COMUNICARTE: A VIRTUAL REALITY GAME TO IMPROVE PUBLIC SPEAKING SKILLS

M. El-Yamri, A. Romero-Hernandez, M. Gonzalez-Riojo, B. Manero

Complutense University of Madrid (SPAIN)

Abstract

Oratory or the art of public speaking with eloquence has been cultivated since ancient times. Nowadays, the Internet has multiplied the importance of this discipline, because today audiences can be counted in millions of people. However, the fear of speaking in public -a disproportionate reaction to the threatening situation of facing an audience- affects to a very important part of the population.

For example, 57% of young people (12-17 years old) consider the fear of speaking in public as the second most feared social situation [1], and 75% of the population is affected by this fear [2].

To alleviate this fear, we created the Project ComunicArte, a virtual reality videogame for training the ability of public speaking. It is built as an environment where the speaker confronts a virtual audience that reacts in real time to the speaker's features, such as voice, gestures and bio-metric parameters (heart rate or body language, among others). This videogame is focused on the audience, since in real life, the only feedback we receive when we speak in public is provided by our listeners [3]. The novelty of this tool are the emotions extracted from the orator features, which influence directly in the audience reactions. Thus, the videogame includes an audience that reacts in real time to the way the speakers gives the speech, to give them the possibility to react and adapt their speech accordingly.

This paper details the design and creation of this game, its general characteristics and the modelling of its reactive virtual audience.

Keywords: public speaking, virtual reality, educational videogames, emotion analysis.

1 INTRODUCTION

Communicating orally and, specifically, speaking in public, is an important skill in different daily life tasks. When a person communicates, either to another person, or in front of an audience, the message is conveyed by what is said and how it is said. When we communicate, two main factors come into play: verbal communication, and non-verbal communication.

Verbal communication (VC) is the type of communication in which linguistic signs are used to transmit a message. The signs are mostly arbitrary and / or conventional, since they express what is transmitted and are also linear; each symbol goes one after the other.

On the other hand, **Non-verbal communication** (NVC) [4], [5] is the process of communication by sending and receiving messages without words, that is, by signs and cues. NVC arises with the beginnings of the human species before the evolution of the language itself [6]. Animals also show certain types of non-verbal communication. In this communication we can distinguish between body language and gestures, and paralanguage. For the purposes of this work, what interests us most is the last one.

According to Laukka [7], when a message is transmitted, it is coupled with the emotions of the individual. These emotions affect the respiratory system of the speaker, producing the voice tone to be modified. From these changes in the voice tone and without taking into account the message content, we can establish what is the emotion that is affecting this speech. Obviously, not all the emotions that the speaker feels are transmitted or reflected in the voice but, for the purposes of this work, what interests are those emotions that the audience can perceive and, consequently, react to.

There are many recent investigations that have focused on the automatic recognition of emotions in voice. Most of these works [8] are based on the use of neural networks, vector support machines or Gaussian models that train with large volumes of previously classified audio recordings. The emotion generated by the information that the speaker wants to convey has a decisive influence on the selection of the words and the structure of the sentences expressed.

Furthermore, there are different methodologies to detect emotions through text (content), which can be divided into four sections[9]: keyword detection, lexical affinity, statistical processing of natural language and methods based on knowledge of the real world. In the detection of keywords in text, emotions are extracted based on the presence of words that refer to emotions or affective words.

However, analysing emotions is a complex and expensive process. Therefore, the goals of this project do not include the implementation of an emotional analysis system, but several already implemented ones are used.

Using computer tools to improve the ability to speak in public is not new. However, the vast majority of the tools consulted are limited to offering a space to practice a speech, or in some cases, to giving a very limited feedback (such as raise the voice's tone) at the end of the experience [10]–[12]. The reason that there are no tools capable of doing a deeper evaluation is the difficulty involved in this task. Deciding that a speech is right or wrong is a very complex, subjective and tremendously subtle human ability.

There is a precedent work by Chollet el al. [13] that introduces the concept of reactive audiences to give a feedback to the speaker without breaking the illusion of the game world. We used that concept to create our reactive audience. However, our audience only reacts to the speaker' emotions detected by their verbal and non-verbal features.

2 OBJECTIVES

The main objective of this work is to create a virtual environment in the form of a videogame capable of improving a speaker's communication skills through an emotions-responsive virtual audience that provides real-time feedback.

Oral communication can be understood as a continuous phenomenon of adaptation of the speakers to the feedback received by the listener. In any technological tool that pretends to teach oral communication skills, the feedback of the listener must be crucial for the speaker to know if the message being effective or not.

