. A person who is positively engaged in a job interview, therefore, might have the characteristics of an extroverted personality. Future studies may test this hypothesis.

Finally, during my period abroad in Thessaloniki, a computational system was constructed to predict the behaviours identified as central to the definition of Physical Mutual Engagement (PME). The ability to automatically identify these behaviours opens up many possibilities, both for candidates and recruiters. One is to build software that takes advantage of this system and provides feedback to both recruiters and candidates. This feedback could be provided in real-time or after the interview. If provided in real-time, both parties could modify their behaviour to improve their engagement during the conversation. If provided after the fact, they could use this information both to assess the performance of the interview and to improve their own performance in the future.

10. Conclusions

The objectives of this study were to gather evidence on non-verbal cues that effectively convey that the speaker is involved in the interaction when this interaction takes place in a videoconferencing system. Given the extensive use that has been made of these systems in recent years and the anticipated future use for years to come, both in the interests of accurate assessment when this takes place via videoconferencing and of interaction-free from misunderstanding, awareness and mastery of these cues is therefore very useful. Certainly, a tool that can detect these cues could be of great help to those involved in recruitment. In a world that increasingly envisages working and collaborating remotely, understanding such dynamics, however, seems to be useful for any type of worker. Their promise is to work faster but also away from the subjective bias of a human recruiter. But who are we filtering and hiring, good-looking and compliant minorities? People with a high ability to look at themselves through the eyes of the other (looking-glass self)? The ability to see and adapt to the perspective of the other could lead to success but also be the basis for manipulation and deception. Courses on how to look and talk to raise the level of candidates on that portion of performance represented only by video experience and easily overcome with a minimum of training. Mutual symmetry of intent and levelling of the point of enunciation is simply achieved through transparency, as this thesis postulates. The necessary training is simply that in which the few cues are revealed and pointed. Software that tests performance with annotations is sufficient to avoid misunderstandings and bases semi-automatic processes on ethical principles putting not only the candidate but also the recruiter under scrutiny. There is nothing magical in the elaboration of the algorithm, only the identification of a combination of polysemous cues.

Equally important is to clarify that no bias is introduced into the AI because of the method by which the classification was constructed (Srinivasan & Chander, 2021). In fact, HireVue has been targeted by a research centre in Washington DC, the Electronic Privacy Information Center, which has filed an official complaint with the US Federal Trade Commission that it does not disclose how candidates' personal data is used; in countries such as the EU, this is considered necessary before processing it. Specific regulations have been proposed to oblige Ai-based recruitment tools to disclose the criteria used in their algorithms.

References

Amalfitano, J. G., & Kalt, N. C. (1977). Effects of eye contact on the evaluation of job applicants. *Journal of Employment Counseling*, *14*(1), 46–48.

Andersen, J. F. (1979). Teacher Immediacy as a Predictor of Teaching Effectiveness. *Annals of the International Communication Association*, *3*(1), 543–559. https://doi.org/10.1080/23808985.1979.11923782

Andersen, J. F., Andersen, P. A., & Jensen, A. D. (1979). The Measurement Of Nonverbal Immediacy. *Journal of Applied Communication Research*, *7*(2), 153–180. https://doi.org/10.1080/00909887909365204

Anderson, N., & Shackleton, V. (1990). Decision making in the graduate selection interview: A field study. *Journal of Occupational Psychology*, *63*(1), 63–76. https://doi.org/10.1111/j.2044-8325.1990.tb00510.x

Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., & Lawless, M. (2019). Using Zoom Videoconferencing for Qualitative Data Collection: Perceptions and Experiences of Researchers and Participants. *International Journal of Qualitative Methods*, *18*, 1–8. https://doi.org/10.1177/1609406919874596

Arminen, I., Licoppe, C., & Spagnolli, A. (2016). Respecifying Mediated Interaction. *Research on Language and Social Interaction*, *49*(4), 290–309. https://doi.org/10.1080/08351813.2016.1234614

Bailenson, J. N. (2021). Nonverbal overload: A theoretical argument for the causes of Zoom fatigue. *Technology, Mind, and Behavior*, *2*(1). https://doi.org/10.1037/tmb0000030

