Public Speaking Training with a Multimodal Interactive Virtual Audience Framework - Demonstration

Mathieu Chollet[1], Kalin Stefanov[2], Helmut Prendinger[3] and Stefan Scherer[4]
[1]Japan Society for the Promotion of Science, National Institute of Informatics, Tokyo, Japan
[2]KTH Royal Institute of Technology, Stockholm, Sweden
[3]National Institute of Informatics, Tokyo, Japan
[4]USC Institute for Creative Technologies, Los Angeles, CA, USA
mchollet@enst.fr, kalins@kth.se, scherer@ict.usc.edu, helmut@nii.ac.jp

ABSTRACT

We have developed an interactive virtual audience platform for public speaking training. Users' public speaking behavior is automatically analyzed using multimodal sensors, and ultimodal feedback is produced by virtual characters and generic visual widgets depending on the user's behavior. The flexibility of our system allows to compare different interaction mediums (*e.g.* virtual reality *vs* normal interaction), social situations (*e.g.* one-on-one meetings *vs* large audiences) and trained behaviors (*e.g.* general public speaking performance *vs* specific behaviors).

Categories and Subject Descriptors

H.1.2 [**User/Machine systems**]: [Human information processing];
I.5.4 [**Pattern Recognition Applications**]: Computer Vision, Signal Processing;
K.3.1 [**Computers and Education**]: Computers Uses in Education

Keywords

Virtual audience; public speaking training; automatic behavior recognition

1. INTRODUCTION

Interpersonal skills such as public speaking are essential assets for a large variety of professions and in everyday life. Nonverbal communication (affect, demeanor, posture, eye contact, speech tone and fluency) is a key aspect of successful public speaking and interpersonal communication [5]. Audiences provide indirect feedback during presentations by signaling nonverbally, as they continuously rate and sense the presenter's speaking style, such as nodding and leaning forward in presentations they enjoy, or averting their gaze when they are not interested [3]. Paying attention to these feedback behaviors allows speakers to improve their performance.

ICMI '15 November 09-13, 2015, Seattle, WA, USA

ACM 978-1-4503-3912-4/15/11.
http://dx.doi.org/10.1145/2818346.2823294.

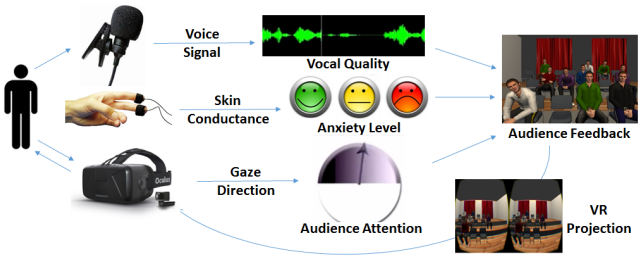

Figure 1: Architecture of our public speaking training system.

However, an actual human audience is not always available or sometimes too intimidating for an anxious speaker. Virtual audiences have already been used successfully in Virtual Reality Therapy (VRT) to mitigate public speaking anxiety [4]. Pushing further, we want to investigate if public speaking performance can be improved using virtual training.

In previous work [1], we built an interactive virtual audience framework for public speaking training which could provide indirect feedback to a user. It relied on a Wizard of Oz to provide input on the user's performance. In this demonstration, we present a fully automatic version of our virtual audience for public speaking training. Multimodal data is obtained from a variety of sensors (e.g. microphone, kinect, physiological sensors), and multimodal feedback is produced by the audience and visual widgets. Additionally, our system was updated to be compatible with a virtual reality headset.

2. SYSTEM ARCHITECTURE

Our public speaking training system consists of two main components: a multimodal signals perception framework, and a virtual environment populated with virtual characters acting as a virtual audience. We use a messaging system allowing for these two components to be distributed over multiple computers.

2.1 Perception framework

Our multimodal signals perception framework integrates signals obtained from a variety of sources. Raw data is analyzed to detect relevant signals or events, such as the oc-

currence of a gesture or the presence or absence of voice. Additionally, these signals can then feed classifiers to provide higher level information, such as the degree of attention the user gives to the audience, or an estimation of the presenter's anxiety. Examples of data sources that have been integrated include:
- *Depth* (*e.g.* Microsoft Kinect): posture, gestures, body activity, ...
- *Audio*: presence or absence of speech, Mel-frequency cepstral coefficients, vocal quality, ...
- *Video* (when not using a virtual reality headset): eyebrow movements, smiles, emotions, ...
- *Physiological*: heart rate, skin conductance, anxiety, ...

