

A Multimodal Positive Computing System for Public Speaking

Evaluating User Responses to Avatar and Video Speaker Representations

Fiona Dermody and Alistair Sutherland
School of Computing, Dublin City University, Dublin, Ireland

Keywords: Real-time Feedback, Multimodal Interfaces, Positive Computing, HCI, Public Speaking.

Abstract: A multimodal Positive Computing system with real-time feedback for public speaking has been developed. The system uses the Microsoft Kinect to detect voice, body pose, facial expressions and gestures. The system is a real-time system, which gives users feedback on their performance while they are rehearsing a speech. We wish to compare two versions of the system. One version displays a live video-stream of the user. The other displays a computer-generated avatar, which represents the user's body movements and facial expressions. Visual feedback is displayed on both versions in proximity to the speaking modality it relates to. In all other aspects, the two versions are identical. We found that users rated the video version of the system more distracting as they focussed on their physical appearance rather than their speaking performance when using it.

1 INTRODUCTION

The prevalent fear of public speaking can impact on a person's success in education or enterprise (Dwyer and Davidson, 2012), (McCroskey et al., 1989), (Harris et al., 2002). In the Positive Computing framework, self-awareness is described in the context of reflection and getting to know oneself. In regard to public speaking, this implies an awareness of how a speaker appears to an audience while speaking. For instance, some speakers may not be aware of the importance of using gestures when speaking to engage an audience (Toastmasters International, 2011). With awareness comes self-knowledge, the power to choose to develop yourself and realise your full potential (Morgan, 2015).

We wish to compare two versions of a system for increasing users' awareness of their public speaking performance. The system is a real-time system which gives users feedback on their performance while they are rehearsing a speech. One version of the system displays a live video-stream of the user. The other displays a computer-generated avatar which represents the user's body movements and facial expressions. In all other aspects, the two versions are identical. Feedback is displayed using visual icons in both versions of the system in proximity to the speaking modality it relates to as shown in Figures 2 and 4. The objective

of the study was to see which version of the system made the users more aware of their performance.

2 RELATED WORK

There are other systems that have focused on awareness in the context of public speaking and communications skills development. AwareMe utilises a wristband that provides speakers with haptic and visual feedback as they are speaking on speaking rate, voice pitch and filler words (Bubel et al., 2016). Cicero:Virtual Audience Framework utilises a virtual audience comprising avatars to convey non-verbal feedback to speakers (Batinca et al., 2013), (Chollet et al., 2015a), (Chollet et al., 2015b). In other systems, the user is represented using an avatar or video stream. A Virtual Rehearsal educational application gives feedback to users in real-time on open and closed gestures using the Microsoft Kinect 2 skeletal view avatar (Barmaki and Hughes, 2015), (Barmaki, 2016). Presentation Trainer represents the user using video and provides them with real-time feedback on one nonverbal speaking modality at a time with a gap of at least six seconds between feedback displays (Schneider et al., 2015). The feedback provided by these systems can make users aware of their speaking behaviour and this awareness can aid in the development of communication skills.

3 POSITIVE COMPUTING

Positive Computing is a paradigm for human-computer interaction, whose objective is to increase the well-being of the user (Calvo and Peters, 2015), (Calvo and Peters, 2014), (Calvo and Peters, 2016). It has a number of themes including competence, self-awareness, stress-reduction and autonomy. In this paper we will concentrate mainly on self-awareness. We will look at some of the other themes in Future Work. If the user is to increase their competence in public speaking, they must first become aware of aspects of their speaking performance. For example, body posture, gestures and gaze direction are all important aspects of public speaking (Toastmasters International, 2011), (Toastmasters International, 2008). Hitherto, the only way to get this awareness was either to practise in front of a human mentor or in front of a mirror. Both of these can cause stress or anxiety for the user. The objective of our system is to allow the user to gain this awareness in private without being exposed to stress or anxiety. The users will get this awareness by looking at a representation of themselves as they speak. They can choose to view themselves as an avatar or a live video stream. Real-time feedback is superimposed on their chosen representation. The research question posed in this paper is, which of these two makes the user more aware of their performance?

