2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)

Emotions-responsive audiences for VR public speaking simulators based on the speakers' voice

Meriem El-Yamri*, Alejandro Romero-Hernandez†, Manuel Gonzalez-Riojo‡ and Borja Manero§
Software Engineering and Artificial Intelligence Department,
Computer Science and Engineering Faculty,
Complutense University of Madrid
Madrid, Spain
*melyamri@ucm.es, †alerom02@ucm.es, ‡manuel.gonzalez@ucm.es, §bmanero@ucm.es

***Abstract*—Oratory or the art of public speaking with eloquence has been cultivated since ancient times. However, the fear of speaking in public -a disproportionate reaction to the threatening situation of facing an audience- affects a very important part of the population. This work arises from the need to help alleviate this fear through a tool where to train the ability of public speaking. To this purpose, we built a virtual reality system that offers the speaker a safe environment to practice presentations. Since the audience is the only way to receive feedback when giving a speech, our system offers a virtual audience that reacts and gives real-time feedback based on the emotions conveyed by three parameters: voice tone, speech content and speaker's gaze. In this paper, we detail the modelling of the audience just focusing on the speakers' voice tone: 1) by presenting an algorithm that controls the audience reactions based on the emotions beamed by the speaker, and 2) by carrying out an experiment comparing the reactions generated by the agents with those of a real audience to the same speech, in order to refine the given algorithm.**

***Index Terms*—public speaking; educational videogame; virtual reality; emotion analysis**

I. Introduction

Since ancient times, the need of public speaking has been a constant for human kind [1]. Public speaking has been used throughout history to persuade, convince, teach or even to conduct the thinking of others.

Oral communication also includes other things besides from the very content of the discourse itself: both, non-verbal communication [2](body expression, movement, gestures, etc.) and everything that gives meaning to the content (rhythm, voice tone, etc.).

Public speaking is a cross-disciplinary practice to different areas of human life: apart from giving a conference before an audience, it can also be put into practice to speak at a neighbors' meeting, intervene in class, give a point of view and defend it, speak in front of an HR manager at a job interview or even give a speech at a wedding.

However, the cultivation of this discipline has faced an obstacle that has also accompanied human kind throughout history: the fear of public speaking. In order to alleviate this fear, which affects a generalized part of the population (75%) [3], it is necessary to practice and train this skill. This is the main objective of the tool presented in this paper, which allows practicing public speaking in a virtual environment facing a reactive audience composed of software agents.

This project has been partially funded by BBVA foundation (ComunicArte project: PR2005-174/01).

A. Related Work

Throughout history, the fear of public speaking has been treated in different ways: theater or improvisation activities, behavioral therapies, workshops for public speaking, etc. And in the early 90s, tools that make use of technology to address fears or train certain skills begin to proliferate.

It is necessary to know that the applications that were pioneers in training using virtual reality were the simulators. They have also recently been used to treat stage fright and fear of public speaking. The first applications just offered a virtual stage where the player can rehearse his discourse [4], [5]. The experiment conducted by Pertaub et al. [6] concluded that speakers reacted in a similar way whether the audiences were virtual of real. This led to some researchers to create audiences with certain behaviors assigned manually. Fukuda et al. [7] created a virtual classroom that determined, with the help of experts in the field, the behavior of the agents based on 6 emotional states, and Kang et al. [8] conducted a similar experiment building a virtual audience with subtle variations in their reactions.

More recently, Chollet et al. [9] presented a demonstration of a platform to teach public speaking where the audience reacted to some parameters of the speech. This systems were not based on virtual reality, the audience was presented in 2D screens. We want to highlight a promising work of the same author [10] involving a reactive virtual audience, although they did not provide much insight of how the system was implemented.

The aim to create reactive audiences is comprehensible, since in the real world, the only feedback that the speakers receive when giving a talk is the audience reactions. That is what makes them vary their actions in real time. Clearly, a simulator for public speaking must include a reactive audience, otherwise we would find ourselves in front of a dead audience that would only help the speaker become familiar with the stage.

In the tool we developed, the speakers can practice and improve their speech and their ability in front of an audience,

2161-377X/19/$31.00 ©2019 IEEE
DOI 10.1109/ICALT.2019.00108

as they receive instant feedback from them. Our virtual audience of software agents does not have a hardwired behavior as in some of the cases studied, but rather reacts based on an analysis carried out in real time according to the speaker's actions. In some of the previous works [11], the feedback provided is based on some parameters (e.g. voice, body, gaze) with descriptors. However, what we intend to analyze in this work is in a different layer: the emotions layer. Although we analyze more or less the same speakers parameters, we do not generate reactions based on numeric values for those parameters, but rather detect which emotion is present in those parameters and if the speaker is transmitting coherent emotions to the audience.

