Noname manuscript No.
(will be inserted by the editor)

Training public speaking with virtual social interactions: Effectiveness of real-time feedback and delayed feedback

Mathieu Chollet · Stacy Marsella · Stefan Scherer

Received: date / Accepted: date

Abstract Social signal processing and virtual social interaction technologies have allowed the creation of social skills training applications, and initial studies have shown that such solutions can lead to positive training outcomes and could complement traditional teaching methods by providing cheap, accessible, safe tools for training social skills. However, these studies evaluated social skills training systems as a whole and it is unclear to what extent which components contributed to positive outcomes. In this paper, we describe an experimental study where we compared the relative efficacy of real-time interactive feedback and after-action feedback in the context of a public speaking training application. We observed that both components provide benefits to the overall training: the real-time interactive feedback made the experience more immersive and improved participants' motivation in using the system, while the after-action feedback led to positive training outcomes when it contained personalized feedback elements. Taken in combination, these results confirm that both social signal processing technologies and virtual social interactions are both contributing to social skills training systems' efficiency. Additionally, we observed that several individual factors, here the subjects' initial level of public speaking anxiety, personality and tendency to immersion significantly influenced the training experience. This finding suggests that social skills training systems could benefit from being tailored to participants' particular individual circumstances.

Keywords public speaking · oral competence · social signals processing · virtual social interactions

1 Introduction

Oral communication skills are essential in many modern professions and are also widely useful in various situations of our personal lives. Indeed, oral communication and presentation skills have been identified as a core skill for graduates across disciplines [35,12]. Unfortunately, public speaking can be a very anxiety-provoking situation, and it is the most commonly cited stressful situation by sufferers of social phobia [25].

Training oral competence is a complex topic, and a wide variety of pedagogical components may be implemented in a training curriculum, such as modeling [4] (*i.e.* observing models of experts or peers demonstrating appropriate behavior) or feedback. Surprisingly, a common pedagogical framework for oral competence training accounting for the effectiveness of these different components in different contexts is still lacking [10, 21,9,26]. Still, a plethora of studies have been devoted to the evaluation of different sub-components, and there is clear evidence that several pedagogical components

This material is based upon work supported by the U.S. Army Research Laboratory under contract number W911NF-14-D-0005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Government, and no official endorsement should be inferred.

M. Chollet
IMT Atlantique, 4 rue Alfred Kastler, 44300, Nantes, France
E-mail: mathieu.chollet@imt-atlantique.fr

S. Marsella
University of Neuroscience and Psychology, University of Glasgow, 62 Hillhead St, Glasgow G12 8AD, United Kingdom
E-mail: stacy.marsella@glasgow.ac.uk

S. Scherer
Institute for Creative Technologies, University of Southern California, 12015 E Waterfront Dr, Playa Vista, CA 90094, USA
E-mail: scherer@ict.usc.edu

are beneficial for training, such as having the opportunity to practice public speaking or receiving personalized feedback [26]. These activities can however be challenging to integrate in modern demanding teaching environments, as they can be costly, time-consuming, difficult to standardize, and as participants subject to social anxieties might be reluctant to participate actively in such settings.

Technological innovations in the domains of virtual reality and automated social signal processing have recently been combined to create social skills training applications which hold the potential to provide solutions to these challenges. Virtual simulations of social situations can be created and used as training environments mimicking their real counterparts, such as a simulated job interview with virtual recruiters [2,32]. Additionally, trainees' behaviors can be automatically measured using speech recognition, facial recognition and other social signal processing techniques [54,55, 33], and these measurements can subsequently be used to generate personalized feedback to the trainees. Such solutions could provide efficient training tools, complementing traditional teaching approaches in the context of increasing student populations and increasing pressure for cost-effective solutions [27].

However, in order to realize this potential, it is crucial to assess systematically the efficacy of these applications and the relative contributions of their different components. Perhaps due in part to the complexity of creating and evaluating such systems, and in part to the lack of overall pedagogical framework for designing social skills training curricula, few studies systematically compared the effectiveness of different technological components of social skills training applications. In this article, we describe a study aimed at disentangling the relative efficacy of real-time interactive feedback and after-action feedback. We leverage the Cicero public speaking training framework, which features an interactive virtual audience and social signal processing components. We consider real-time feedback produced by the simulation's virtual characters and delayed feedback through an after-action report consisting of the users' video-taped presentation enriched with personalized feedback on several aspects of the users' performance. In the following section, we begin by discussing existing research which evaluated social skills training applications featuring various forms of feedback. In section 3, we present the Cicero public speaking training framework which we used in this paper. Our experiment is then described in section 4 and we discuss the implications of our results in section 5.

2 Related work

Social signal processing and virtual humans are very active research topics, and their resulting technologies are becoming robust enough to be integrated into software applications. In particular, social skills training stands to benefit strongly from them as they can be used to create simulations of the social situation to train and to automatically assess trainees. Using cameras, microphones or other sensors, the trainee's vocal, verbal and non-verbal behaviors can be detected using computer vision and signal processing techniques, and then analyzed to compute scores and metrics describing the trainee's performance [46,57,53,43,15]. Additionally, feedback can be provided to the trainee, either in real-time during their presentation - for instance using graphical icons [20,52] or through the behavior of the virtual humans populating the simulation [2,17] - or through an additional after-action report step. This usually consists of an interface displaying the video of the trainee's performance along with personalized feedback information, and allows trainees to reflect on their performance after the training simulation [24,1,53,32, 2].

2.1 Virtual social interactions as social skills training environments

Interactive systems using virtual humans as social interaction partners have shown success in improving social skills, such as in the domain of job interview training with the MACH [32] and TARDIS [19] systems. MACH was tested with 90 undergraduate students (53 female and 37 males) from the MIT campus [32]. The experiment design consisted of three phases where the participants first interacted with a human counselor (considered the baseline assessment), then depending on condition they interacted with a specific version of MACH (with feedback or not) or watched a 30 minute educational video. Finally they interacted with the same human counselor for post-intervention assessment. The human counselors were blind to the conditions, which consisted of a control condition where participants watched the educational video, a condition using MACH without feedback and a condition using MACH with feedback. The results showed a significant improvement in job interview skills between the MACH no-feedback and MACH with-feedback conditions and between the MACH with-feedback and the control condition. The TARDIS job interview serious game was evaluated using a similar protocol with three phases [19]. A first baseline assessment was performed in a mock interview with a human counselor. In a second phase, half of the

participants interacted with the TARDIS system while the others read training material about job interview. Finally, a third phase consisted of another mock interview with a human counselor blind to the training condition to assess improvement after training. The participants interacting with the system improved significantly more than the group using a traditional method on a variety of measures including overall performance.

A particular interface paradigm for public speaking training is the virtual audience. Such a system aims at reproducing a public speaking situation with high fidelity, using an environment that is typical of public speaking situations (*e.g.* a conference room) and populating it with virtual characters acting as the audience.