The speaker reacts the same way to a virtual audience as to a real one, and an audience that gives a negative feedback negatively affects the speaker (and vice versa) [14]. If we add to these facts the importance of the listener's feedback mentioned in the previous paragraph, we reach the main objective of this work.

Thus, to achieve our main objective, our work will focus on three sub-objectives: 1) Create an emotions extractor to evaluate speakers' emotions through the voice and speech content, 2) Track speakers' gaze to "direct" the detected emotions, and 3) Create an audience capable of reacting in real time based on the emotions emitted by the speaker.

In this article, we describe the design the first prototype of the emotions-based virtual trainer.

3 TOOL DESIGN

In this section we describe the design route that we followed to create the videogame and carry out the two objectives mentioned in the previous section. First, in section 3.1 we outline how was the virtual audience designed. In section 3.2 we describe the extractor concept and how has the emotion recognition been included in this application.

3.1 General Tool Design

The project has been designed in such a way that it is divided into two decoupled environments that work together. On the one hand, there is the Virtual Environment (VE), where the speaker's action takes place and where the speakers features are collected. On the other hand, there is the Analysis Environment (AE), where these features are processed to generate reactions in the virtual audience and draw conclusions about the effectiveness of the speaker's speech.

In Fig. 1 you can see a diagram of the general functioning of the two environments, from the features collection until when a reaction is generated and sent to the VE, as well as the components involved in the whole process.

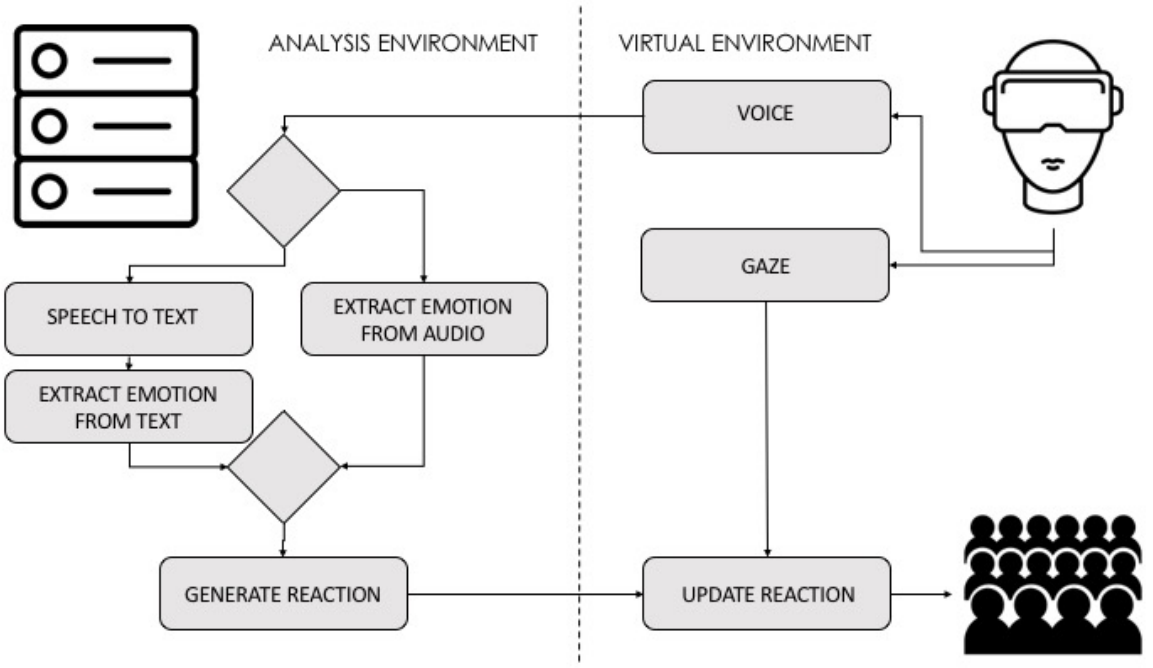

Fig 1. General functioning of the two environments

In our system, the speakers' features can be analysed. These features can be internal or external, Internal Features refer to the parameters of the speaker that the audience does not see. They are more related to biometrical parameters (e.g. sweat or heart rate). These features do not have to affect the reaction of the audience, since the audience is not aware of them, but they are interesting to know how the internal state of the speaker affects his actions and his performance as a speaker.

However, what interests us to generate feedback are the external features, which are external to the speaker. They are those that are clearly appreciable by the audience and, therefore, will have an impact on the degree of its attention. We can distinguish between different types of features according to the parameter that is being analysed at each moment: voice, content or gaze.

Our system has been designed in a modular way to allow the insertion of extraction or analysis components in a simple way, requiring the minimum configuration. The implementation has been done always bearing in mind the scalability of the project. For example, it is possible to create new types of features simply by adding a new type of *XFeature* with new attributes, which will be included in each case in *ExternalFeature* or *InternalFeature*.