Basch, J. M., Melchers, K. G., Kurz, A., Krieger, M., & Miller, L. (2020). It Takes More Than a Good Camera: Which Factors Contribute to Differences Between Face-to-Face Interviews and Videoconference Interviews Regarding Performance Ratings

and Interviewee Perceptions? *Journal of Business and Psychology*, *Im*. https://doi.org/10.1007/s10869-020-09714-3

Bjornsdottir, R. T., & Rule, N. O. (2017). The Visibility of Social Class From Facial Cues. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspa0000091.supp

Braun, V., Clarke, V., & Gray, D. (2017). *Collecting Qualitative Data* (V. Braun, V. Clarke, & D. Gray, Eds.). Cambridge University Press. https://doi.org/10.1017/9781107295094

Burgoon, J. K., Birk, T., & Pfau, M. (1990). Nonverbal Behaviors, Persuasion, and Credibility. In *Human Communication Research* (Vol. 17, Issue 1).

Chen, L., Feng, G., Leong, C. W., Lehman, B., Martin-Raugh, M., Kell, H., Lee, C. M., & Yoon, S. Y. (2016). Automated scoring of interview videos using Doc2vec multimodal feature extraction paradigm. *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 161–168. https://doi.org/10.1145/2993148.2993203

Coker, D. A., & Burgoon, J. (1987). The Nature of Conversational Involvement and Nonverbal Encoding Patterns. *Human Communication Research*, *13*(4), 463–494. https://doi.org/10.1111/j.1468-2958.1987.tb00115.x

Corbin, J. M., Strauss, A. L., & Strauss, A. L. (2008). *Basics of qualitative research : techniques and procedures for developing grounded theory*. 379.

Cristel, R. T., Demesh, D., & Dayan, S. H. (2020). Video conferencing impact on facial appearance: Looking beyond the COVID-19 pandemic. In *Facial Plastic Surgery and Aesthetic Medicine* (Vol. 22, Issue 4, pp. 238–239). Mary Ann Liebert Inc. https://doi.org/10.1089/fpsam.2020.0279

Edinger, J. A., & Patterson, M. L. (1983). Nonverbal involvement and social control.

Psychological Bulletin, *93*(1), 30–56. https://doi.org/10.1037/0033-2909.93.1.30

Forbes, R. J., & Jackson, P. R. (1980). Non-verbal behaviour and the outcome of selection interviews. *Journal of Occupational Psychology*, *53*(1), 65–72. https://doi.org/10.1111/j.2044-8325.1980.tb00007.x

Frank, K. (2007). *Posture and Perception in the Context of the Tonic Function Model of Structural Integration copy* (pp. 27–35). IASI Yearbook.

Gifford, R., Ng, C. F., & Wilkinson, M. (1985). Nonverbal Cues in the Employment Interview. Links Between Applicant Qualities and Interviewer Judgments. *Journal of Applied Psychology*, *70*(4), 729–736. https://doi.org/10.1037/0021-9010.70.4.729

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The Illusion of Transparency: Biased Assessments of Others' Ability to Read One's Emotional States. *Journal of Personality and Social Psychology*, *75*(2), 332–346. https://doi.org/10.1037/0022-3514.75.2.332

Grondin, F., Lomanowska, A. M., Békés, V., & Jackson, P. L. (2020). A methodology to improve eye contact in telepsychotherapy via videoconferencing with considerations for psychological distance. *Counselling Psychology Quarterly*, *00*(00), 1–14. https://doi.org/10.1080/09515070.2020.1781596

Hassin, R. R., Aviezer, H., & Bentin, S. (2013). Inherently ambiguous: Facial expressions of emotions, in context. In *Emotion Review* (Vol. 5, Issue 1, pp. 60–65). https://doi.org/10.1177/1754073912451331

Henry, A., & Shellenbarger, T. (2020). To Zoom or not to Zoom? Choosing a videoconferencing platform. *Nurse Author & Editor*, *30*(4), 30–34. https://doi.org/10.1111/nae2.9

Iacono, V. lo, Symonds, P., & Brown, D. H. K. (2016). Skype as a tool for qualitative research interviews. *Sociological Research Online*, *21*(2), 1–15. https://doi.org/10.5153/sro.3952

Imada, A. S., & Hakel, M. D. (1976). Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews. *Journal of Applied Psychology*, *62*(3), 295–300. https://doi.org/10.1037/0021-9010.62.3.295