Virtual audiences were first investigated to treat public speaking anxiety. North *et al.* found that virtual audiences were effective in inducing stress and reducing public speaking anxiety [29]. Researchers also investigated the effect of three different types of virtual audiences, namely a neutral, a positive, and a negative audience, consisting of eight virtual characters [42]. They found that in all three settings, participants experienced a significant level of anxiety, even in participants who did not report being particularly anxious about public speaking. A randomized clinical trial evaluating psychotherapy with virtual audiences was conducted by Safir *et al.* [48]. Here, 88 participants were randomly assigned to one of 3 conditions: a waiting list, cognitive-behavior therapy (CBT) with imagination (participants had to imagine a public speaking situation), and CBT with virtual reality exposure including virtual audiences. Using self-rating anxiety questionnaires, the authors found a statistically significant reduction of anxiety in both CBT groups, which was maintained for both groups a year after the intervention. While there was no difference in anxiety reduction between the two CBT groups, the virtual reality CBT group suffered much lower attrition rates, suggesting an additional benefit to CBT with virtual audiences over regular approaches. Additionally, [45] conducted a large scale clinical trial comparing a group undergoing CBT with virtual reality exposure with virtual audiences and a group undergoing group exposure therapy and found similar results: both groups benefited from a significant reduction in public speaking anxiety, maintained a year after the treatment. No difference was found between the virtual audience condition and traditional group therapy. In a related domain, Bissonnette *et al.* studied the use of virtual audiences for reducing musicians' performance anxiety, successfully reducing it after 6 sessions spread over 3 weeks [7].

Beyond exposure therapy for socially anxious individuals, Chollet *et al.* proposed to use virtual audience systems for training public speaking ability in the general population [17]. They developed an experimental paradigm that compared a passive virtual audience, a passive audience enriched with graphical feedback and an interactive audience providing feedback on the trainee's behavior (as measured through a Wizard-of-Oz paradigm) through its non-verbal behavior (nodding and leaning forward to indicate positive performance, leaning back and shaking heads for negative performance). Experts rated the performance of 45 users before and after training with the different conditions and found that the graphical feedback paradigm fared worst, and that the passive and interactive audiences led to positive learning outcomes. However, while the interactive audience was rated as more engaging, challenging and useful by users, it did not lead to better outcomes than the passive audience. Still, the system was limited in that it was not fully automatic (user behaviors were manually detected) and the interactive audience's behavioral feedback had not been validated. In a further study [14], they compared the user experiences and training outcomes for native *vs* non-native English speakers using their system, and found that the level of language fluency impacted both the efficacy of the system as well as the quality of the user experience and user ratings of the system's quality.

2.2 Automated public speaking feedback

Several researchers have experimented with providing direct feedback to public speaking trainees [52, 20, 50]. In a study involving 30 students interacting with the Rhema public speaking training system, researchers evaluated the efficacy of continuous (*e.g.* line plots displaying continuous values) as well as sparse feedback (*e.g.* textual hints appearing when a condition is met, *e.g.* "LOUDER" when users speak too quietly) for such training systems [52]. All participants gave three presentations (with an average duration of 3 minutes) with systems providing continuous, sparse, or no feedback. The participants preferred the sparse feedback system which only provided brief periodical feedback over a feedback system providing continuous information. A post-hoc survey was conducted but differences in training outcomes between the feedback strategies were not found.

Logue [20] is a similar system that provides real-time feedback to presenters on their speech rate, body openness and body energy using a set of functional icons displayed on a Google Glass worn by the presenters. An experimental study demonstrated that Logue was well received by users and that observers rated participants using Logue with a higher openness than participants

in a control condition where the feedback icons were deactivated.

Barmaki and Hugues presented a system for training teachers to adopt better body postures, using a virtual classroom populated with manually controlled virtual students [5]. The authors used TeachLivE as their testbed environment, a virtual reproduction of a classroom setting populated with virtual students controlled manually, complementing it with an independent feedback application using a Microsoft Kinect to detect body postures and a screen displaying a green (respectively orange) stimulus when the detected posture was open (resp. defensive).

Schneider *et al.* introduced the Presentation Trainer system [50] which uses visual and haptic feedback (using a vibrating wristband) to provide feedback to trainees on mistakes (*e.g.* crossing arms, inappropriate volume) they make while presenting. During the training, the user's image is mirrored on a screen. When mistakes are detected, the user is first notified then interrupted using wristband vibrations and visual feedback displayed on the screen, explaining how to correct their mistake. The authors evaluated their system by comparing a group training with the full system to a control group where the system only mirrored the user's image without providing feedback. Participants receiving feedback were found to make significantly less mistakes than participants in the control condition.

The ROC Speak framework [58] allows public speaking trainees to record a video performance on which to receive combined automated non-verbal behavior feedback - *i.e.* graphical representations of the user's automatically detected smiles, movements and vocal behavior - with subjective feedback - *i.e.* comments from other trainees having watched the user's performance, automatically evaluated in terms of their helpfulness. The system was evaluated in a longitudinal study where trainees interacted with ROC Speak or with a control condition 5 times over 10 days [58]. The control condition consisted of a similar system without automated non-verbal feedback and automated ranking of other trainees' comments. The authors observed that ROC Speak led to significant training improvements, both in general and in comparison to the control condition. However, it is unclear what respective contributions are attributable to the automated behavior feedback or to the automated user comments ranking.

Another interactive system aimed a public speaking improvement is Automanner introduced by Tanveer *et al.* [53]. It is singular compared to the other systems we presented in the fact that it focuses on detecting and providing feedback on presenters' mannerisms, *i.e.* body movements frequently exhibited by a speaker, often unbeknownst to them. Automanner uses an automatic pattern extraction to detect users' mannerisms from their presentations recorded with a Microsoft Kinect. An interface is provided to the users after their speech to make them aware of their mannerisms. Analyses of users' self-reports show that even though they found the detection of mannerisms to be somewhat inaccurate, they still found value in the system as they report it made them aware of some of their mannerisms. However the authors did not investigate if the users actually exhibited fewer mannerisms after using the system compared to their initial state.

Mihoub and Lefebvre [40] proposed a model for describing the complex relationship between the cognitive state of a presenter, their public speaking performance, and the adequate feedback to produce to them in a training context. This model was represented using a Dynamic Bayesian Network (DBN) and relied on a combination of input signals from wearable devices. Using a set of rules derived empirically or from existing references, multimodal signals describing the user's performance are evaluated (*e.g.* low, good, excessive gesturing) using a 10s sliding window and fed into the DBN which provides the type of feedback to produce as an output. In an evaluation study using the model to provide feedback after participants' presentations, the system was judged by participants as relevant and useful for training, however actual training outcomes were not evaluated.

3 Cicero: an Interactive Public Speaking Training Framework

In this study, we leveraged the Cicero public speaking training platform. The Cicero system combines a public speaking simulation interactive system with an after-action report interface. The simulation system consists of an interactive virtual audience which reacts to the performance of the speaker. During interactions with the virtual audience, the trainee's behavior is monitored and a performance score is computed based on it. This score is then used to alter the non-verbal behavior of the virtual audience.

The Cicero system was developed over a series of iterations and is based on the Virtual Human Toolkit. Several components were reused or extended for the development of the Cicero system while others were developed specifically for its purposes. In Figure 1, the architecture of the system is outlined along with the relationships between the various modules. We describe those in more detail in the following subsections.