Likewise, one or several extractors can be added to analyse this newly added feature, generating reactions that will be added to the sum of reactions sent to the VE, improving the reactions of the virtual audience and giving a more realistic feedback to the speaker.

3.2 Extractors: Automatic Emotion Recognition

This project is pioneer in using automatic recognition of emotions to generate feedback from the virtual audience. The analysis of the emotions transmitted by the speaker is used to subsequently generate reactions in real time in the virtual audience, so that the speaker receives feedback and can modify his speech (on the fly) accordingly.

This is the main functionality of the extractor modules, which use the tools of emotion recognition and, thus, determine the emotion present in the speaker's voice or in speech content. Once these emotions are processed and analysed, our feedback system generates a percentage of the speaker's effectiveness. This percentage of effectiveness is later translated into reactions of the virtual audience, which provides feedback to the speaker in real time about how effective is the speech.

Below we detail how are audio and text features processed, as those features are the ones that we can apply emotion recognition to.

- Voice Feature: Fragment of audio of short duration (5 - 10 seconds) recorded of the speech that the speaker is doing. This type of features is recorded in a superimposed way, with a recording

window (10 initial seconds are processed and then an analysis is generated every 5 seconds). These features are used to analyse the voice tone, without taking into account the content of the speech, and after that, detect emotions transmitted by the speaker from these fragments.

- Content Feature: This type of feature focuses on the content of the speech and it is obtained from the audio fragments mentioned above. These fragments are processed to convert the audio into the text. Subsequently, they will be divided into sentences and analysed to determine the emotion of the speaker according to the content of the speech.

3.3 Virtual Audience Design

As we have seen, one of the most important aspects in oral communication is the feedback of the interlocutor (the person with whom you are talking, or an audience). Public speaking is an artisanal skill, the speaker must have clues to know when the speech is effective and when it is failing. Also, when we communicate we transmit emotions, which are those that the interlocutor is able to detect, and based on these emotions, our listeners are able to provide such feedback.

The Virtual Environment is a project developed with a videogame engine and it is where the speaker's presentation takes place. It is built with a scenario on which the speaker stands and can move freely.

Opposite to the speaker, there is the virtual audience, composed of a set of ACMs (Audience Character Model) that are agents that react in real time to the actions of the speaker: variations in the speech, areas to which the speaker's gaze is directed, voice tone, etc.

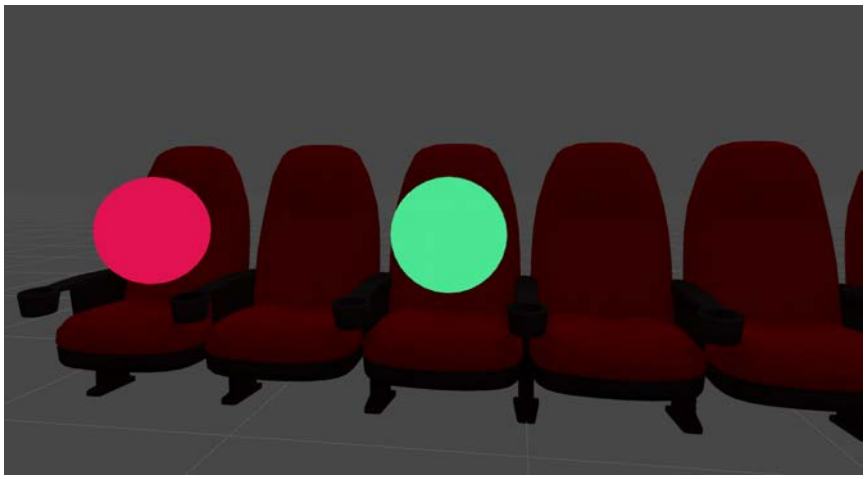

Fig 2. Audience Character Models (ACMs)

The audience appears in a pseudo-random way in the seats and each individual in the audience is an ACM. In addition, at all times, the speaker can see at the back of the room data about how is his speech progressing: percentage of effectiveness, number of attentive ACMs and the emotions perceived in the speech.

The reactions are shown in this first prototype with variations in colour, making an interpolation of red (boring, not attentive) to green (very attentive).

The spherical shape (see figure) of the ACMs has been designed in this way since 3D modelling is an expensive process and because there are studies [15] that show that actors in a virtual world or in a game need not be realistic. The important matter is that these actors have a consistent behaviour.