Moreover, other inputs or events from the virtual environment can be integrated as perceptual inputs. For instance, the virtual reality headset can provide the head direction of the user and therefore an approximation of his or her gaze direction. Interactions with an integrated slideshow can be used as an indication as whether the presenter is running late or is still on time with the presentation total time.

2.2 Virtual audience

A virtual environment consisting of a presentation room with a slide screen and three layers of chairs has been created within the Unity 3D engine. The layout of the virtual audience (*e.g.* amount of characters, their appearances and their distribution) can be easily customized.

The virtual audience's characters' behavior is driven by the multimodal inputs presented earlier. Each character is assigned a behavior profile, which is a set of rules of the following tuple: $< Descriptor; Signals; Conditions >$. The *Descriptor* refers to the ID of one of the perceptual inputs presented earlier. The *Signals* item represents a behavioral response to be produced by the virtual character when the value of *Descriptor* meets the defined *Conditions*. For instance, in the following example, the character will shake its head when the user has looked at the audience less than half of the time:
$< gaze\_audience,\ < head\ type = "shake" / >,\ in(0, 0.5) >$

Generic visual widgets can also be used to provide direct feedback to the user about his or her performance. For instance, a colored bar can be displayed on top of the screen, directly reflecting the user's performance (*e.g* a full green bar can indicate good performance, whereas a short red bar indicates poor performance), similarly to [2]. Finally, a post-hoc printout of the measured behaviors can be produced to allow the user to reflect on his or her performance.

2.3 Demonstration scenario

For this demonstration, we have designed a public speaking training scenario aimed at training for research conference presentations. Custom slides can be loaded within the virtual environment, allowing a user to train presenting his or her own material. The virtual characters react to the following aspects:

1-*Amount of gaze directed at the audience*: presenters should look at the audience rather than at their slides. Head direction data obtained from the VR headset allows to compute whether the user hasn't looked at the audience for a long time. When that happens, virtual characters signal the user by clearing their throats.

2-*Body activity*: an appropriate amount of gesturing and movement is better while speaking in public. Our perception framework provides a measure of body activity derived from depth data. This measure is used to affect the virtual character's postural behavior: if the user is too active or too passive, the characters lean back in disagreement.

3-*Heart rate*: The user's heart rate or anxiety level can be directly displayed using a visual feedback item (*e.g.* smiling face if the heart rate is in a normal range, sad face if it is elevated).

3. CONCLUSION

We presented a public speaking training system based on an interactive virtual audience framework. Its design is focused on flexibility allowing to easily modify training conditions, audience layouts, trained behaviors and feedback types. In particular, the multiplicity of integrated multimodal sensors allow to detect various types of behaviors, in turn being interpreted into various measures of performance or affect. Multimodal feedback can be produced in realtime according to these behaviors and higher level measures, by virtual characters and generic visual widgets, or after the presentation with a post-hoc printout. Additionally, the system can be used with regular displays or with virtual reality headsets. In future work, we will use this prototype to investigate the efficacy of virtual audiences for the training of public speaking and other related social skills, such as job interviews or conducting meetings. In particular, we will compare and assess which training conditions (e.g. virtual reality *vs* regular interaction) and modalities of feedback (*e.g.* normal audience behavior *vs* exaggerated behavior, such as falling asleep) are the most efficient for improving users' skills.

4. ACKNOWLEDGMENTS

This material is partly supported by the JSPS (Japan Society for the Promotion of Science) and the National Science Foundation under Grant No. IIS-1421330.

5. REFERENCES

[1] M. Chollet, T. Wörtwein, L.-P. Morency, A. Shapiro, and S. Scherer. Exploring feedback strategies to improve public speaking: An interactive virtual audience framework. In *Proceedings of UbiComp'15*, 2015.

[2] I. Damian, C. S. S. Tan, T. Baur, J. Schöning, K. Luyten, and E. André. Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In *Proceedings of CHI'15*, pages 565–574. ACM, 2015.

[3] P. D. MacIntyre, K. A. Thivierge, and J. R. MacDonald. The effects of audience interest, responsiveness, and evaluation on public speaking anxiety and related variables. *Communication Research Reports*, 14(2):157–168, 1997.

[4] D.-P. Pertaub, M. Slater, and C. Barker. An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and Virtual Environments*, 11(1):68–78, Feb. 2002.

[5] E. Strangert and J. Gustafson. What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations. In *Interspeech*, page 1688–1691. ISCA, 2008.