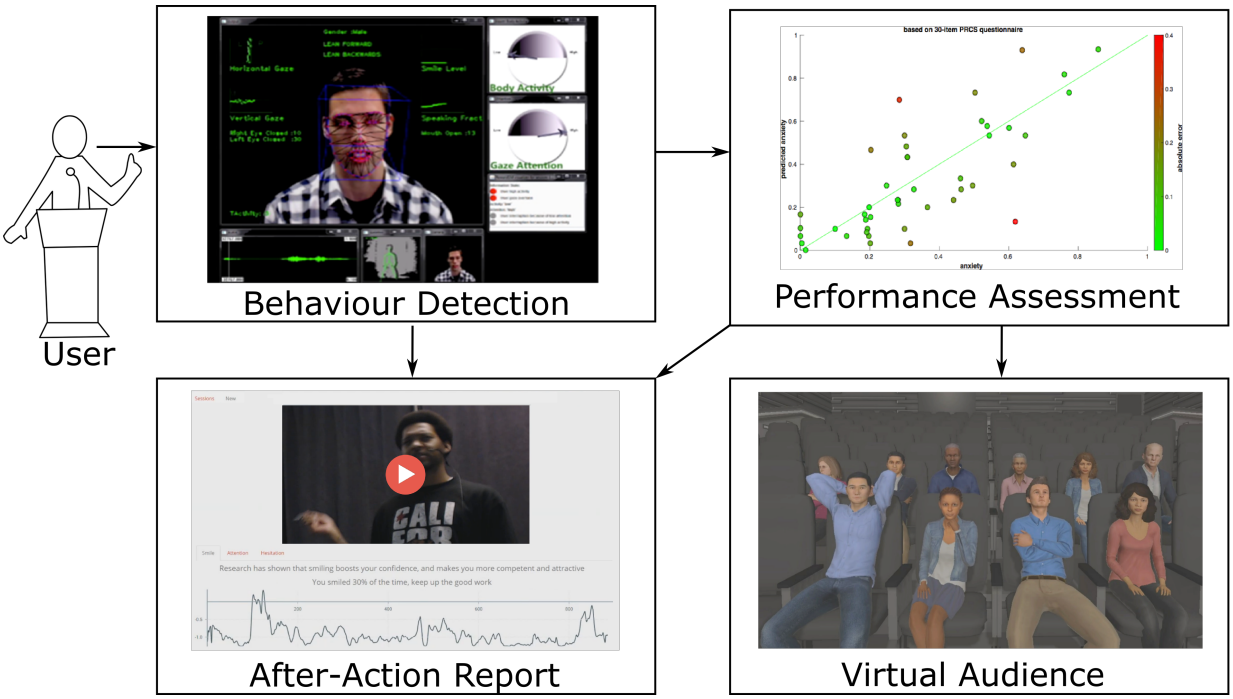

Fig. 1 General architecture of our public speaking training system.

3.1 Behavior Detection

The Multisense module is a behavior detection module part of the VHToolkit. Based on the Social Signal Interpretation (SSI) framework [55], it combines different specialised submodules in order to detect multimodal behaviors in real-time, such as facial expressions, head gaze, or postures. In our case, the head gaze and facial expressions module was based on the OpenFace framework [3]. Multisense outputs its detection results in the Perception Markup Language (PML) format [49] periodically over the network, in our case once per second. Once the interaction is over, Multisense saves the recorded audio and video files of the presentations which are then used when generating the after-action report.

3.2 Performance Assessment

The *Performance Assessment* module is used to regularly compute a public speaking performance score based on the behavioral inputs received from Multisense. We created an *ad-hoc* module which computes a performance score every 5 seconds using the Multisense perception data in the PML format. The task of evaluating a public speaking performance is inherently subjective and multimodal, as it is necessary to combine information from a presenter's contextual speech, their voice, non-verbal behaviors such as gaze and posture, as well as their mastery of the subject. In [57, 46, 43], these authors created multimodal public speaking performance assessment models; however, these models incorporate information from many behavioral signals, and it would be difficult for trainees to simultaneously keep a track of all of them while training. Therefore, we decided to simplify the problem by focusing on specific qualities of a presentation, namely *eye contact* and *facial expressions* only. The choice of these two cues was based on technical and experimental factors as well as discussions with public speaking coaches. During intial project phases, we organised meeting with experts from the international Toastmasters public speaking training association to analyze public speaking training practices and exercices. From those interviews, we derived a variety of cues seen by experts as indicative of good public speaking performance, such as eye contact, varied vocal expressivity, absence of disfluencies, appropriate gestures, positive expressions, and more. When considering which and how many features to consider, we opted for a simplified scheme with a reduced amount of cues. While such a model would not account for all the complexity of public speaking, it would hold the experimental benefit of being easy to understand and to use as a reference for a public speaking training exercise. Additionally, including too many feedback elements in the experiment might have introduced confouding elements, including different levels of complexity of feedback, or varied confidence of the recognition algorithms

used. Finally, when considering which signals to integrate, eye contact and facial expressions were selected since sofware modules for facial expressions and head gaze recognition had already been integrated in other systems of our laboratory, which enabled for rapid deployment of the experiment.

Consequently, we designed an automated behavior assessment tool based on eye contact and facial expression signals. This tool relies on the estimation of two variables, based on the participant's head gaze direction and facial action units detected by the OpenFace module, which is competitive in head gaze estimation and facial action units detection [3]. Specifically, the system updates the trainee's performance score every 5 seconds: if the participant looked at the audience more than 80% of the time (condition 1), and if the participant showed a positive expression (defined by an activation of AU6 or AU12) more than 20% of the time and did not show a negative expression (defined by an activation of AU4 or AU15) more than 20% of the time (condition 2), then the performance of the participant is rated as positive. If only one of the condition is met, the performance is neutral, and if both conditions are not met, it is negative. Those values were determined through preliminary tests, where we asked several team members to act out public speaking performances with high (resp. low) eye contact and positive facial expressions. From those examples, the eye contact and facial expression value thresholds were determined as appropriate compromise between simplicity and ability to separate the positive and negative examples. During runtime, the variables for audience eye contact and facial action units are obtained directly from the PML data stream outputted from Multisense.

3.3 Virtual Audience

We used the *Cicero* interactive virtual audience system that was introduced in [17] and evaluted in a series of studies [13,16]. This virtual audience contains an audience director model which controls how to express different levels of arousal and valence through audience non-verbal behaviors such as head movements, postures, gaze patterns, and facial expressions. This model was evaluated through a series of crowdsourcing studies, in which participating subjects created variations of audience members' behaviors, which were then further used to generate a large number of videos of virtual audiences with varying behavior. The resulting videos of this dataset were then evaluated by other subjects for the overall perceived arousal and valence of the virtual audience. The results showed that the behavior of the audience could be adjusted to display incremental changes in arousal and valence that would be perceived by users. In our case, when a new performance score is computed by the *Performance Assessment* module, this score further influences the audience feedback behavior. If the performance is positive, then the audience's behavior is modified to appear slightly more positive and engaged. If the performance is neutral, the audience stays in the same state. Finally, when the performance is negative, the audience is modified to appear slightly more negative and less engaged.

The feedback produced by the virtual audience on the user's performance is of a very different nature compared to the feedback systems presented in 2.2, which involved different forms such as graphical or textual feedback, and different temporalities from notifications [20,52] to corrective interruptions [50] and after-action reports [53]. By using implicit, social cues, we anticipate that this form of feedback will not distract users from their performance - as opposed to some forms of graphical feedback which can take attentional and cognitive resources [17] or require interruptions [50]. Since users were able in previous studies to perceive the general audience impression conveyed through its socio-emotional cues [13], we assume that they should be able to implicitly infer the general quality of their performance, even if subconsciously. However, this assessment will be imprecise and is not likely to enable users to draw precise conclusions on how to adapt their behavior; they might however perceive the experience as more emotionally engaging, and the audience social cues could influence participants' behavior through subtle reinforcement of good performances and inhibition of poor performances.

3.4 After-Action Report

Finally, we designed a web-based *After-Action Report* interface. Whenever a presentation is over, the application retrieves the presentation's audio and video files. A transcription of the presentation is obtained automatically through IBM Watson's Speech-To-Text service[1]. When the transcript is received, the web page displaying the presentation's video along with feedback is presented to the trainee. We created two alternatives of the after-action report, a personalized and a non-personalized version; in the personalized version, graphs illustrating the trainees' eye contact and facial expressions are displayed along with personalized hints which vary depending on the behaviors of the users (*e.g. "you looked at the audience* 63% *of the time: you are*

[1] https://www.ibm.com/watson/services/speech-to-text/

doing great but you can still improve!"). In the non-personalized version, general hints are provided along with the presenter's video but no graphs or personalized feedback is presented.