Irani, E. (2019). The Use of Videoconferencing for Qualitative Interviewing: Opportunities, Challenges, and Considerations. *Clinical Nursing Research*, *28*(1), 3–8. https://doi.org/10.1177/1054773818803170

Janghorban, R., Roudsari, R. L., & Taghipour, A. (2014). Skype interviewing: The new generation of online synchronous interview in qualitative research. *International Journal of Qualitative Studies on Health and Well-Being*, *9*(1). https://doi.org/10.3402/qhw.v9.24152

Lippa, R. (1998). The Nonverbal Display and Judgment of Extraversion, Masculinity, Femininity, and Gender Diagnosticity: A Lens Model Analysis. *Journal of Research in Personality*, *32*(1), 80–107. https://doi.org/10.1006/jrpe.1997.2189

Marhefka, S., Lockhart, E., & Turner, D. A. (2020). Achieve Research Continuity During Social Distancing by Rapidly Implementing Individual and Group Videoconferencing with Participants: Key Considerations, Best Practices, and Protocols. *AIDS and Behavior*, *24*(7), 1983–1989. https://doi.org/10.1007/s10461-020-02837-x

McGinley, H., LeFevre, R., & McGinley, P. (1975). The influence of a communicator's body position on opinion change in others. *Journal of Personality and Social Psychology*, *31*(4), 686–690. https://doi.org/10.1037/0022-3514.31.4.686

McGovern, T. V. (1976). *The making of a job interviewee: The effect of nonverbal behavior on an interviewer's evaluations during a selection interview.* Southern Illinois University at Carbondale.

Mehrabian, A. (1968a). Relationship of attitude to seated posture, orientation, and distance. *Journal of Personality and Social Psychology*, *10*(1), 26–30. https://doi.org/10.1037/h0026384

Mehrabian, A. (1968b). Some referents and measures of nonverbal behavior. *Behavior Research Method and Instruction*, *1*(6), 203–207.

Mehrabian, A. (1969). *Significance of Posture and Position in the Communication of Attitude and Status Relationships*. *71*(5), 359–372.

Mehrabian, A. (1971). Silent Messages. In *Unfriendly Fire*. https://doi.org/10.2307/j.ctt20fw8nb.9

Mehrabian, A., & Friar, J. T. (1969). Encoding of attitude by a seated communicator via posture and position cues. *Journal of Consulting and Clinical Psychology*, *33*(3), 330–336. https://doi.org/10.1037/h0027576

Mirick, R. G., & Wladkowski, S. P. (2019). Skype in qualitative interviews: Participant and researcher perspectives. *Qualitative Report*, *24*(12), 3061–3072.

Moran, N., Hadley, L. v., Bader, M., & Keller, P. E. (2015). Perception of "back-channeling" nonverbal feedback in musical duo improvisation. *PLoS ONE*, *10*(6). https://doi.org/10.1371/journal.pone.0130070

Muralidhar, S., Nguyen, L. S., Frauendorfer, D., Odobez, J. M., Mast, M. S., & Gatica-Perez, D. (2016). Training on the job: Behavioral analysis of job interviews in hospitality. *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 84–91. https://doi.org/10.1145/2993148.2993191

Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2015). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*. https://doi.org/10.1109/FG.2015.7163127

Nakano, Y. I., & Ishii, R. (2010). Estimating user's engagement from eye-gaze behaviors in human-agent conversations. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 139–148. https://doi.org/10.1145/1719970.1719990

Neff, M., Wang, Y., Abbott, R., & Walker, M. (2010). Evaluating the Effect of Gesture and Language on Personality Perception in Conversational Agents. *Intelligent Virtual Agents*, 222–235. http://dx.doi.org/10.1007/978-3-642-15892-6_24

Nguyen, L. S., Frauendorfer, D., Mast, M. S., & Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*, *16*(4), 1018–1031. https://doi.org/10.1109/TMM.2014.2307169

Nguyen, L. S., & Gatica-Perez, D. (2016). Hirability in the Wild: Analysis of Online Conversational Video Resumes. *IEEE Transactions on Multimedia*, *18*(7), 1422–1437. https://doi.org/10.1109/TMM.2016.2557058

Rasipuram, S., Das, R., Pooja Rao, S. B., & Jayagopi, Di. B. (2018). Online peer-to-peer discussions: A platform for automatic assessment of communication skill. *2017 7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2017*, *2018-Janua*, 68–73. https://doi.org/10.1109/ACIIW.2017.8272588

