Smart Learning Environments

RESEARCH Open Access

Designing a VR game for public speaking based on speakers features: a case study

Meriem El-Yamri*, Alejandro Romero-Hernandez, Manuel Gonzalez-Riojo and Borja Manero

*Correspondence:
melyamri@ucm.es
Software Engineering and Artificial Intelligence Department, Computer Science and Engineering Faculty, Complutense University of Madrid, Madrid, Spain

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 offer s 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 a behavioural-realistic 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. In this experiment, the audience subjects are asked to fill a questionnaire - level of engagement and perceived emotions - for a speech performed by professional actors representing different emotions. Afterwards, we compared the reactions of said audience with the ones generated by our algorithm, and used the results to improve it.

Keywords: Public speaking, Educational video game, Virtual reality, Emotion analysis

Introduction

Since ancient times, the need of public speaking has been a constant for human kind (Aristotle & Rhetoric, 2009). 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 (Mehrabian, 2017)(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. Many experts convey that the only way to alleviate this fear, which affects a generalized part of the population (75%) (Gratacós, 2018), is by practicing and training this skill. Nevertheless, practice also requires feedback. Practicing alone in front of a mirror helps, but it is not enough to find your best public speaker. To achieve that, we need an external feedback on if we are doing well or not. The problem is that finding experts capable of evaluating a speech, and what is more important, capable of providing an effective feedback, is not an easy task. The main goal of this project is to build an effective technological tool capable of recreate an environment to practice, and a realistic (in the behaviour) audience that provides feedback to the speaker in real time. We are aware, as many authors have noted, that creating an audience capable of evaluating all the parameters that we humans do, requires much work. Besides, analyzing those parameters to give a right feedback is an enormous challenge. Maybe we will need time to answer questions like: Is a speech good or bad? or What features should a certain speaker improve to get better results? However, the system that we present in this paper tries to shed some light in this matter in order to ease the way of creating a system that understands the rules of human rhetoric. Thus, the main objective of the tool presented in this paper is to present the modelling of a reactive audience composed of software agents based on the speaker's voice tone.

This paper is structured as follows: the present section outlines the previous work and the objectives of this paper; the next section details the tool design as a game. After that, we review the creation of the reactive audience; then, we include a detailed review of the experiment carried out to adjust our tool; finally, the last two sections present the conclusions and the future work, respectively.

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.

There are a lot of situations when simulations are used to improve skills. Videogames are one of them, and in the last decades, there has been an emerging trend: serious games. These type of games (from now on educational video games) are a proven solution as a learning tool (Brom et al., 2011; Hwang et al., 2012). Educational video games are applied in a multitude of teaching fields such as: mathematics (Bos & Shami, 2006; Lowrie & Jorgensen, 2011), computer science (Papastergiou, 2009), social sciences (López & Cáceres, 2010) or geography (Tüzün et al., 2009).

To contextualize a little bit more, it is necessary to know that the applications that were pioneers in training using virtual reality were the simulators. There are many simulators dedicated to training and learning. Below, we describe some of the most important areas in which applications of this type are used.

In the military field, simulators have been used for decades. Since the 1950s, the United States already had military training systems (Page & Smith, 1998), where soldiers confronted various combat situations in safe environments or practiced skills like piloting aircrafts, tanks or navy ships.

In the health area, where a failure also implies critical consequences, these tools are also being used more and more to simulate risky surgical operations (Delingette et al., 1999) or learn to perform daily tasks of patient treatment and diagnosis (Halan et al., 2018).

They have also recently been used to treat stage fright and fear of public speaking. The first applications just offer a virtual stage where the player can rehearse a discourse (M. North et al., 2015; Anderson et al., 2005). The experiment conducted by Pertaub et al. (2002) concluded that speakers reacted in a similar way whether the audiences were virtual or real. This led to some researchers to create audiences with certain behaviors assigned manually. Fukuda et al. (2017) 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. (2016) conducted a similar experiment building a virtual audience with subtle variations in their reactions.

More recently, Chollet et al. (2015) 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 (Chollet et al., 2015) 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, 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 (Batrinca et al., 2013), 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 speaker's 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 (2005), 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 (Sauter et al., 2010). 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 (Peterson, 2005), 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.

Objectives

The ultimate goal, as said above, 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 (Buschmeier & Kopp, 2018). Public speaking is a one-to-many dialogue where the audience takes the place of the listener. Thus, the audience feedback should lead the speech of the good public speaker.