4 Experimental study

The current study utilized a between-subjects design to assess differences across social skills training components in terms of training outcomes as well as user experiences. Specifically, we looked at whether real-time feedback and personalized after-action feedback contributed to positive training outcomes and user experiences. Additionally, we investigated whether there was any influence of participants' personality, immersive tendencies, and prior level of public speaking anxiety on participants' experiences and perceptions of the system as a whole. Our research questions were the following:

Q1 Does real-time audience behavioral feedback contribute to positive training outcomes and user experiences?
Q2 Does providing personalized delayed feedback, here in the form of an after-action report, contibute to positive training outcomes and user experiences?
Q3 Are there individual factors which lead to differences in user experiences and training effectiveness?

4.1 Protocol

A between-subjects design was used for the current study. Participants were randomly assigned to 4 conditions in which the interactivity of the audience and the level of personalization of the after-action report were manipulated. The audience could be configured to provide real-time feedback to the user or to be passive, while the after-action report could be personalized, non-personalized, or absent. The conditions and the associated configurations of the audience and the after-action report are described in Table 1. The study room itself was outfitted with a large (60") screen on which the audience was displayed. Over this screen, a video camera was attached which captured the participants' presentations. Additionally, participants were equipped with a clip-on microphone to capture their speech.

The participants were asked to give 3 short presentations (about 2 to 3 minutes) on pre-determined topics. Before each one, the participants were given up to 5 minutes to prepare for their presentations. The topic for the first (pre-test) and last (post-test) presentations, was to give a description of Los Angeles, the city in which the study took place. The topic was the same in both conditions to allow for future comparisons of participants' behavior before and after training. For the middle (training) presentation, participants were instructed to briefly speak about themselves, as if they were introducing themselves in a job interview. Additionally, the participants had been briefed about the presentation tasks and topics prior to them coming to the lab, allowing them to prepare speech in advance if they desired; however, no participants reported having prepared a presentation before the experiment. The training presentations were followed by an interaction with the after-action report tool, configured according to the participant's assigned condition. The feedback was designed to provide hints on the public speaking performance of the individual. More specifically, the participant received advice relating to eye contact, flow of speech and pause fillers, and facial expressions. The generic advice was obtained from publicly available resources from public speaking training association websites (Toastmasters). In the personalized after-action report condition, graphs and quantifications of the participants' behavior were included with this advice.

4.2 Questionnaires

The following questionnaires were administered before and after the task to gain an understanding of stable predispositional traits of the participant, prior to their interaction with our system and participants' overall experience after interacting with our system.

4.2.1 Public Report of Confidence as a Speaker

The Public Report of Confidence as a Speaker (PRCS) is a questionnaire administered to assess fear of public speaking [41]. Several other instruments have been developped which are relevant to assess public speaking anxiety, such as the Personal Report of Communication Apprehension [39] or the Self Statements during Public Speaking [31], or social anxiety more broadly, such as the Liebowitz Social Anxiety Scale [36] used in [48]. We chose the PRCS as it had been utilized in previous studies featuring virtual audiences as a backdrop for public speaking tasks [42,6,17], . It consists of 30 yes-no questions, such as "I get anxious when I think about a speech coming up". We used this questionnaire to obtain a measure of public speaking anxiety in the form of a score ranging from 0 to 30.

4.2.2 Brief Big Five Personality Inventory

The Big 5 Inventory [47] is a widely used instrument to assess an individual along five factors of personality

Condition	Audience	Report		n (Genders)	Ages	PRCS
P-PA	**Passive**	*Personalized*	**Q1**	14 ($7M, 7F$)	45.9 ($SD = 14.1$)	9.7 ($SD = 8.5$)
P	**Interactive**	**Personalized**		14 ($9M, 5F$)	45.8 ($SD = 13.5$)	11.1 ($SD = 8.2$)
NP	*Interactive*	**Generic**	**Q2**	13 ($6M, 7F$)	40.8 ($SD = 14.3$)	8.3 ($SD = 6.2$)
NoR	*Interactive*	**None**		16 ($9M, 7F$)	40.5 ($SD = 13.1$)	7.8 ($SD = 5.4$)
All				57 ($31M, 26F$)	43.2 ($SD = 13.6$)	9.2 ($SD = 7.1$)

Table 1 Overview of the 4 conditions of our experiment and the associated configurations of the system, with participant distributions in gender, age, and PRCS public speaking anxiety score. ***P-PA***: ***P***ersonalized after-action report, ***P***assive ***A***udience. ***P***: ***P***ersonalized after-action report, interactive audience. ***NP***: ***N***on-***P***ersonalized after-action report, interactive audience. ***NoR***: ***No*** after-action ***R***eport, interactive audience.

(e.g., openness, conscientiousness, extraversion, agreeableness, and neuroticism) [47]. This 10-item version uses a five point Likert scale with anchors at 1, "Disagree Strongly" to 5, "Agree Strongly". While the Big Five model is the dominant model of personality in the literature, several alternatives exist for which a review can be found here [22].

4.2.3 Immersive Tendencies Questionnaire

The Immersive Tendencies Questionnaire (ITQ, [56]) is a commonly used scale in the presence literature and was used to identify an individual's propensity to become immersed in simulations. This instrument consists of 32 items which break down into subscales that measure tendencies regarding involvement, focus, and gaming. Along the ITQ, Witmer and Singer present a Presence Questionnaire; while the latter has been criticized for its subjective nature, the ITQ is commonly used in studies to extract important individual factors in studies related to immersive systems [51,37,7].

4.2.4 Immersive Experience Questionnaire

In order to evaluate participants' immersion in the training simulation as well as their evaluation of the task, we used the immersion questionnaire published by Jennett *et al.* [34]. The questionnaire consists of 31 items consisting of scales from 1 to 5. These items correspond to the following dimensions: basic attention (4 questions), temporal dissociation (6 questions), transportation (6 questions), challenge (6 questions), emotional involvement (5 questions) and enjoyment (4 questions). Four questions that did not fit the particular experience of using our system were excluded. There are many alternatives to Jennett's questionnaire especially for determining feelings of presence, such as Witmer and Signer's presence questionnaire (presented alongside the ITQ presented above) [56], the Temple Presence Inventory [38]. However, the nature of the system that we evaluate here make Jennett's questionnaire's an ideal match for this study. Indeed, we are not only interested in the subjective experience of presence of our study participants, but also by how much they generally become engaged and involved with the training tasks afforded by the system.

4.2.5 Attrakdiff Questionnaire

AttrakDiff is a questionnaire that measures user perceptions of the *hedonic* and *pragmatic* qualities of software [30]. The pragmatic quality of a product related to its general usability and perceived effectiveness to achieve its proclaimed goals. The hedonic quality is further divided into two sub dimensions, *hedonic - identity* and *hedonic - stimulation*. The hedonic - stimulation dimension relates to how much users feel a software is motivating or stimulating, while the hedonic - identity dimension measures users' ability to express themselves through a product. AttrakDiff contains three sets of seven scales consisting of with opposite adjective pairs that measure user perceptions of the software, such as "Simple - Complicated" or "Novel - Ordinary". Each of these scales consists of a 7-point Likert scale with values ranging from -3 to 3 where zero represents a neutral point between the opposite adjectives.