According to Laukka [12], both when speaking and when we give a talk, the message transmitted is accompanied by our emotions. From the changes in the voice tone we can establish what emotion is affecting the speech [13]. Obviously, not all the emotions that the speaker feels are transmitted or reflected in the voice tone but, for the purposes of this work, what interests us are those emotions that the audience can perceive and, consequently, react to.

In other works, each parameter has values that indicate whether it is positive or negative, but the rest of the parameters that are being analyzed at the same time are not taken into account. However, one of the most important things in oral communication is coherence [14], that is, an effective communication is one in which the speaker transmits the same emotions with what he says and how he says it. For example, if a speaker has a tone of voice that conveys sadness, but the body movement indicates aggression, his speech is not being coherent, and therefore, his communication is not natural or effective. On the contrary, if the same emotion is detected in the set of parameters, it means that the speaker is reaching his audience and is communicating effectively.

B. Objectives

The main goal of this work is to create a virtual environment capable of improving a speaker's communication skills by creating a reactive audience that gives feedback in real time. In the dialogue between humans, the feedback of the listener is a phenomenon with one of the most important functions in the conversation coordination. It works both to regulate the flow and to create and ensure understanding between the two interlocutors. This makes feedback an interesting mechanism to apply also in the conversation between agents or in human-agent interaction [15].

Our objective is to use virtual reality to create an environment in which the speaker can face a public speaking situation that resembles real life as much as possible. To this purpose, our work is focused on creating an audience of software agents capable of reacting in real time, based on the actions of the speaker (voice tone, speech content and gaze direction). We attempt to create reactions in the same way that happens in a conversation between two people or when a person is sitting among an audience.

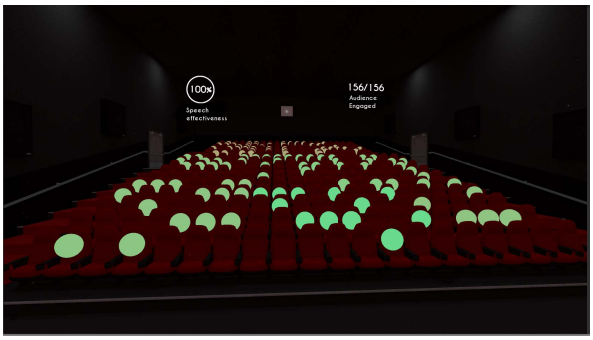

Fig. 1. Game environment

II. Tool Design

Next, we describe the decisions that were made when designing the tool and how this process was carried out, from the general mechanics to the particular design of the virtual audience agents.

A. Game Environment

The virtual environment tries to put the user in diverse public speaking situations as realistic as possible (i.e. a large audience to give a lecture, a class or a job interview) (i.e. a large audience to give a lecture, a class or a job interview), allowing the speakers to train their speech and develop their skills to better cope with this type of situations in real life (see figure 1).

The game is created as virtual reality experience, in order to improve the immersion feeling and the realism sensation that the user perceives. In addition, we chose to present the system as a videogame to motivate the users in the objectives achievement and thereby improve their learning. This is intended to ensure that the players have the feeling of being in an environment in which they are protected. As Johan Huizinga pointed out in his Homo Ludens [16], when we play, we get inside a magic circle where our actions have no impact outside that circle (in real life). This allows us, as they do in other sectors to educate their workers, to lose the fear of making mistakes. And a lack of fear is crucial for the learning process to occur. Within our safe environment, the speakers will not be afraid to make a fool of themselves, which will make them able to train for the real life scenario.

The reaction of an audience to a presentation -or certain to parts of it - offers the speakers the opportunity to adapt and react in real time, just as they would do in front of a real audience, but with the advantages of doing it in a safe environment. This allows the speakers to prepare in advance and modify their speeches to improve the virtual audience reaction.

Well-implemented gamification improves the learning process [17]. However, we have to be very careful with the elements to gamify. On many occasions, the desire to gamify everything, takes the player out of the game experience [18].

In our system, we make that audience react in real time through the analysis of different parameters of the speaker -voice tone and projection, speech content and gaze direction-. Thereby, we get a gamified environment where the speaker has to get the attention of the maximum number of possible attendees.