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.

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.

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), allowing the speakers to train their speech and develop their skills to better cope with this type of situations in real life (see Fig. 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, a place where they can fail. As Johan Huizinga pointed out in his Homo Ludens (Huizinga, 2014), 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 face a 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

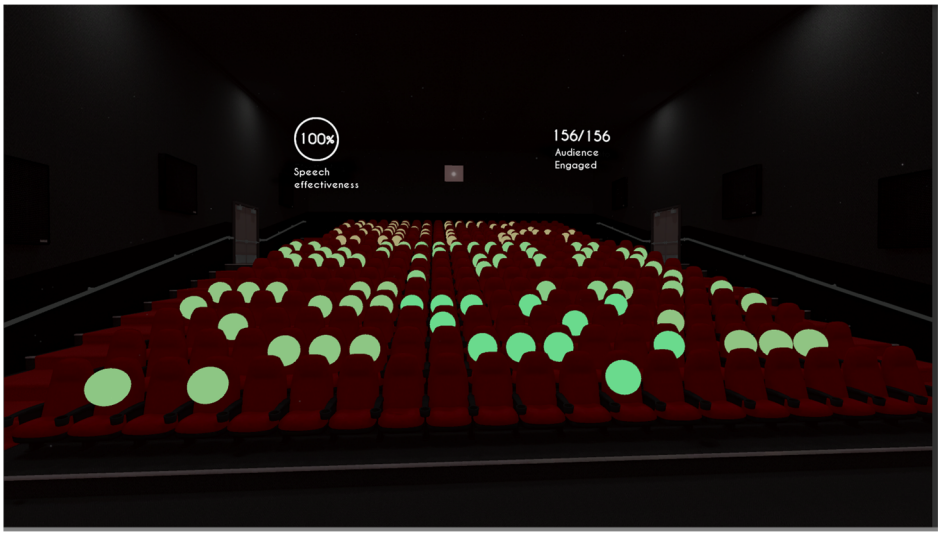

Fig. 1 Game environment. Screen capture of the game environment

speakers to prepare in advance and modify their speeches to improve the virtual audience reaction.

Well-implemented gamification improves the learning process (Kapp, 2012). 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 (Fernández Vara, 2009).

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 speaker's 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 speaker's voice, 2) the content of the speech, and 3) the speaker's 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.

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 (Vinayagamoorthy et al., 2005) that show that characters apearance in a virtual world or in a game do not necessarily have to be realistic. The important point 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 character's appearance is not realistic.

Reactive audience

In this section we will first detail the behavior of an ACM. We will also describe the voice parameter and how we use it in order to extract emotions.

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 speaker's 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 Fig. 2.

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 the data provided by the API, 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 speaker's 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.

Fig. 2 ACM Behaviour. ACM color variations according to their reaction

Given that $AudioEffectiveness = AE, AudioWeight = AW, TextEffectiveness = TE, TextWeight = TW, FocusEffectiveness = FE, FocusWeight = FWandACMSeverity = AS$, the ACM reaction is generated as follows:

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

In the next Section, we describe how we calculate the effectiveness percentages formula for each one of the features(*AE, TE and FE*). 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 (Carniegie, 2017), 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 speaker's gaze direction affects the agent reaction in a 20%.

Audio feature

We iteratively record a short audio fragment (5 - 10 seconds) of the speaker's 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, 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 (see Table 1), with values ranging between 0 and 100, which indicate how likely is that emotion to produce an effective speech. This weights were at first based on the authors common sense.

In addition to the emotion detected, the third-party Emotion Analysis API provides a Confidence Score (CS), 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

Table 1 Emotion weights base on voice tone

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

First approach

that the purpose of the reactive virtual audience is to resemble in as much as possible the reactions of a real audience.

Therefore, the main objective of the experiment that we detail next, is to see the degree of similarity between the reactions of a real audience, and the reactions of the agents of the virtual audience, generated based only on the voice factor.

Experiment

Through this section we will detail the objective, participants, instruments, design of the experiment, and the results that helped us to improve our virtual audience. At the end of the section we will discuss the achieved results.

Objective

The aim of this experiment was to adjust the weights of each emotion extracted from voice tone 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 were actors who were in charge of interpreting in front of the real audience speeches with emotion changes.