4 UTILISING VIDEO

It has been found beneficial in social skills development that individuals observe their own behaviour on video. The video allows the user to see their own body pose, facial expressions, gaze direction and gestures in granular detail. However, not everyone reacts well to seeing themselves on video. 'The cognitive dissonance that can be generated from the discrepancies between the way persons think they come across and the way they see themselves come across can be quite emotionally arousing and, occasionally, quite aversive' (Dowrick and Biggs, 1983). Furthermore, Dowrick and Biggs also note that people may become aware of their nonverbal communication when observing themselves on video and may not be happy with what they observe (Dowrick and Biggs, 1983). Also they become distracted by the physicality of their appearance, their perceived level of physical attractiveness or lack thereof may become their focus, as they observe themselves on video, as opposed to their behaviour. If the person has a negative self-perception of themselves on video, then their reaction to the vi-

deo will not be positive (Dowrick, 1999).

4.1 Video and Public Speaking

Live video stream has been utilised in innovative multimodal systems for public speaking. For instance, Presentation Trainer provides users with real-time feedback using a live video representation of the user (Schneider et al., 2015). The Presentation Trainer uses the Microsoft Kinect 2 to track the users' voice and body. Feedback is presented on one nonverbal speaking modality at a time with a gap of at least six seconds between feedback displays. The feedback is interruptive and directive because it directs the user to stop and adjust their speaking behaviour if they are deemed to have exhibited an undesirable speaking behaviour. The feedback is displayed as text eg. 'reset posture'.

5 UTILISING AN AVATAR

An avatar presents an abstract representation of the user and this form of abstract representation allows the user to see their body pose, gestures and facial expressions in 3D. Given the issues noted using video in Section 4, this form of abstract representation could be advantageous because the user is less likely to be distracted by details of their physical appearance.

5.1 Avatars and Public Speaking

Studies of fear of public speaking have shown that people do respond favourably to virtual agents 'even in the absence of two-way verbal interaction, and despite knowing rationally that they are not real' (Garau, 2006), (Pertaub et al., 2002). Virtual agents have been used effectively in multimodal systems for public speaking, most notably, (Chollet et al., 2015b). In the aforementioned system, virtual agents were used to represent an audience that responded to the user's speaking performance. In the system described in this paper, the avatar represents the user themselves.

6 SYSTEM DESCRIPTION

In this paper we present a multimodal Positive Computing system which gives feedback in real-time on different speaking modalities simultaneously. The term 'multimodal' refers to the fact that the system detects multiple speaking modes in the speaker such as their gestures, voice and eye contact. The user can









Visual Feedback Icon	Meaning
	Agitation
	Arms folded
	Hands joined
	Look left
	Look right
	Happy
	Surprised
	Voice

Figure 1: Visual feedback icons.

select if they want to receive feedback on all speaking modes or a subset of them. The objective of the system is to enable the user to speak freely without being interrupted, distracted or confused by the visual feedback on screen. A more detailed description of the system can be found here (Dermody and Sutherland, 2016). The system consists of a Microsoft Kinect 1 connected to a laptop. The system uses the Microsoft Kinect to sense the user’s body movements, facial expressions and voice. The user stands in front of the system and speaks. Feedback is given on a laptop screen in front of the user.

6.1 System Feedback

Real-time visual feedback is displayed as follows, see Figure 1:

Arrows around the user’s or avatar’s head to prompt the user to change their view direction. A rolling graph is displayed above the user’s head which displays the pitch of the user’s voice. The frequency of the peaks in the graph indicate the speech rate. This allows the user to see how fast they are talking. The slope of the graph indicates rising and falling tones which allows the user to gauge whether they are talking in a monotone voice or using a lot of vocal variety. An icon is displayed to indicate when the user’s hands are touching. The icon is located over the avatar’s hands. An icon is displayed when the user has crossed their arms. An icon is displayed to indicate if the user is agitated or moving too quickly. The icon is located next to the avatar’s body. An icon is displayed to indicate whether the user is smiling or

surprised. The icon is located next to the avatar’s face.