4 CONCLUSIONS

Oral communication is inherent to human beings, and allows us to transmit emotions, desires, facts or ideas our daily life. In an oral communication there are two types of underlying communications: verbal and non-verbal. When communicating, the attitude of the listener is crucial, since this is what provides us with feedback so that we can evaluate the effectiveness of our speech.

In this paper we have presented the design of a videogame that helps improve the ability public speaking using virtual reality. We extract the emotions that the speaker experience and transmit with

the help of market available tools. The system sends those emotions to the virtual audience taking into consideration the speakers' gaze.

In addition, we created a virtual audience capable of analysing the emotions transmitted by the speaker with his speech and react based on that analysis. Thus, the speaker receives real-time feedback based on the emotions extracted from the analysis of his both verbal and non-verbal factors.

By playing with the weights assigned to each of the factors introduced in the algorithm, we can model different types of audiences: friendlier or more severe, areas of the audience that focus only on one factor or even audiences that are learning from the speaker's speech.

5 FUTURE WORK

The main future work in this project will be to improve and expand the feedback algorithm of the feedback system with new features. Currently, this algorithm takes into account some of the speaker's features (voice tone, discourse content and where he looks at) and assigns different weights according to the emotion the speaker projects with each of these features. The improvement in this sense would be given by adding new factors to analyse, and include a learning process about the speaker, so that the system will learn from the speaker's actions and react accordingly.

Currently, our team is working on the inclusion of biometric factors to apply emotion recognition such as: heart rate, electromyogram, body temperature or galvanic skin response (GSR). Also, another one of the improvements being considered is the use of automatic question generators so that the audience can ask the speaker question based on the speech in an automatic manner.

ACKNOWLEDGEMENTS

This project has been partially funded by BBVA foundation (ComunicArte project: PR2005-174/01), and Ministry of Science, Innovation and Universities of Spain (Didascalias, RTI2018-096401-A-I00)

REFERENCES

[1] C. A. Essau, J. Conradt, and F. Petermann, "Frequency and comorbidity of social phobia and social fears in adolescents," *Behav. Res. Ther.*, vol. 37, no. 9, pp. 831–843, 1999.

[2] M. Gratacós, "DO YOU SUFFER FROM GLOSSOPHOBIA?," 2019. [Online]. Available: http://www.glossophobia.com/index.html. [Accessed: 09-May-2019].

[3] 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 the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015, pp. 1143–1154.

[4] A. Mehrabian, *Nonverbal communication*. Routledge, 2017.

[5] M. Knapp, *Comunicación no verbal*. Paidós, 1999.

[6] J. C. Gómez, *Los inicios de la comunicación: estudio comparado de niños y primates no humanos e implicaciones para el autismo*, vol. 106. Ministerio de Educación, 1995.

[7] P. Laukka, P. Juslin, and R. Bresin, "A dimensional approach to vocal expression of emotion," *Cogn. Emot.*, vol. 19, no. 5, pp. 633–653, 2005.

[8] M. Milošević and Z. Đurović, "Challenges in emotion speech recognition," in *3rd International Conference on Electrical, Electronic and Computing Engineering. IcETRAN*, 2015.

[9] V. Francisco Gilmartín, "Identificación Automática del Contenido Afectivo de un Texto y su Papel en la Presentación de Información." Universidad Complutense de Madrid, Servicio de Publicaciones, 2009.

[10] M. M. North, S. M. North, and J. R. Coble, "Virtual reality therapy: an effective treatment for the fear of public speaking," *Int. J. Virtual Real.*, vol. 3, no. 3, pp. 1–6, 2015.

[11] P. L. Anderson, E. Zimand, L. F. Hodges, and B. O. Rothbaum, "Cognitive behavioral therapy for public-speaking anxiety using virtual reality for exposure," *Depress. Anxiety*, vol. 22, no. 3, pp. 156–158, 2005.

[12] L. Batrinca, G. Stratou, A. Shapiro, L.-P. Morency, and S. Scherer, “Cicero-towards a multimodal virtual audience platform for public speaking training,” in *International workshop on intelligent virtual agents*, 2013, pp. 116–128.

[13] M. Chollet, K. Stefanov, H. Prendinger, and S. Scherer, “Public speaking training with a multimodal interactive virtual audience framework,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 367–368.

[14] M. Slater, D.-P. Pertaub, C. Barker, and D. M. Clark, “An experimental study on fear of public speaking using a virtual environment.,” *Cyberpsychol. Behav.*, vol. 9, no. 5, pp. 627–33, 2006.

[15] V. Vinayagamoorthy, A. Steed, and M. Slater, “Building characters: Lessons drawn from virtual environments,” in *Proceedings of Toward Social Mechanisms of Android Science: A CogSci 2005 Workshop*, 2005, pp. 119–126.