Experimental design

To achieve our goal, we compare the reactions generated by our virtual audience (VA) with the reactions of a real audience (RA). The experiment worked as follows: first, one actor gave a 5 to 7 min prepared speech playing different emotions along it. The actor was in a stage in front of the audience, with a chronometer in the background. During that speech, the audience was asked to: 1) Write down the exact time in which they detect any change in the emotion transmitted by the actor (the chronometer in the room is to that aim), 2) Identify the emotion present in the actor speech -among a set of 5 possible emotions-, and 3) Evaluate if they were engaged and the effectiveness of the speech on a scale of 1 to 10 at that precise moment. Later, the actors repeated this design 4 times playing different emotions. Figure 3 illustrates this design.

Instruments and materials

For the purposes of this experiment, we use only the analysis environment of our tool, an API that can extract from audio features an emotion, and generate a reaction on the virtual audience based on that analysis. We also use a microphone and a video camera to record the actor performances. Each member of the audience had a questionnaire (see Table 2) to fill. It aims to know how the actors speeches affected the individuals in the audience, in order to allow a comparison with our virtual audience reactions.

The questionnaire measures three main values: the second where there is a change of emotion in the speech, engagement of the audience member with the speech at that time and the emotion perceived from the speech.

We have used abbreviations to qualify groups of emotions that are showed in Table 3.

As Table 2 summarizes, we aggregated the values of the questionnaire as follows: for the time value, we established 15-second time slots to count as single changes; for the engagement question ('Yes' or 'No' answer), we summed the values, where each 'Yes' counted as

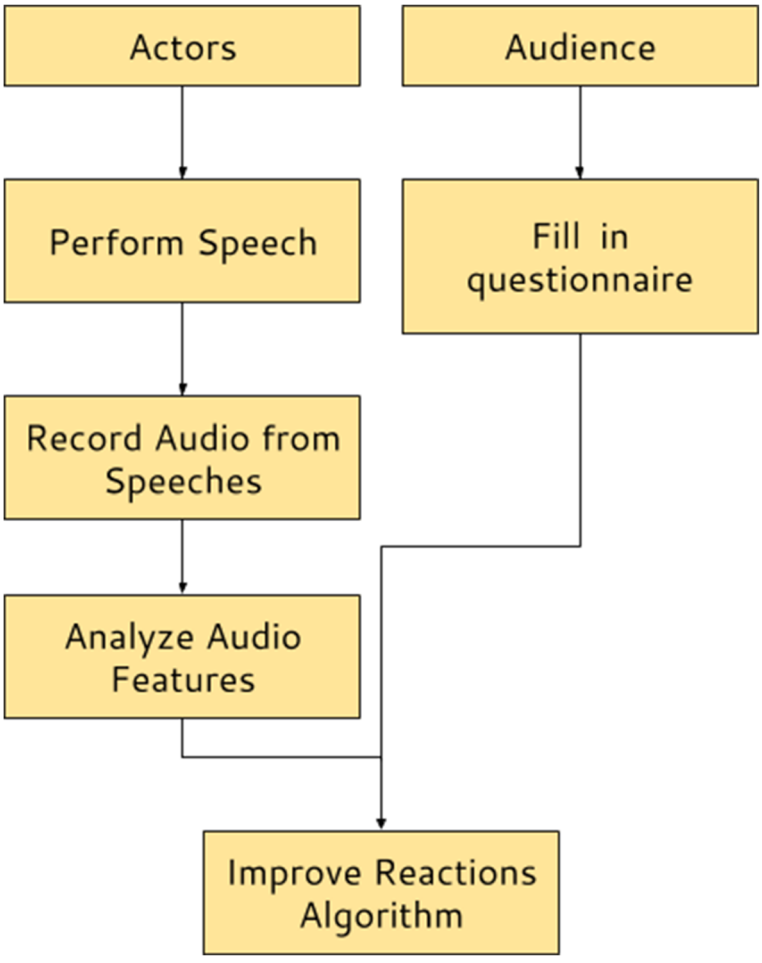

Fig. 3 Experimental design. Diagram of the experimental design

1 and each 'No' as 0; for the engagement level (from 0 to 10), we aggregated the results with an average of all of them; and, for the emotion value, we aggregated the results by doing a mode over all the emotions detected by the audience.