4.2.6 Additional Questions

Finally, we included two questions pertaining to how participants perceived the system to be useful for helping train public speaking. These questions allow us to assess participants' evaluations of the task-specific potential of our system on training public speaking. The questions were presented with 5-point Likert scales with the anchors "1, Strongly disagree" to "1, Strongly agree", and are reported below:

- "Would you train with this tool to improve your public speaking skills, if given the chance?"
- "How useful do you think this tool can help you with improving your public speaking skills?"

4.3 Participants

We recruited 63 participants for this study. Out of this total number, 6 sessions had to be removed because of

technical issues (failed recording, missing audio-visual frames, missing questionnaire data). The participants had an average age of 43.2 ($SD = 13.6$). A breakdown of the age, gender, and PRCS distributions across conditions is also presented in Table 1.

4.4 Training Outcomes Assessment

In order to evaluate the quality of the participants' performances, we ran an online evaluation study using the Amazon Mechanical Turk crowdsourcing platform. The workers were instructed to watch a randomly chosen participant video (from the pre-test and post-test videos) and to assess the presenters' performance quality on 10 aspects of public speaking performance. These 10 aspects were chosen through discussion with public speaking experts and used in related work [6]; they included categories related to the speakers' bodily behavior (*e.g.* body posture, eye contact), vocal behavior (*e.g.* flow of speech, pause filler avoidance) as well as higher order variables (*e.g.* confidence). These aspects were rated using 7-point Likert scales. Each video was rated by four different annotators, and it was ensured that no annotator would rate both the pre-test and post-test video for a single participant. The annotators were not forced to rate all available videos; in practice, a minority of workers rated many videos (8 workers rated 28 videos or more, accounting for nearly 60% of the total ratings), while a majority annotated just a handful of videos.

In order to assess the reliability of the crowdsourced annotations, we computed Intra-Class Correlations. Indeed, the public speaking aspects ratings data was ordinal in nature, and every video was judged by a different set of raters, pooled randomly from a larger population of raters. Different variants of the ICC exist depending on the particularities of a dataset; in our case, we use a one-way, absolute, average-measures ICC following the guidelines from [28]. We employ the *irr* package of the *R* statistical framework to compute the ICCs for each public speaking aspect. Using Cicchetti's cutoff values as a reference [18], the resulting ICCs ranged from fair (Stage usage: $ICC = 0.42$, Speech intonation: $ICC = 0.48$, Pause fillers: $ICC = 0.55$, Flow of speech: $ICC = 0.58$, Presentation structure: $ICC = 0.59$), to good (Body posture: $ICC = 0.61$, Eye contact: $ICC = 0.63$, Overall performance: $ICC = 0.65$, Confidence: $ICC = 0.66$), and excellent (Gesture usage: $ICC = 0.78$).

4.5 Results

In this section, we present the results of statistical tests that were undertaken to answer research questions *Q1*, *Q2* and *Q3*. For *Q1* and *Q2*, Student t-tests were realised to test for differences between conditions. For *Q3*, we calculated Pearson's correlation coefficients between individual variables of interest and response measures.

Q1 - Real-time audience feedback - First, we investigated if real-time audience behavioral feedback benefited participants in terms of training outcomes. We compared the training outcomes between the passive audience (*P-PA*) and interactive audience (P) condition by comparing the difference between the average ratings of the 10 aspects for the post-training presentation video and for the pre-training videos. We did not find any significant difference across all 10 aspects. Results are illustrated in Figure 2.

Second, we compared participants' user experiences between the passive audience and interactive audience condition. While we did not find any significant difference in terms of training outcomes, we found that users receiving real-time feedback from the interactive audience rated their experience to be more immersive ($p < 0.0005$), more enjoyable ($p = 0.049$), and that were more likely to indicate they would use the system for training if they had the chance ($p = 0.033$). However, we did not find any significant difference in terms of how the conditions were rated on AttrakDiff questionnaires dimensions of pragmatic ($p = 0.295$) and hedonistic (identity: $p = 0.439$; stimulation: $p = 0.272$) quality.

Q2 - Delayed feedback personalization - With our second research question, we compared the effects of delayed feedback provided with the after-action report component of our system. In particular, we investigated whether personalizing feedback contributes to training outcomes or whether providing trainees with a recording of their performance is sufficient. To this end, we compared training outcomes on the 10 annotated performance aspects between the personalized after-action feedback condition (P), the non-personalized after-action feedback condition (NP), and the condition without after-action report (NoR). Results are presented in Figure 3. Significant differences in training outcomes were observed between the personalized feedback (P) and non-personalized feedback (NP) condition for the *Confidence* ($p = 0.047$), *Eye Contact* ($p = 0.031$), and *Presentation structure* ($p = 0.028$), while trends for also observed for *Flow of Speech* ($p = 0.072$) and *Overall Performance* ($p = 0.066$). Additionally, significant differences between the personalized feedback (P) condition and the condition without after-action report

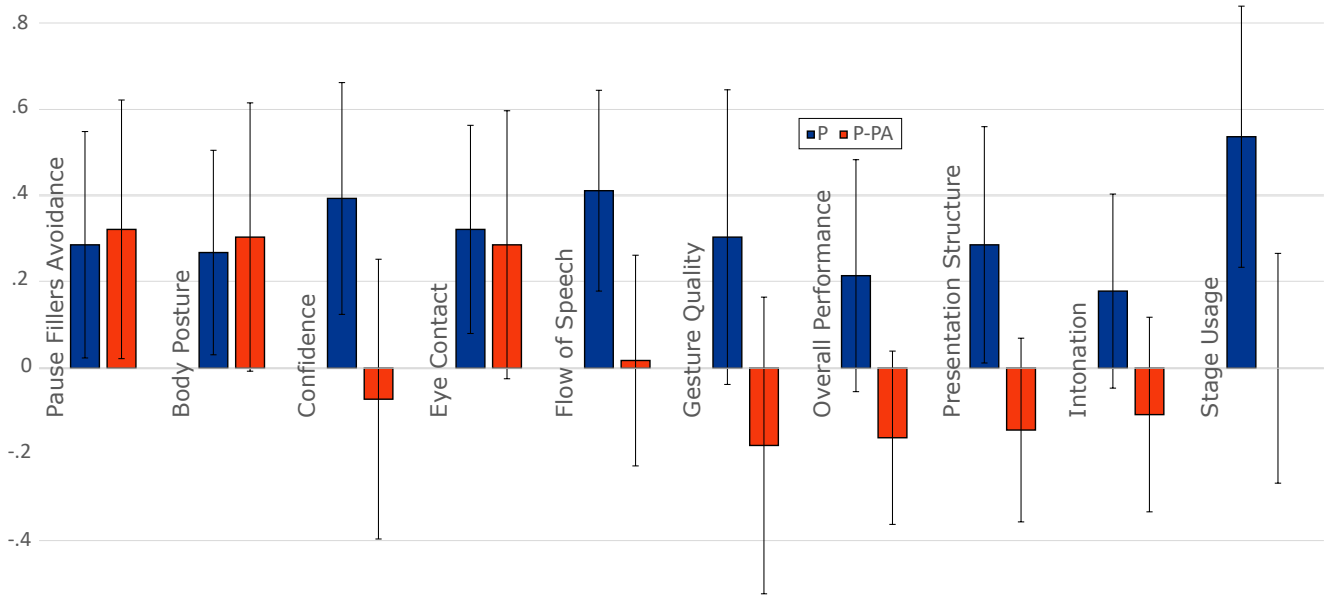

Fig. 2 Comparison of training outcomes between the interactive audience (***P***, in dark blue) and non-interactive audience (***P-PA***, in red) real-time feedback conditions.