Rasipuram, S., & Jayagopi, D. B. (2016). Automatic assessment of communication skill in interface-based employment interviews using audio-visual cues. *2016 IEEE*

International Conference on Multimedia and Expo Workshop, ICMEW 2016. https://doi.org/10.1109/ICMEW.2016.7574733

Rasipuram, S., & Jayagopi, D. B. (2019). A comprehensive evaluation of audio-visual behavior in various modes of interviews in the wild. *ACM International Conference Proceeding Series*, 94–100. https://doi.org/10.1145/3316782.3321528

Rice, S. M., Graber, E., & Kourosh, A. S. (2020). A Pandemic of Dysmorphia: "zooming" into the Perception of Our Appearance. In *Facial Plastic Surgery and Aesthetic Medicine* (Vol. 22, Issue 6, pp. 401–402). Mary Ann Liebert Inc. https://doi.org/10.1089/fpsam.2020.0454

Richmond, V. P., Gorham, J. S., & Mccroskey, J. C. (1987). The Relationship Between Selected Immediacy Behaviors and Cognitive Learning. *Annals of the International Communication Association*, *10*(1), 574–590. https://doi.org/10.1080/23808985.1987.11678663

Schmid Mast, M., Gatica-Perez, D., Frauendorfer, D., Nguyen, L., & Choudhury, T. (2015). Social Sensing for Psychology: Automated Interpersonal Behavior Assessment. *Current Directions in Psychological Science*, *24*(2), 154–160. https://doi.org/10.1177/0963721414560811

Schroff, F., & Philbin, J. (2015). *FaceNet: A Unified Embedding for Face Recognition and Clustering*.

Sedgwick, M., & Spiers, J. (2009). The Use of Videoconferencing as a Medium for the Qualitative Interview. *International Journal of Qualitative Methods*, *8*(1), 1–11. https://doi.org/10.1177/160940690900800101

Singh, R., & Soumya, A. (2020). Updated comparative analysis on video conferencing platforms- Zoom, Google Meet, Microsoft Teams, WebEx Teams and

GoToMeetings. *EasyChair: The World for Scientists*, 1–9. https://easychair.org/publications/preprint/Fq7T

Spears, R., & Lea, M. (1992). *Social influence and the influence of the "social" in computer-mediated communication ANGCALM: A Multi-Nation Comparison of Theories of Anger/Calm Judgements View project A normative perspective of intergroup discrimination View project*. http://martinlea.com

Sproull, L., & Kiesler, S. (1991). Connections: New ways of working in the networked organization. *MIT Press*.

Srinivasan, R., & Chander, A. (2021). Biases in AI systems. In *Communications of the ACM* (Vol. 64, Issue 8, pp. 44–49). Association for Computing Machinery. https://doi.org/10.1145/3464903

Szafir, D., & Mutlu, B. (2012). Pay attention! Designing Adaptive Agents that Monitor and Improve User Engagement. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (*, 11–20. https://doi.org/10.1080/20507828.2017.1350794

Todorov, A., & Oh, D. W. (2021). The structure and perceptual basis of social judgments from faces. In *Advances in Experimental Social Psychology* (Vol. 63, pp. 189–245). Academic Press Inc. https://doi.org/10.1016/bs.aesp.2020.11.004

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, *23*(1), 3–43. https://doi.org/10.1177/009365096023001001

Washburn, P. V, & Hakel, M. D. (1973). Visual cues and verbal content as influences on impressions formed after simulated employment interviews. *Journal of Applied Psychology*, *58*(1), 137.

Wen, Z., Lin, W., Wang, T., & Xu, G. (2021). *Distract Your Attention: Multi-head Cross Attention Network for Facial Expression Recognition*. http://arxiv.org/abs/2109.07270

Wexley, K. N., Fugita, S. S., & Malone, M. P. (1975). An applicant's nonverbal behavior and student-evaluators' judgments in a structured interview setting. *Psychological Reports*, *36*(2), 391–394.

Whittaker, S. (2002). *Theories and Methods in Mediated Communication*. https://www.researchgate.net/publication/2587655

Young, D. M., & Beier, E. G. (1977). The role of applicant nonverbal communication in the employment interview. *Journal of Employment Counseling*, *14*(4), 154–165.