Results

Since not all participants detect and record an emotional change at the same time, we compared the results by establishing 15-second slots. That is, if one participant in the audience annotates a change in the second 0:16, and another does it in 0:26, it is considered that both have detected the same change.

Table 2 Experiment questionnaire

Measured variable	Type	How is it calculated	Range
Change second	Time in M:S	We use 15-second slots to detect single changes	00:00 - 07:00
Engagement (Engaged?)	Yes or No	Each Yes adds 1 to result	0 - 1
Engagement (Level)	One choice	Average value	1 - 10
Emotion	Multiple choice	All values Mode	A, C, E, F, N

Table 3 Emotion groups abbreviations

Tag	Emotion group
A	Bored / Insecure / Sad
C	Calm
E	Stress / Angry / Aggressive
F	Happy / Enthusiastic / Friendly
N	Neutral

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.

Table 4 shows the intermediate results that we obtained for the speech of the first actor. In this table, we reflect the aggregated responses from all the members of the real audience (RA) during the experiment. This responses are the engagement -if they were engaged or not, and on what level from 1 to 10- and the perceived group of emotions from the speech. The groups of emotions are labeled as shown in Table 3 (The value N/A indicates that there was no answer or not sufficient information to calculate the value).

In order to compare between the emotion values reflected by the RA and the ones calculated by the VA, the table also includes a column with the group of emotions perceived by the virtual audience of our tool (VA), and extract conclusions on how emotions affect the level of engagement on the audience.

The rest of the results for the other three speeches are listed in Tables 5, 6 and 7 at the end of this paper.

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. 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 emotions that were similar to the one transmitted in the audio fragment.

This experiment also shows that each member of our 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.

Table 4 Experiment: Results from speech #1

Timestamp	*Engagement* Engaged?	*Engagement* Level	Real Audience	Virtual Audience
0:15	N/A	N/A	C	E
0:30	N/A	6,75	F	C
0:45	N/A	N/A	F	F
1:00	N/A	N/A	F	F
1:15	N/A	7,33	A	N
1:30	N/A	7,33	N	N
1:45	N/A	8,75	E	E
2:00	N/A	6,75	N/A	E
2:15	N/A	7,4	A	N
2:30	N/A	N/A	A	E
2:45	N/A	N/A	F	E
3:00	N/A	7	F	F

Table 5 Experiment: Results from speech #2

Timestamp	*Engagement* Engaged?	*Engagement* Level	Real Audience	Virtual Audience
0:15	Sí	8,75	E	E
0:30	Sí	4	A	E
0:45	N/A	N/A	N/A	N/A
1:00	N/A	6	C	N/A
1:15	N	N/A	N	E
1:30	N/A	5	N/A	N/A
1:45	N/A	7,5	N/A	N/A
2:00	Sí	7,5	F	E
2:15	N/A	5	F	E
2:30	N/A	N/A	N/A	E
2:45	Sí	7,5	F	E
3:00	N	N/A	E	E
3:15	Sí	7	E	E
3:30	N	5,8	A	E
3:45	N	7,66	A	C
4:00	N	5	N/A	N/A
4:15	N	N/A	N/A	N/A
4:30	N	4	A	C
4:45	N/A	N/A	N/A	C
5:00	Sí	7	A	E
5:15	Sí	5,57	A	C
5:30	N/A	5	N/A	N/A
5:45	N/A	4	A	E
6:00	N	5	N	N
6:15	N	2,5	A	N
6:30	N	9	A	N

Table 6 Experiment: Results from speech #3

Timestamp	*Engagement* Engaged?	*Engagement* Level	Real Audience	Virtual Audience
0:15	Sí	7,6	A	E
0:30	Sí	N/A	A	E
0:45	Sí	6,5	C	E
1:00	Sí	7,14	F	F
1:15	Sí	4	C	N/A
1:30	N/A	7,33	F	N
1:45	N	3	N	N
2:00	N	6,75	A	N
2:15	N	5	A	E
2:30	N/A	3,33	A	E
2:45	Sí	7,5	E	E
3:00	N/A	8,66	E	E
3:15	Sí	8	E	F
3:30	Sí	7,71	A	A
3:45	Sí	N/A	A	A
4:00	N/A	9	N/A	N/A
4:15	Sí	7,4	C	N
4:30	Sí	8	C	N
4:45	Sí	5,66	C	N
5:00	N/A	8,8	C	C
5:15	Sí	5,57	A	A