(*NoR*) were found for *Flow of Speech* ($p = 0.010$) and a trend was observed for *Presentation structure* ($p = 0.079$) In comparison, we did not observe any significant differences in training outcomes between the non-personalized feedback (*NP*) condition and the no feedback condition (*NoR*).

We also compared participants' responses to the AttrakDiff and presence questionnaires in order to assess whether the different after-action report configurations led to any differences in user experience. No significant difference was observed, both between the personalized report (*P*) and non-personalized (*NP*) report, as well as between the personalized (*P*) report and the absent report (*NoR*) condition.

Q3 - Individual Differences - We ran Pearson's correlation statistical tests in order to test for associations between variables describing individual traits and output variables. We investigated the impact of initial public speaking anxiety, personality, and immersive tendencies.

Public speaking anxiety - There was no effect of the PRCS public speaking anxiety score on user evaluations of the quality of the system. However, the PRCS scores were found to be highly correlated with perceived challenge ($\rho = .51, p < 0.001$) and negatively correlated with enjoyment of the task ($\rho = -.27, p = 0.0422$).

Personality - When computing correlations between measures of the five OCEAN personality dimensions and dependent variables, we found an effect of neuroticism ($\rho = .41, p = 0.0016$) as well as extraversion ($\rho = -0.48, p < 0.001$) on perceived challenge.

Immersive tendencies - The total score obtained with the immersion experience questionnaire was highly correlated with the participants' immersive tendency scores ($\rho = .53, p = 0.001$). Looking at the underlying factors of the immersion questionnaire, we found that participants' immersive tendency was particularly correlated with transportation ($\rho = .49, p = 0.003$) and emotional involvement ($\rho = .39, p = 0.021$).

5 Discussion

In this experimental study, we set out to compare the relative contributions to training outcomes and user experiences of two components of a public speaking training application featuring virtual humans and social signals processing technologies. Specifically, we assessed the impact of real-time feedback provided by virtual humans, and the impact of an after-action report component providing personalized feedback. These elements are common features of many social skills training applications [2,32,17], however their respective impact on the system effectiveness is usually not compared.

With *Q1*, we compared a condition involving a interactive audience providing real-time feedback reflecting the user's performance to a passive audience only displaying minimal behaviors and not providing any feedback. We did not observe any effect on the training outcomes of providing real-time feedback in behavioral form. This seems to indicate that real-time implicit feedback is not particularly useful in terms of improving the public speaking behavior of trainees. It is

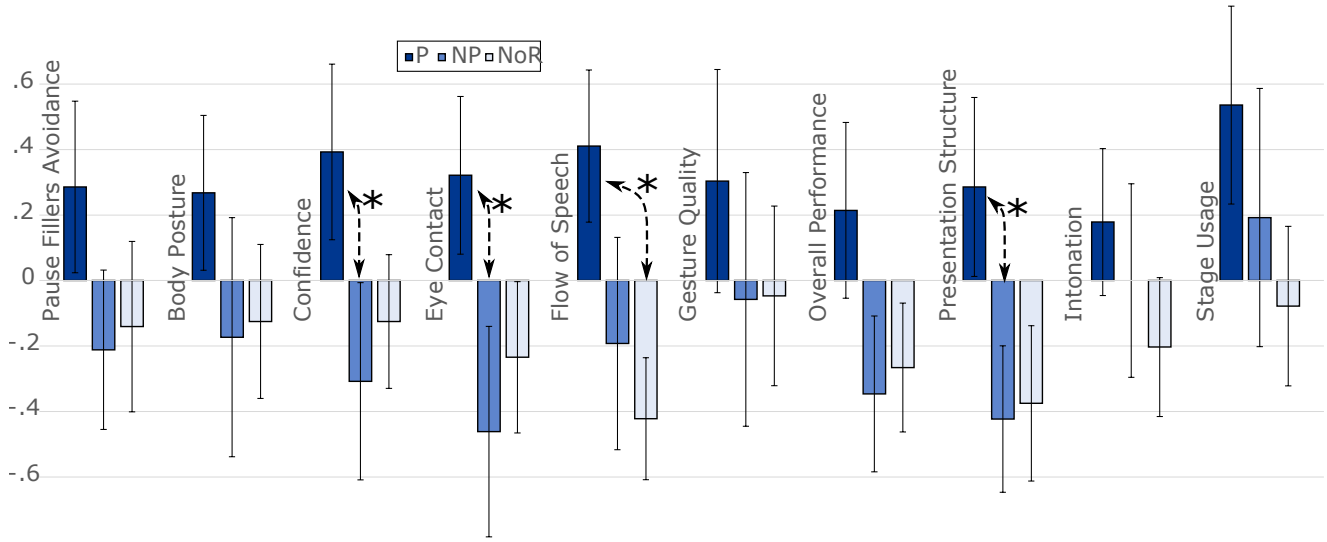

Fig. 3 Comparison of training outcomes between the personalized (***P***, in dark blue), non-personalized (***NP***, in medium blue), and absent (***NoR***, in light blue) After-Action Report conditions. Significant differences at the $p < 0.05$ level are highlighted with a $*$ sign.

important to note that the feedback was here designed to reflect the trainee's current behavior performance as defined by simple models of public speaking assessment; it is indeed possible that different forms of behavioral feedback (*e.g.* vocal feedback) could be more beneficial in terms of training outcomes. In parallel, we observed that the interactive audience contributed positively to the user experience in terms of immersion, motivation and enjoyment of the experience.

In contrast, with *Q2*, we compared the impact of using an after-action report including personalized feedback or generic feedback. We observed that a personalized after-action report tool contributed to improving training outcomes. This seems to indicate that public speaking skills training benefits from delayed feedback, but providing the trainees with recordings of their presentations is insufficient and that it is necessary to provide some guidance to the user in order for them to analyze and reflect on their own behavior. However, the overall user experience was unaffected by the addition of this component.

Taken together, these two results suggest that both the interactive, real-time audience feedback component and the after-action delayed feedback components of a social skills training system provide benefits to the trainees, although in two very different dimensions. The implicit feedback provided by virtual humans as part of a simulated interaction does not seem to directly benefit training, however it creates a more motivating experience and leads users to declare they would be more likely to engage with such a system for training. This suggests that including a form of interactivity during the training task could improve the user experience, potentially leading to higher training motivation and more training over time, which could translate to stronger training outcomes. On the other hand, delayed personalized feedback directly impacted training outcomes, as judged by third party annotators, even on the very short training session of our study. However, this did not seem to translate into an improved user experience.

Finally, we investigated with *Q3* whether some individual factors impacted the user experience of training with such a social skills training system. As expected, we found an effect of individual differences on user's experience and on participants' evaluations of the system. Our first finding regarding participants' public speaking anxiety score, obtained through the PRCS questionnaire, influenced the user experience. Specifically, PRCS scores were highly correlated with perceived challenge and negatively correlated with enjoyment of the task, meaning that highly anxious individuals found the task to be more difficult and less enjoyable. A similar result was elicited when comparing neuroticism, a personality trait related to emotional instability, and perceived task challenge. While these findings are unsurprising, they are noteworthy because they concern the population of speakers that stand to benefit the most from training [17,48]; as such, this highlights the necessity to tailor simulated social interactions to individuals who suffer from higher levels of social anxiety, especially for the first interactions with such a system, in order to foster higher motivation and thus higher likelihood of usage of such systems by these participants. We also observed an effect of participants' immersive tendency scores on several

output measures. Unsurprisingly, the immersive tendencies score was highly correlated with participants' total immersion scores - more interestingly, we found that participants with higher tendencies to immersion tended to report a stronger sense of feeling transported inside the simulation and stronger emotional involvement. This was further confirmed by the finding that participants with higher tendency to immersion rated the software as more stimulating than other participants. In particular, the stronger emotional response of participants with higher tendency to immersion seems noteworthy for applications related to training social skills; indeed, subjects' appraisals of evaluative social situations (such as a public speaking situation) have a strong influence on their behavior, performance and internal states [8]. Therefore, participants who score higher on the immersive tendency scale might be more receptive to strong socio-emotional stimuli displayed by virtual agents, which in turn might influence the overall training experience and its training outcomes.