These particular speaking behaviours were chosen because they have been rated as important by experts in public speaking (Toastmasters International, 2011), (Toastmasters International, 2008).

6.2 Avatar and Video Stream

To provide users with more autonomy, a central tenet of positive computing, the system can be configured according to users’ preferences [6], (Calvo and Peters, 2012), (Calvo and Peters, 2014), (Calvo and Peters, 2016), (Calvo et al., 2014). A user can configure the system to use an avatar, see Figure 2, or video stream, see Figure 3.

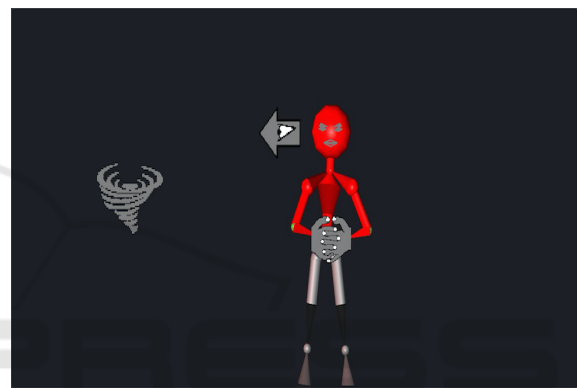


Figure 2: System Avatar with indicative visual feedback on gaze direction, agitation and hands touching. The avatar represents the user.



Figure 3: System Video Stream with visual feedback on gaze direction and hands-touching. The user is represented in the video.

7 STUDY DESIGN

The study had 10 participants (4M, 6F). Participants were drawn from the staff and student body at our university. They came from a range of faculties across

the Humanities, Science, Computing and Business. The study was designed to be a one-time recruitment with a duration of 25 minutes per participant. The participants completed a preliminary questionnaire on demographic information and a post-questionnaire. Participants were all novice speakers who had done some public speaking but wished to improve their skills in this area. None of the participants had used a multimodal system for public speaking previously. The post-questionnaire consisted of eight items. It was based on the ACM ICMI Multimodal Learning Application Evaluation and Design Guide¹. User experience was evaluated using three questions on naturalness, motivation to use the application again and stress experienced using the application. Awareness was evaluated using four questions on distraction, body awareness, awareness of feedback and awareness of speaking behaviour. An open question was added for additional comments. Following a pretest, it was decided to use a white background for the video stream as the participants stated that they could not discern their dark clothing against a black background.

7.1 Study Format

Each session opened with an introduction consisting of an overview of the study format. Each participant was given a demonstration of both the video and the avatar versions of the system. Participants were asked to speak for one minute on a subject of their choice using each version. Five of the participants used the avatar version first followed by the live video. The other five participants used the video version first followed by the avatar version. Speakers completed the post-questionnaire immediately after using each version. The post-questionnaires contained the same items each time. Afterwards, there was a brief closing interview.

The questionnaire asked them to rate different aspects of the version, which they had just used, on a scale of 1 to 10. Users could also add optional written comments after each question.

8 RESULTS

For each question, we compared a boxplot of the responses for the avatar version with a boxplot of the responses for the video version. The most dramatic difference was for the question on distraction, see Figure

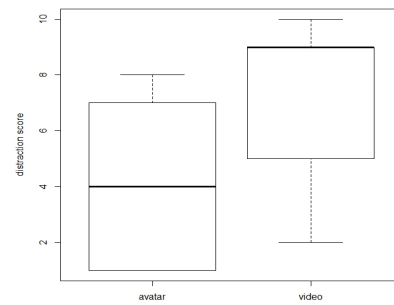


Figure 4: Boxplots of the responses for the avatar version and the video version in answer to the question of distraction. The higher the score, the more distracting was the version. As can be seen, participants reported that the video version was more distracting.

4. Users rated the video version as more distracting than the avatar version. This is consistent with the scores for awareness of speaking performance and for awareness of feedback. Users tended to score the avatar system higher for awareness of speaking performance and awareness of feedback.