Table 7 Experiment: Results from speech #4

Timestamp	*Engagement* Engaged?	*Engagement* Level	Real Audience	Virtual Audience
0:15	N/A	6,50	A	A
0:30	Sí	6,50	N	N
0:45	N/A	N/A	N	N/A
1:00	N	4,50	A	C
1:15	N	3,50	A	A
1:30	N/A	4,50	N/A	N/A
1:45	N/A	8,00	N/A	N/A
2:00	Sí	6,67	N	N
2:15	N	3,33	N	E
2:30	N	N/A	N	E
2:45	N	5,33	A	N
3:00	N/A	3,50	N	E
3:15	N/A	3,00	N/A	N/A
3:30	Sí	9,00	F	C
3:45	N/A	N/A	F	N
4:00	N/A	N/A	F	E
4:15	Sí	7,17	F	E
4:30	Sí	N/A	F	E
4:45	Sí	7,50	N/A	N/A
5:00	N/A	N/A	N/A	N/A
5:15	N/A	N/A	N/A	N/A
5:30	Sí	8,00	F	E
5:45	N/A	N/A	N/A	N/A
6:00	Sí	7,00	C	C
6:15	Sí	7,50	E	E
6:30	Sí	8,50	N/A	E
6:45	N/A	6,00	E	E

Discussion

Based on the results obtained, we changed the values of emotion weights in order to get a more behavioural-realistic audience in our virtual system. New values are shown in the Table 8. These weights have been adjusted taking into account the engagement values reflected by the real audience to the various emotions detected. As we saw in Table 1, the weights were based on common sense. After the experiment, we were able to see how certain emotions engaged the real audience, and according to those engagement values, we modified the emotion weights.

As an 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

Table 8 Emotion weights base on voice tone

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

Second approach

assumed it was counterproductive 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 five -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 the virtual audience agents have been seeing and hearing during a speaker's speech, they can learn and react differently to the same emotion transmitted.

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 behavioural-realistic reactive audiences is critical in order to achieve effective environments to train public speaking. First, we have presented a virtual reality video game 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.

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 speaker's 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.

One of the most important factors that influences a public speaking performance is non-verbal behaviour, mainly body language. There is work on body language recognition, but it is in very early stages. Also, it is very complex to recognize body language in a period of time, in order to detect gestures. Because of that, currently, the researchers of this project are working on extending the features analyzed to detecting and extracting emotion from body language features, such as gestures or pose.

Thus, 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.

Acknowledgments
Not applicable.

Authors' contributions
The main author (ME) designed and developed the tool described in this paper and performed the research and the writing of this paper. AR, MG and BM provided insight, edited and reviewed the paper. All authors read and approved the final manuscript.

Funding
This project has been partially funded by BBVA foundation (ComunicArte-Ayudas Fundación BBVA a Equipos de Investigación Científica 2017: PR2005-174/01) and by the Ministry of Science, Innovation and Universities (Didascalias project: RTI 2018-096401-A100).

Availability of data and materials
The datasets generated and/or analyzed during the current study are not publicly available due to privacy reasons but are available from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Received: 10 September 2019 Accepted: 29 October 2019
Published online: 11 November 2019

References

Anderson, P.L., Zimand, E., Hodges, L.F., Rothbaum, B.O. (2005). Cognitive behavioral therapy for public-speaking anxiety using virtual reality for exposure. *Depression and anxiety*, *22*(3), 156–158.

Aristotle (2009). *Rhetoric*. United States: A & D Publishing.

Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., Scherer, S. (2013). Cicero-towards a multimodal virtual audience platform for public speaking training, In *International Workshop on Intelligent Virtual Agents*. https://doi.org/10.1007/978-3-642-40415-3_10 (pp. 116–128): Springer.

Bos, N., & Shami, N.S. (2006). Adapting a face-to-face role-playing simulation for online play. *Educational Technology Research and Development*, *54*(5), 493–521.

Brom, C., Preuss, M., Klement, D. (2011). Are educational computer micro-games engaging and effective for knowledge acquisition at high-schools? a quasi-experimental study. *Computers & Education*, *57*(3), 1971–1988.

Buschemeier, H., & Kopp, S. (2018). 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 (AAMAS 2018), Stockholm, Sweden (pp. 1213–1221).

