A Multimodal System for Public Speaking with Real Time Feedback

Fiona Dermody
Dublin City University
Dublin,Ireland
Fiona.Dermody3@mail.dcu.ie

Alistair Sutherland
Dublin City University
Dublin, Ireland
Alistair.Sutherland@dcu.ie

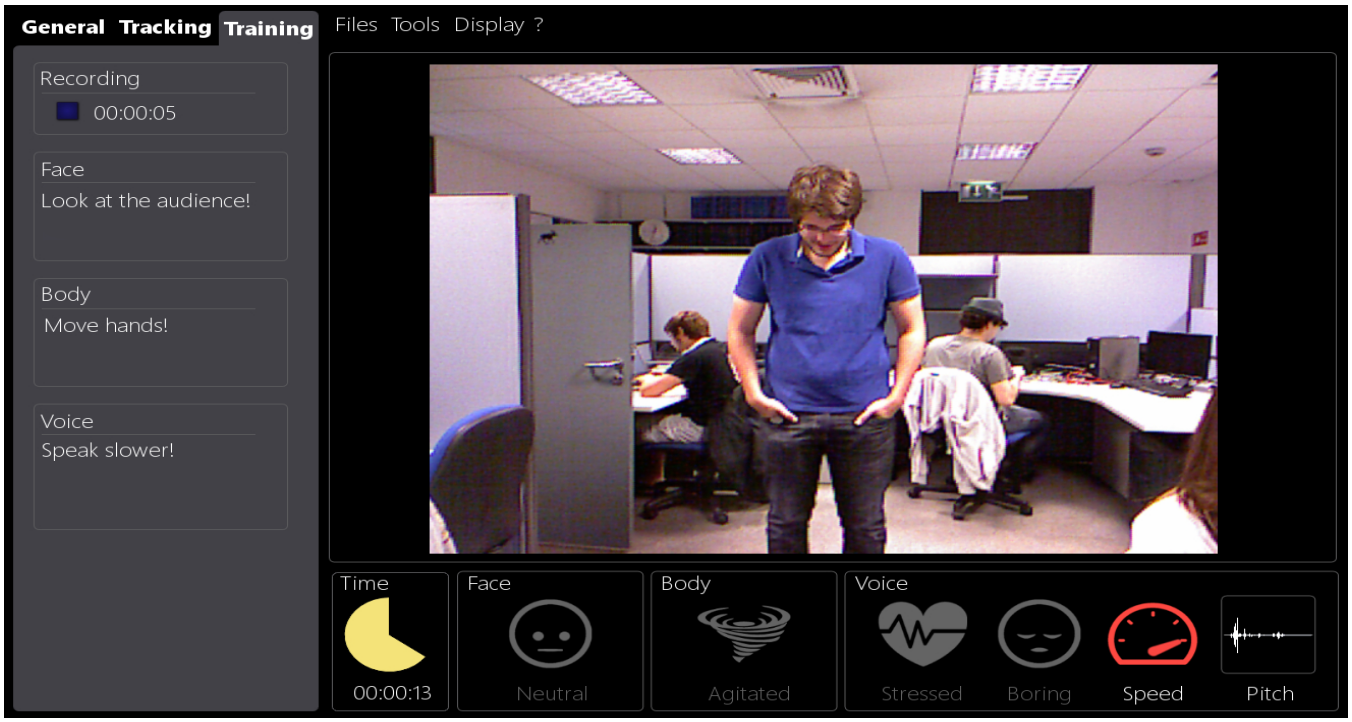

Figure 1: Indicative real time feedback displayed to the user

ABSTRACT

We have developed a multimodal prototype for public speaking with real time feedback using the Microsoft Kinect. Effective speaking involves use of gesture, facial expression, posture, voice as well as the spoken word. These modalities combine to give the appearance of self-confidence in the speaker. This initial prototype detects body pose, facial expressions and voice. Visual and text feedback is displayed in real time to the user using a video panel, icon panel and text feedback panel. The user can also set and view elapsed time during their speaking performance. Real time feedback is displayed on gaze direction, body pose and gesture, vocal tonality, vocal dysfluencies and speaking rate.

ICMI '15, November 09-13, 2015, Seattle, WA, USA
ACM 978-1-4503-3912-4/15/11.
http://dx.doi.org/10.1145/2818346.2823295

Categories and Subject Descriptors

H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces; Evaluation/methodology; Screen Design; User-centered design;

General Terms

Design, Human Factors.

Keywords

Multimodal Interface; Affective Computing; Human Computer Interaction; Social Signal Processing; Public Speaking.