This type of feedback, compared to other possible ones (scores, game alarms, etc.), is much more favorable to immersion, since it simulates what happens in reality when we give a speech. Each one of the agents of the audience generates independent reactions based on the emotions that are extracted from the speakers features. These agents are modeled to try to predict how the actions of the speaker would impact a real person from the audience.

For the purposes of this work, we decided to focus on external speaker features, since they are the ones that the audience is capable of perceiving. And since we pursue to replicate the reaction of a real audience, we chose three external characteristics: 1) the speakers voice, 2) the content of the speech, and 3) The speaker gaze. These three features have been chosen due to three fundamental reasons: 1) they simplify their capture and analysis for this first prototype, 2) their capture is not invasive for the speaker, and 3) they determine whether a speech is good or not.

B. Designing the audience of ACMs

In this first prototype, the audience individuals -ACMs (Audience Character Model)- have the shape of a sphere that changes color according to the reaction it is representing at each moment. The agents have been modeled as simple spheres in order to generate a functional prototype in the shortest time possible and to test the agents reactions, since 3D modeling is an expensive process that would have required more time, and because there are studies [19] that show that characters in a virtual world or in a game do not necessarily have to be realistic. The important thing is that these characters have a consistent and coherent behavior. That is to say, if a character in a virtual environment acts in a consistent way, the player can accept that character as real even though the characters appearance is not realistic.

III. Reactive Audience

In this section we will first detail the behavior of an ACM. We will also describe the voice parameter and an experiment that we did with said parameter.

A. ACM Behaviour

In order to keep this prototype simple, the agents behaviour is very basic. What we aim to reflect with this behaviour is the attention degree the ACM is paying to the speakers speech, in normalized values; being the values close to 0 indicators of the agent not being attentive (non-engaged), and those close to 1 indicators of the agent being attentive to the presentation (engaged). The reactions are shown in this prototype with color variations, making an interpolation from red (bored, not attentive) to green (very attentive). Some of the color variation are shown in figure.

Each one of the agents has a degree of severity that can be assigned as considered. This value, also normalized, indicates how reticent an agent is to be attentive to a speech. The more severe, the more difficult it is to keep the agent green all the time. This parameter has been inserted in the agents to be able to model different "personalities" in the audience, given that in a conference in the real world, not all the audience members have the same predisposition to be attentive.

One of the things that we highly cared about is that the audience reactions could be captured at a glance. To achieve that, we followed this pattern: the general color of the audience or of an area of the auditory can give a clear idea to the speaker about how the presentation is progressing and if there should be any variations in the speech -to modify the voice tone or look more towards a certain part of the audience-.

Once these three features have been processed, we analyze them through a series of APIs that allow to detect the emotion transmitted by the speaker at each moment. With this information, and with the percentage of attention that the speaker gives to each section of its virtual audience, we created an algorithm that calculates a percentage of effectiveness of the speakers speech. This percentage of effectiveness is subsequently translated into reactions of the ACMs, which use this percentage of the analysis, in combination with the severity measure they have, and generate a reaction in the form of a color change. This way, they provide feedback to the speaker in real time about what the speech is like and if they are attentive or not.

Given that *Audio Effectiveness = AE, Audio Weight = AW, Text Effectiveness = TE, Text Weight = TW, Focus Effectiveness = FE, Focus Weight = FW and ACM Severity = AS*, the ACM reaction is generated as follows:

$$ACMReaction = \frac{AE \cdot AW + TE \cdot TW + FE \cdot FW}{100} - AS$$

The effectiveness percentages for each one of the features are described below. The weights assigned to each one of them are editable and allow us to tune the ACMs to be more focused on one feature or another as desired. By default, and taking into account the literature about public speaking [20], we have given more weight to the reaction obtained from the voice feature, with 50%, the speech content is assigned a 30% weight, and the speakers gaze direction affects the agent reaction in a 20%.

B. Audio Feature

We iteratively record a short audio fragment (5 - 10 seconds) of the speakers speech. These fragments will later be used to analyze the voice tone and detect emotions transmitted by the speaker.

To this end, we used an API for the analysis of emotions in the voice, which allows us to analyze the audio fragments -with a certain degree of certainty- and extract the predominant emotion or group of emotions. Based on the emotions detected,

Emotion	Weight
Boredom	30
Stress	50
Neutral	100
Calmness	80
Happiness	100

TABLE I
EMOTION WEIGHTS BASE ON VOICE TONE. FIRST APPROACH

Emotion	Weight
Boredom	50
Stress	100
Neutral	80
Calmness	100
Happiness	100

TABLE II
EMOTION WEIGHTS BASE ON VOICE TONE. SECOND APPROACH.

we created a rules based formula that, according to the emotion that the speaker is transmitting at each moment, assigns a percentage of effectiveness to it.