Carniegie, D. (2017). *How to Develop Self-confidence and Influence People by Public Speaking*. Gallery Books, United States: Simon and Schuster.
Chollet, M., Wörtwein, T., Morency, L.-P., Shapiro, A., Scherer, S. (2015). 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. UbiComp '15*. https://doi.org/10.1145/2750858.2806060 (pp. 1143–1154). New York: ACM.
Chollet, M., Stefanov, K., Prendinger, H., Scherer, S. (2015). Public speaking training with a multimodal interactive virtual audience framework, In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15*. https://doi.org/10.1145/2818346.2823294 (pp. 367–368). New York: ACM.
Delingette, H., Cotin, S., Ayache, N. (1999). Efficient linear elastic models of soft tissues for real-time surgery simulation. *Medicine Meets Virtual Reality: The Convergence of Physical & Informational Technologies: Options for a New Era in Healthcare*, *62*, 100.
Fernández Vara, C. (2009). The tribulations of adventure games: integrating story into simulation through performance, PhD thesis.
Fukuda, M., Huang, H.-H., Ohta, N., Kuwabara, K. (2017). Proposal of a parameterized atmosphere generation model in a virtual classroom, In *Proceedings of the 5th International Conference on Human Agent Interaction, HAI '17*. https://doi.org/10.1145/3125739.3125776 (pp. 11–16). New York: ACM.
Gratacós, M. (2018). Glossophobia. Do you suffer from glossophobia? http://http://www.glossophobia.com/. Accessed 10 Nov 2018.
Halan, S., Sia, I., Miles, A., Crary, M., Lok, B. (2018). Engineering social agent creation into an opportunity for interviewing and interpersonal skills training: Socially interactive agents track, In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden (pp. 1675–1683).
Hwang, G.-J., Wu, P.-H., Chen, C.-C. (2012). An online game approach for improving students' learning performance in web-based problem-solving activities. *Computers & Education*, *59*(4), 1246–1256.
Huizinga, J. (2014). *Homo Ludens Ils 86*: Routledge. https://doi.org/10.4324/9781315824161.
Kang, N., Brinkman, W.-P., Birna van Riemsdijk, M., Neerincx, M. (2016). The design of virtual audiences. *Computers in Human Behavior*, *55*(PB), 680–694. https://doi.org/10.1016/j.chb.2015.10.008.
Kapp, K.M. (2012). *The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education*. San Francisco: Pfeiffer: Wiley.
Laukka, P., Juslin, P., Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition & Emotion*, *19*(5), 633–653.
López, J.M.C., & Cáceres, M.J.M. (2010). Virtual games in social science education. *Computers & Education*, *55*(3), 1336–1345.
Lowrie, T., & Jorgensen, R. (2011). Gender differences in students' mathematics game playing. *Computers & Education*, *57*(4), 2244–2248.
Mehrabian, A. (2017). *Nonverbal Communication*. New York: Routledge, Taylor and Francis Group.
M. North, M., M. North, S., R. Coble, J. (2015). Virtual reality therapy: an effective treatment for the fear of public speaking. *International Journal of Virtual Reality (IJVR)*, *03*(3), 1–6.
Page, E.H., & Smith, R. (1998). Introduction to military training simulation: a guide for discrete event simulationists, In *1998 Winter Simulation Conference. Proceedings (Cat. No. 98CH36274), vol. 1*. https://doi.org/10.1109/wsc.1998.744899 (pp. 53–60): IEEE.
Papastergiou, M. (2009). Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Computers & education*, *52*(1), 1–12.
Pertaub, D.-P., Slater, M., 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. https://doi.org/10.1162/105474602317343668.
Peterson, R.T. (2005). An examination of the relative effectiveness of training in nonverbal communication: Personal selling implications. *Journal of Marketing Education*, *27*(2), 143–150.
Sauter, D.A., Eisner, F., Calder, A.J., Scott, S.K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology*, *63*(11), 2251–2272.
Tüzün, H., Yılmaz-Soylu, M., Karakuş, T., İnal, Y., Kızılkaya, G. (2009). The effects of computer games on primary school students' achievement and motivation in geography learning. *Computers & Education*, *52*(1), 68–77.
Vinayagamoorthy, V., Steed, A., Slater, M. (2005). Building characters: Lessons drawn from virtual environments, In *Proceedings of Toward Social Mechanisms of Android Science: A CogSci 2005 Workshop* (pp. 119–126). Stressa, Italy: COGSCI 2005.

Publisher's Note