The results of this work, however, should be interpreted in light of several limitations. First, the interpretation of the results was realised by third-party annotators recruited on crowdsourcing platforms. The subject of crowdsourcing reliability is a very active field of research and while certain challenges and pitfalls come with it, it is not necessarily unreliable [11]. In our particular context, the opinion of non-specialists is of interest since convincing crowds is often the goal of public speaking. However, domain specialists, such as public speaking coaches, therapists or specialized researchers could provide more objective judgments, in particular when considering finer aspects of public speaking behavior. More generally, we used the ratings realised by annotators observing performances after their realisation and outside of their social and spatial context. However, several studies have shown that the context in which individuals observe actions has major implications on their judgments [23, 44]. For instance, observing a public speaker from a first-person perspective (being addressed by the speaker) or a third-person perspective (watching a video of the speaker addressing someone else) has implications on the cognitive mechanisms employed when analyzing the performance. In our context, it would be relevant to compare the impressions of observers witnessing a performance first-hand as members of the audience, to observers juding a recording of the performance online.

Another drawback concerns the experimental design and particularly the set of conditions. It would have been ideal to conduct a full study with all possible combinations of the passive or interactive audience and the personalized, non-personalized, or absent after action-report. Additionally, a control condition without any feedback, or a control condition with another traditional form of training (*e.g.* watching instructional videos on public speaking), could have been considered to provide a baseline. While we considered having such an the experimental protocol, the overall scope of the study unfortunately had to be reduced due to constraints in terms of time and resources. Finally, another important and related limitation concerns the possibility of Type I errors in the statistical tests comparing the experimental conditions. Indeed, the low amount of participants in relationship to the amount of conditions combined with the number of variables tested highly increases the likelihood of Type 1 error in the tests presented above. Therefore, it is essential not to consider the statistical tests of this study as definite arguments in favor or against a particular question, but as one data point to consider in light of the broader and growing body of knowledge encompassing other similar studies.

6 Conclusion

Social signal processing and virtual social interaction technologies have allowed to create applications which stand to provide many benefits for training and assessing social skills or for helping manage social anxieties. Initial studies have shown that such solutions can lead to positive training outcomes and could complement traditional teaching methods by providing cheap, accessible, safe tools for training social skills. These studies usually evaluated social skills training systems as a whole and it is unclear to what extent which components contributed to positive training outcomes and user experiences, however this knowledge is necessary to properly understand how to create the most efficient training systems.

We conducted an experimental study where we compared the relative efficacy of real-time interactive feedback and after-action feedback in the context of public speaking training. We observed that both of these components provide benefits to the overall training: the real-time interaction led to a more immersive experience and improved participants' motivation in using the system, while the after-action feedback led to positive training outcomes when it contained personalized feedback elements. Taken in combination, these results confirm that both social signal processing technologies and virtual social interactions can be useful to create social skills training systems. Additionally, we observed that some individual factors, here the subjects' initial level of public speaking anxiety, personality, and immersive

tendencies, also significantly influenced the training experience. This finding suggests that social skills training systems and experiences could benefit from being tailored to participants' particular individual circumstances.

References

1. Ali, M.R., Crasta, D., Jin, L., Baretto, A., Pachter, J., Rogge, R.D., Hoque, M.E.: LISSA - Live Interactive Social Skill Assistance. In: Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, pp. 173–179. IEEE (2015)
2. Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssafidou, E., Damian, I., Ennis, C., Egges, A., Gebhard, P., et al.: The tardis framework: intelligent virtual agents for social coaching in job interviews. In: International Conference on Advances in Computer Entertainment Technology, pp. 476–491. Springer (2013)
3. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 59–66. IEEE (2018)
4. Bandura, A.: Self-efficacy: The exercise of control. Macmillan (1997)
5. Barmaki, R., Hughes, C.E.: Embodiment analytics of practicing teachers in a virtual immersive environment. Journal of Computer Assisted Learning (2018)
6. Batrinca, L., Stratou, G., Shapiro, A., Morency, L.P., Scherer, S.: Cicero - towards a multimodal virtual audience platform for public speaking training. In: Intelligent Virtual Agents, p. 116–128 (2013)
7. Bissonnette, J., Dubé, F., Provencher, M.D., Moreno Sala, M.T.: Evolution of music performance anxiety and quality of performance during virtual reality exposure training. Virtual Reality **20**(1), 71–81 (2016). DOI 10.1007/s10055-016-0283-y
8. Blascovich, J., Mendes, W.: Challenge and threat appraisals: The role of affective cues. in j. forgas (ed.), feeling and thinking: The role of affect in social cognition (pp. 59–82) (2000)
9. Bower, M., Cavanagh, M., Moloney, R., Dao, M.: Developing communication competence using an online video reflection system: Pre-service teachers' experiences. Asia-Pacific Journal of Teacher Education **39**(4), 311–326 (2011)
10. Brown, T., Morrissey, L.: The effectiveness of verbal self-guidance as a transfer of training intervention: its impact on presentation performance, self efficacy and anxiety. Innovations in Education and Teaching International **41**(3), 255–271 (2004)
11. Casler, K., Bickel, L., Hackett, E.: Separate but equal? a comparison of participants and data gathered via amazon's mturk, social media, and face-to-face behavioral testing. Computers in human behavior **29**(6), 2156–2160 (2013)
12. Chan, V.: Teaching oral communication in undergraduate science: Are we doing enough and doing it right?. Journal of learning design **4**(3), 71–79 (2011)
13. Chollet, M., Chandrashekhar, N., Shapiro, A., Morency, L.P., Scherer, S.: Manipulating the perception of virtual audiences using crowdsourced behaviors. In: International Conference on Intelligent Virtual Agents, pp. 164–174. Springer (2016)
14. Chollet, M., Prendinger, H., Scherer, S.: Native vs. non-native language fluency implications on multimodal interaction for interpersonal skills training. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 386–393. ACM (2016)
15. Chollet, M., Scherer, S.: Assessing public speaking ability from thin slices of behavior. In: to appear in Proceedings of the IEEE International Conference on Face and Gesture Recognition (2017)
16. Chollet, M., Scherer, S.: Perception of virtual audiences. IEEE computer graphics and applications **37**(4), 50–59 (2017)
17. Chollet, M., Wörtwein, T., Morency, L.P., Shapiro, A., Scherer, S.: Exploring Feedback Strategies to Improve Public Speaking: An Interactive Virtual Audience Framework. In: Proceedings of UbiComp 2015. ACM, Osaka, Japan (2015)
18. Cicchetti, D.V.: Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological assessment **6**(4), 284 (1994)
19. Damian, I., Baur, T., Lugrin, B., Gebhard, P., Mehlmann, G., André, E.: Games are better than books: in-situ comparison of an interactive job interview game with conventional training. In: International Conference on Artificial Intelligence in Education, pp. 84–94. Springer (2015)
20. Damian, I., Tan, C.S.S., Baur, T., Schöning, J., Luyten, K., André, E.: Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, pp. 565–574. ACM, New York, NY, USA (2015). DOI 10.1145/2702123.2702314
21. De Grez, L., Valcke, M., Roozen, I.: The impact of goal orientation, self-reflection and personal characteristics on the acquisition of oral presentation skills. European Journal of Psychology of Education **24**(3), 293 (2009)
22. Feher, A., Vernon, P.A.: Looking beyond the big five: A selective review of alternatives to the big five model of personality. Personality and Individual Differences p. 110002 (2020)
23. Flanagan, J.R., Johansson, R.S.: Action plans used in action observation. Nature **424**(6950), 769–771 (2003)
24. Fung, M., Jin, Y., Zhao, R., Hoque, M.E.: Roc speak: semi-automated personalized feedback on nonverbal behavior from recorded videos. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1167–1178. ACM (2015)
25. Furmark, T., Tillfors, M., Everz, P.O., Marteinsdottir, I., Gefvert, O., Fredrikson, M.: Social phobia in the general population: prevalence and sociodemographic profile. Social psychiatry and psychiatric epidemiology **34**(8), 416–424 (1999)
26. van Ginkel, S., Gulikers, J., Biemans, H., Mulder, M.: Towards a set of design principles for developing oral presentation competence: A synthesis of research in higher education. Educational Research Review **14**, 62–80 (2015)
27. van Ginkel, S., Gulikers, J., Biemans, H., Noroozi, O., Roozen, M., Bos, T., van Tilborg, R., van Halteren, M., Mulder, M.: Fostering oral presentation competence through a virtual reality-based task for delivering feedback. Computers & Education **134**, 78–97 (2019)
28. Hallgren, K.A.: Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in quantitative methods for psychology **8**(1), 23 (2012)