The experimental design was a standard hypothesis test, in which the independent variable was the version of system (avatar or live video) and the dependent variable was the level of distraction. Scores ranged from 0 for no distraction to 10 for very distracting.

The order in which the participants used the versions (avatar first or video first) could potentially be a confounding variable. Therefore, the users were divided into two equal-sized groups (avatar first and video first), in order to measure any effect that this variable might have. An Analysis of Variance was carried out and showed that there was no statistically significant difference between the groups.

The p-value for the difference between the levels of distraction for the video version and the avatar version was 0.04, which indicates that there was a significant difference between the responses to the two versions.

We conclude that the live video image was distracting users from their performance and the system feedback. This was confirmed by the users' remarks in the closing interview. Nine of the ten users said that the live video stream was very distracting. They said that they were more aware of aspects of their personal appearance than of their performance or the feedback. All participants reported that they would be motivated to use the system again, both avatar and video versions. Participants reported that they experienced more stress while using the video version of the system.

¹<http://sigmla.org/mla2015/ApplicationGuidelines.pdf> accessed on January 2016

9 DISCUSSION

The fact that all participants reported that they would use the system again highlights the need for multimodal systems for communication skills development. Participants, on the whole did not like seeing themselves in video mode. They made comments such as ‘I did not like seeing myself in this situation’ and ‘I felt awkward looking at myself’. However, three participants did report that they could discern their facial expressions more clearly in video mode. Furthermore, one participant reported that using the system in avatar mode ‘made the act of talking feel too disembodied and therefore harder to relate to the information/feedback provided by the system’. This participant also stated ‘that I was less focused on my speaker behaviour and more inclined to feel distracted by the avatar’. The fact that one user had such an adverse reaction to the avatar illustrates the need for interactive multimodal systems to provide users with the autonomy to choose the appearance of the interface that they are using. The participant who disliked the avatar had a very different experience of using the system in video mode reporting that ‘the systems seems easy and fun to use, I would find this very helpful for preparing lectures and conference presentations’. Three participants reported that they would like to be able to change the height of the display screen and have a bigger screen. The participants who were taller requested the screen height adjustment.

10 CONCLUSION

The main conclusion from our study is that most users prefer to see an avatar rather than a live video stream. Most users find it distracting or even unpleasant to see themselves. The avatar is more abstract and the user can concentrate on body movements and facial expression rather than being distracted by details of personal appearance. We can also conclude that all the users in the study found both versions of the system beneficial and were motivated to use it again. Some users found it difficult to assimilate the feedback while they were giving the speech in real time. They asked if it was possible to record the output from the system (including the feedback) and review it afterwards off-line. This is actually possible with our system and in future work we will be testing to see how users evaluate watching an off-line recording.

11 FUTURE WORK

Work will focus on the following areas. Does the colour of the feedback affect the users evaluation? Would users prefer a more human-looking avatar? Some research shows that users find such avatars more sympathetic especially if the avatar resembles the user themselves (Baylor, 2009), (Suh et al., 2011). We intend to conduct a longitudinal study to evaluate how the users’ speaking performance changes if they use the system over a period of time. We want to investigate how users evaluate a statistical analysis of their performance e.g. recording how many times the system identified particular events such as hands touching, arms crossed, changing view direction etc.

We will evaluate other aspects of Positive Computing with respect to our system: Do users have autonomy i.e. can they select the feedback that they want when they want it? With regard to stress, is the system stressful to use? Does it become less stressful with practice? Does it reduce the stress of live public performance? Does the users’ competence in public speaking increase with use of the system?

ACKNOWLEDGEMENTS

This material is based upon works supported by Dublin City University under the Daniel O’Hare Research Scholarship scheme. System prototypes were developed in collaboration with interns from École Polytechnique de l’Université Paris-Sud and l’École Supérieure d’Informatique, Électronique, Automatique (ESIEA) France.