1. INTRODUCTION

Previous systems such as [1] have provided feedback on all speaking modalities such as voice, gestures and stance but not in real time. We believe that public speaking is a multimodal task that requires a multimodal system with real time feedback [2].

2. OVERVIEW OF PROTOTYPE

An initial prototype has been developed based on the results of a user survey. The prototype can give feedback to the users while they are speaking in real time. The user can select if they want textual feedback or visual feedback or both. We can easily vary

the size and color of the textual feedback. The user can also select which combination of speaking modalities they want to receive feedback on. For example, they may elect to receive feedback on their facial expressions only. Alternatively, they may receive feedback on facial expressions, voice and gestures simultaneously. The user can also record their speech and review it afterwards getting more detailed feedback offline. The user can choose to view their speech as live video or as an animated avatar. This enables the user to develop an awareness of how they would appear speaking before an audience. This multimodal interface prototype also contains video tutorials on good speaking practices. It contains example sequences, which the user can imitate and the system gives feedback of the accuracy of their imitation. For example, they could be asked to perform a certain sequence of gestures or they could be asked to vary the pitch of their voice at key moments in the speech. Users will develop the skill of speaking within a certain time limit as the interface has a clock. The clock will give them a visual warning as they approach the end of their allotted timespan.

PROTOTYPE COMPONENTS

2.1 Video Panel

1. The video panel is in the center of the screen and can display the following things
2. A live video of the user performing in front of the Kinect
3. A 3D computer-graphic avatar, which mimics the user's movements. The avatar also has a face which can mimic the user's facial expressions. The limbs of the avatar change color, in order to show whether the Kinect is detecting them or not.
4. A 2D stick man, which tracks the user's movements. This is standard software provided by the Kinect but most users find it difficult to interpret.
5. Recorded videos of the user's performance. The user can use these to review their performance and receive more detailed feedback. The user can pause and rewind the video. They can also review past performances to see how much they have improved.
6. Pre-recorded instructional videos created by a human expert. The user can also view example speeches made by famous orators.
7. The user can select which of the above will be displayed.

2.2 Text Feedback Panel

This sidebar on the left-hand side of the screen contains three textboxes- one for feedback on the face, one for the body and one for the voice. Each box can display several lines of text.

2.3 Icon Feedback Panel

The panel on the bottom of the screen displays feedback in the form of visual icons. There is an icon for the each of face, body and voice. There is also

1. a speedometer, which displays the user's word rate in words per minute. It changes color, if the user is speaking too fast or too slow.
2. a stopwatch, which displays how many minutes and seconds to the end of the allotted time. It changes color to alert the user, when there is, say, one minute left.
3. a pitch-spectrum, which displays the pitch levels of the user's voice

3. TYPES OF FEEDBACK

3.1 Gaze Direction

By detecting the user's gaze direction the system can assess whether the user is engaging different sections of the audience. Gaze direction could also be used to express emotions e.g. looking up could express aspiration, looking down could express humility.

3.2 Body Pose and Gestures

Currently the system can recognize simple body poses such as "hands touching/ hands apart". This allows the system to give feedback on whether the user is making the appropriate gesture at the right time. It will also assess whether the user is moving around on the stage or just standing rigidly in one spot. It will assess whether the user is engaging different sections of the audience by turning the body in different directions. In contrast, the system can also assess whether the user is moving too quickly or is too agitated – perhaps swaying from side to side.

3.3 Vocal Tonality and Dysfluencies

The system can detect monotonous voice. If there is little variation in the pitch of the user's voice, the system will display feedback. It can also detect vocal dysfluencies or 'crutch words' such as 'um' and 'ah'. Speaking rate is also detected as the system can detect the number of words per minute. Feedback on speaking rate is displayed via the speedometer.

4. CONCLUSION

We have developed a multimodal prototype interface to enable a user to develop their skill in public speaking. The initial prototype has been designed using the Microsoft Kinect. It can detect body pose, facial expressions and voice. This multimodal interface will give feedback to the user on their speaking performance in real time.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] Batrinca, L. *et al*. 2013. Cicero - Towards a Multimodal Virtual Audience Platform for Public Speaking Training. Intelligent Virtual Agents. R. Aylett *et al*., eds. Springer Berlin Heidelberg. 116–128.

[2] Dermody, F. *et al*. 2015. A Multi-Modal System for Public Speaking - Pilot Study on evaluation of real-time feedback. INTERACT 2015, Part IV, LNCS 9299. J. Abascal *et al*., eds. Springer Berlin Heidelberg. 3–5. DOI: 10.1007/978-3-319-22723-8_47.