We assigned a weight to each one of the emotions detected by the emotion analysis API from voice I, with values ranging between 0 and 100, which indicate how that emotion affects the speech of the speaker.

In addition to the emotion detected, the Emotion Analysis API provides a confidence score, which ranges from 0 to 100 and indicates how confident is the API that the detected emotion is the correct one.

With these data (*Emotion Weight = EW, Confidence Score = CS*) and a simple formula, we generate a percentage of speech effectiveness from voice (*AE*).

$$AE = \frac{EW \cdot CS}{10000}$$

Once the behavior algorithm of the agents in terms of the speaker's voice was developed, it was necessary to do an experiment that allowed us to verify how similar the reactions of the agents were compared with those of a real audience based on this factor, given that the the purpose of the reactive virtual audience is to resemble in as much as possible the reactions of a real audience.

1) Experiment: **Objective**

The aim of this experiment was to adjust the weights of each emotion in our algorithm to better simulate the reaction of a real audience.

Participants

In the experiment participated a total of 19 people. 16 of them were the audience and 3 actors who were in charge of interpreting in front of the real audience speeches with emotion changes.

Experimental Design

To achieve our goal, we compared compare the the reactions generated by our system with the reactions of a real audience.

First, one actor gave a 5 7 minutes prepared speech playing different emotions along it. The audience was asked to: 1) Write down the exact time in which they detect any change in the emotion transmitted by the actor (there is a chronometer in the room to that aim), 2) Identify the emotion, and 3) Evaluate the effectiveness of the speech on a scale of 1 to 10 at that precise moment. The actor repeated this design 4 times playing different emotions.

In order to compare the results obtained by the experiment audience, with the reactions generated by the agents of the system, the audios of the actors speeches were processed by the reaction system of the ACMs.

Results and discussion

The main aim of the experiment was to assess the effectiveness of the voice analysis system, and to refine the algorithm based on the behavior of the voice analysis (see section III-B). Space limitations did not allow us to include all the results obtained in the experiment. However, results showed that the voice analysis API does not have a high percentage of effectiveness (around 55% of success) in terms of specific emotions, but it could accurately predict emotion that were similar to the one transmitted in the audio fragment.

It has also shown that each member of a real audience reacts in a very similar way to a certain speech, since there have been very few variations among the subjects in terms of perceived emotions.

Based on these results, we changed the values of emotion weights in order to get a more realistic audience in our virtual system. New values are shown in the table II.

These weights have been adjusted taking into account the engagement values reflected by the real audience to the various emotions detected. As example, we want to highlight the change occurred with the Stress emotion (anger, aggression), which, a priori, has been given a fairly low weight, because we assumed that this emotion should not generate engagement in the audience (on the contrary, we assumed it was counter-productive to the engagement). However, after conducting the experiment, we noticed that the emotion of Stress produces the opposite effect, and generates a direct connection (or re-connection) with the speakers when they begin to transmit this emotion. For this reason, its weight has been changed to 100. Even so, although this emotion gets the attention of the audience if it is used sporadically, a speech with a continuous anger or aggressiveness tone is not pleasant for the audience, and makes them disconnect from the speech after the first few seconds. For this reason, despite the fact that the weight of the Stress emotion is 100, we also included a verification of the 5 (this number may vary) previous emotions that the speaker has transmitted, and if all of them have been this emotion of Stress, its weight is considerably lowered in the speech effectiveness formula. This is a small approach to automatic learning, which implies that based on what they have been seeing and hearing during a speaker's speech, they can learn and react differently to the same emotion transmitted.

IV. CONCLUSIONS

Speaking in public is a discipline that cuts across many aspects of human life, encompassing very diverse tasks: giving a lecture, speaking at a neighbors' meeting or facing a job interview. However, a high percentage of the population is