REFERENCES

- Barmaki, R. (2016). Improving Social Communication Skills Using Kinesics Feedback. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA ’16, pages 86–91, New York, NY, USA. ACM.
- Barmaki, R. and Hughes, C. E. (2015). Providing real-time feedback for student teachers in a virtual rehearsal environment. *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, pages 531–537.
- Batrinca, L., Stratou, G., and Shapiro, A. (2013). Cicero - Towards a Multimodal Virtual Audience Platform for Public Speaking Training. In Aylett, R., Krenn, B., Pelachaud, C., and Shimodaira, H., editors, *Intelligent Virtual Agents*, volume 8108 of *Lecture Notes in Computer Science*, chapter Cicero - T, pages 116–128. Springer Berlin Heidelberg, Berlin, Heidelberg.

- Baylor, A. L. (2009). Promoting motivation with virtual agents and avatars: role of visual presence and appearance. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1535):3559–3565.
- Bubel, M., Jiang, R., Lee, C. H., Shi, W., and Tse, A. (2016). AwareMe: Addressing Fear of Public Speech through Awareness. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 68–73. ACM.
- Calvo, R. A. and Peters, D. (2012). Positive Computing: Technology for a Wiser World. *interactions*, 19(4):28–31.
- Calvo, R. A. and Peters, D. (2014). *Positive Computing: Technology for wellbeing and human potential*. MIT Press.
- Calvo, R. A. and Peters, D. (2015). Introduction to Positive Computing: Technology That Fosters Wellbeing. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, pages 2499–2500, New York, NY, USA. ACM.
- Calvo, R. A. and Peters, D. (2016). Designing Technology to Foster Psychological Wellbeing. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 988–991, New York, NY, USA. ACM.
- Calvo, R. a., Peters, D., Johnson, D., and Rogers, Y. (2014). Autonomy in technology design. *Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems - CHI EA '14*, pages 37–40.
- Chollet, M., Morency, L.-p., Shapiro, A., Scherer, S., and Angeles, L. (2015a). Exploring Feedback Strategies to Improve Public Speaking: An Interactive Virtual Audience Framework. *UbiComp '15: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1143–1154.
- Chollet, M., Stefanov, K., Prendinger, H., and Scherer, S. (2015b). Public Speaking Training with a Multimodal Interactive Virtual Audience Framework - Demonstration. *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, pages 367–368.
- Dermody, F. and Sutherland, A. (2016). Multimodal system for public speaking with real time feedback: a positive computing perspective. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 408–409. ACM.
- Dowrick, P. W. (1999). A review of self modeling and related interventions. *Applied and Preventive Psychology*, 8(1):23–39.
- Dowrick, P. W. and Biggs, S. J. (1983). *Using video: Psychological and social applications*. John Wiley & Sons Inc.
- Dwyer, K. K. and Davidson, M. M. (2012). Is Public Speaking Really More Feared Than Death? *Communication Research Reports*, 29(2):99–107.
- Garau, M. (2006). Selective fidelity: Investigating priorities for the creation of expressive avatars. In *Avatars at Work and Play: Collaboration and Interaction in Shared Virtual Environments*, pages 17–38. Springer.
- Harris, S. R., Kemmerling, R. L., and North, M. M. (2002). Brief virtual reality therapy for public speaking anxiety. *Cyberpsychology & behavior: the impact of the Internet, multimedia and virtual reality on behavior and society*, 5(6):543–550.
- McCroskey, J. C., Booth Butterfield, S., and Payne, S. K. (1989). The impact of communication apprehension on college student retention and success. *Communication Quarterly*, 37(2):100–107.
- Morgan, N. (2015). Why Gesture Is Important - And What You're Not Doing About It.
- Pertaub, D.-P., Slater, M., and Barker, C. (2002). An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and virtual environments*, 11(1):68–78.
- Schneider, J., Börner, D., Van Rosmalen, P., and Specht, M. (2015). Presentation Trainer, your Public Speaking Multimodal Coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 539–546. acm.
- Suh, K.-S., Kim, H., and Suh, E. K. (2011). What if your avatar looks like you? Dual-congruity perspectives for avatar use. *MIS Quarterly*, 35(3):711–729.
- Toastmasters International (2008). *Competent Communication A Practical Guide to Becoming a Better Speaker*.
- Toastmasters International (2011). *Gestures: Your Body Speaks*.