29. Harris, S.R., Kemmerling, R.L., North, M.M.: Brief virtual reality therapy for public speaking anxiety. Cyberpsychology and Behavior **5**, 543–550 (2002)
30. Hassenzahl, M., Burmester, M., Koller, F.: Attrakdiff: A questionnaire to measure perceived hedonic and pragmatic quality. In: Mensch & Computer, pp. 187–196 (2003)
31. Hofmann, S.G., DiBartolo, P.M.: An instrument to assess self-statements during public speaking: Scale development and preliminary psychometric properties. Behavior Therapy **31**(3), 499–515 (2000)
32. Hoque, M.E., Courgeon, M., Martin, J.C., Mutlu, B., Picard, R.W.: Mach: My automated conversation coach. In: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, pp. 697–706. ACM (2013)
33. Hoque, M.E., Picard, R.W.: Rich nonverbal sensing technology for automated social skills training. Computer **47**(4), 28–35 (2014)
34. Jennett, C., Cox, A.L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., Walton, A.: Measuring and defining the experience of immersion in games. International journal of human-computer studies **66**(9), 641–661 (2008)
35. Kerby, D., Romine, J.: Develop oral presentation skills through accounting curriculum design and course-embedded assessment. Journal of Education for Business **85**(3), 172–179 (2009)
36. Liebowitz, M.: Liebowitz social anxiety scale. Modern problems of pharmacopsychiatry **22**, 141–173 (1987)
37. Ling, Y., Nefs, H.T., Brinkman, W.P., Qu, C., Heynderickx, I.: The relationship between individual characteristics and experienced presence. Computers in Human Behavior **29**(4), 1519 – 1530 (2013). DOI https://doi.org/10.1016/j.chb.2012.12.010
38. Lombard, M., Ditton, T.B., Weinstein, L.: Measuring presence: the temple presence inventory. In: Proceedings of the 12th annual international workshop on presence, pp. 1–15 (2009)
39. McCroskey, J.C.: Validity of the prca as an index of oral communication apprehension. Communications Monographs **45**(3), 192–203 (1978)
40. Mihoub, A., Lefebvre, G.: Wearables and social signal processing for smarter public presentations. ACM Transactions on Interactive Intelligent Systems (TiiS) **9**(2-3), 9 (2019)
41. Paul, G.: Insight vs. Desensitization in Psychotherapy: An Experiment in Anxiety Reduction. Stanford University Press (1966)
42. Pertaub, D.P., Slater, M., Barker, C.: An experiment on public speaking anxiety in response to three different types of virtual audience. Presence: Teleoperators and virtual environments **11**(1), 68–78 (2002). DOI 10.1162/105474602317343668
43. Petukhova, V., Mayer, T., Malchanau, A., Bunt, H.: Virtual debate coach design: Assessing multimodal argumentation performance. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, pp. 41–50. ACM, New York, NY, USA (2017). DOI 10.1145/3136755.3136775
44. Potdevin, D., Sabouret, N., Clavel, C.: Intimacy perception: Does the artificial or human nature of the interlocutor matter? International Journal of Human-Computer Studies **142**, 102464 (2020). DOI https://doi.org/10.1016/j.ijhcs.2020.102464
45. Price, M., Anderson, P.L.: Outcome expectancy as a predictor of treatment response in cognitive behavioral therapy for public speaking fears within social anxiety disorder. Psychotherapy **49**(2), 173 (2012)
46. Ramanarayanan, V., Leong, C.W., Chen, L., Feng, G., Suendermann-Oeft, D.: Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In: Proceedings of the ACM International Conference on Multimodal Interaction, ICMI '15, pp. 23–30. ACM, New York, NY, USA (2015). DOI 10.1145/2818346.2820765
47. Rammstedt, B., John, O.: Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. Journal of research in Personality **41**(1), 203–212 (2007)
48. Safir, M.P., Wallach, H.S., Bar-Zvi, M.: Virtual reality cognitive-behavior therapy for public speaking anxiety: one-year follow-up. Behavior modification **36**(2), 235–246 (2012)
49. Scherer, S., Marsella, S., Stratou, G., Xu, Y., Morbini, F., Egan, A., Morency, L.P., et al.: Perception markup language: Towards a standardized representation of perceived nonverbal behaviors. In: International Conference on Intelligent Virtual Agents, pp. 455–463. Springer (2012)
50. Schneider, J., Börner, D., van Rosmalen, P., Specht, M.: Presentation trainer, your public speaking multimodal coach. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15, pp. 539–546. ACM, New York, NY, USA (2015). DOI 10.1145/2818346.2830603
51. Slater, M.: Measuring presence: A response to the witmer and singer presence questionnaire. Presence **8**(5), 560–565 (1999)
52. Tanveer, M.I., Lin, E., Hoque, M.E.: Rhema: A real-time in-situ intelligent interface to help people with public speaking. In: Proceedings of the 20th International Conference on Intelligent User Interfaces, pp. 286–295. ACM (2015)
53. Tanveer, M.I., Zhao, R., Chen, K., Tiet, Z., Hoque, M.E.: Automanner: An automated interface for making public speakers aware of their mannerisms. In: Proceedings of the 21st International Conference on Intelligent User Interfaces, pp. 385–396. ACM (2016)
54. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and vision computing **27**(12), 1743–1759 (2009)
55. Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., André, E.: The social signal interpretation (ssi) framework: multimodal signal processing and recognition in real-time. In: Proceedings of the 21st ACM international conference on Multimedia, pp. 831–834. ACM (2013)
56. Witmer, B.G., Singer, M.J.: Measuring presence in virtual environments: A presence questionnaire. Presence **7**(3), 225–240 (1998)
57. Wörtwein, T., Chollet, M., Schauerte, B., Morency, L.P., Stiefelhagen, R., Scherer, S.: Multimodal public speaking performance assessment. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 43–50. ACM (2015)
58. Zhao, R., Li, V., Barbosa, H., Ghoshal, G., Hoque, M.E.: Semi-automated 8 collaborative online training module for improving communication skills. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**(2), 32 (2017)