afraid to speak in public. In order to master this discipline, training proves to be the key. In addition, when public speaking, the attitude of the audience is very important, since this is what provides feedback to the speakers so they can see the effectiveness of their speech at the exact moment they are giving it. Thus, the creation of realistic reactive audiences is critical in order to achieve effective environments to train public speaking. First, we have presented a virtual reality videogame in which speakers can practice their presentations in a safe environment, facing a virtual audience who reacts based on some features of the speaker's speech: the voice tone, where the speaker looks at and the speech content. In this paper, we have tried to refine the modelling of the audience just focusing the speakers voice tone. To that aim, we have firstly presented an algorithm that controls the audience reactions based on the emotions beamed by the speaker. In order for the reactions to be as similar as a real audience would have, we carried out a small experiment comparing the reactions generated by the agents with those of a real audience to the same speech. Thanks to this experiment it has been possible to refine the algorithm of reactions of the agents. The main contribution of this work is the refined algorithm to generate audience reactions based on the speakers voice tone. Modelling realistic reactive audiences to a subjective task as public speaking is highly complex, even so, we have brought a scalable way to add "intelligence" to them. Moreover, as seen in the related work section, other authors have addressed this problem, but they never explained the used mechanism in the published work. We want to highlight the need of revealing the success or failed experiences with this kind of algorithms in order to build on top of others work. One good example of the refinement process has been when the speaker transmits stress emotion in his voice. Before, we assumed that stress would cause disengagement in the audience. However, the results showed a different reality, what helped us to adjust our algorithm. Through the experiment, we also confirmed that the voice is one of the main factors that affects a real audience, and thus, using it in the agents reactions turned out to be a good decision.

V. Future Work

There is a lot of future work that has to be done on this project and we have already begun to work on it. One of the main goals is to improve the reactions algorithm of the agents. Currently, this algorithm takes into account some features of the speakers speech (voice, discourse content and gaze direction) and assigns different weights to them. The improvement, in this sense, would be given by adding new factors to analyze (i.e. heart rate, skin conductivity, alpha waves, posture, etc.), and including a learning process in the agent, so that it can improve and react accordingly based on the gathered data. Much more experiments are needed in order to: 1) test the effectiveness of our tool to diminish the public speaking fear, and 2) to better understand the paradigm behind the audiences reactions.

References

[1] Aristotle, *Rhetoric*. A D Publishing, 2009.

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

[3] M. Gratacós. Glossophobia. do you suffer from glossophobia? [Online]. Available: http://http://www.glossophobia.com/

[4] M. M. North, S. M. North, and J. R. Coble, "VIRTUAL REALITY THERAPY: AN EFFECTIVE TREATMENT FOR THE FEAR OF PUBLIC SPEAKING," *International Journal of Virtual Reality (IJVR)*, vol. 03, no. 3, pp. 1–6, Dec. 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01530637

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

[6] 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*, vol. 11, no. 1, pp. 68–78, 2002. [Online]. Available: https://doi.org/10.1162/105474602317343668

[7] M. Fukuda, H.-H. Huang, N. Ohta, and K. Kuwabara, "Proposal of a parameterized atmosphere generation model in a virtual classroom," in *Proceedings of the 5th International Conference on Human Agent Interaction*, ser. HAI '17. New York, NY, USA: ACM, 2017, pp. 11–16. [Online]. Available: http://doi.acm.org/10.1145/3125739.3125776

[8] N. Kang, W.-P. Brinkman, M. Birna van Riemsdijk, and M. Neerincx, "The design of virtual audiences," *Comput. Hum. Behav.*, vol. 55, no. PB, pp. 680–694, Feb. 2016. [Online]. Available: http://dx.doi.org/10.1016/j.chb.2015.10.008

[9] 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*, ser. UbiComp '15. New York, NY, USA: ACM, 2015, pp. 1143–1154. [Online]. Available: http://doi.acm.org/10.1145/2750858.2806060

[10] 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*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 367–368. [Online]. Available: http://doi.acm.org/10.1145/2818346.2823294

[11] 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*. Springer, 2013, pp. 116–128.

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

[13] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual cues in nonverbal vocal expressions of emotion," *Quarterly Journal of Experimental Psychology*, vol. 63, no. 11, pp. 2251–2272, 2010.

[14] R. T. Peterson, "An examination of the relative effectiveness of training in nonverbal communication: Personal selling implications," *Journal of Marketing Education*, vol. 27, no. 2, pp. 143–150, 2005.

[15] H. Buschmeier and S. Kopp, "Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 1213–1221.

[16] J. Huizinga, *Homo Ludens Ils 86*. Routledge, 2014.

[17] K. M. Kapp, *The gamification of learning and instruction: game-based methods and strategies for training and education*. John Wiley & Sons, 2012.

[18] C. Fernández Vara, "The tribulations of adventure games: integrating story into simulation through performance," Ph.D. dissertation, Georgia Institute of Technology, 2009.

[19] 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.

[20] D. Carnegie, *How to develop self-confidence and influence people by public speaking.* Simon and Schuster, 2017.