Spontaneous emotional speech recordings through a cooperative online video game

Daniel Palacios-Alonso, Victoria Rodellar-Biarge,
Victor Nieto-Lluis, and Pedro Gómez-Vilda

Centro de Tecnología Biomédica and
Escuela Técnica Superior de Ingenieros Informáticos
Universidad Politécnica de Madrid
Campus de Montegancedo - Pozuelo de Alarcón - 28223 Madrid - SPAIN
email:daniel@junipera.datsi.fi.upm.es

Abstract. Most of emotional speech databases are recorded by actors and some of spontaneous databases are not free of charge. To progress in emotional recognition, it is necessary to carry out a big data acquisition task. The current work gives a methodology to capture spontaneous emotions through a cooperative video game. Our methodology is based on three new concepts: novelty, reproducibility and ubiquity. Moreover, we have developed an experiment to capture spontaneous speech and video recordings in a controlled environment in order to obtain high quality samples.

Keywords: Spontaneous emotions; Affective Computing; Cooperative Platform; Databases; MOBA Games

1 Introduction

Capturing emotions is an arduous task, above all when we speak about capturing and identifying spontaneous emotions in voice. Major progress has been made in the capturing and identifying gestural or body emotions [1]. However, this progress is not similar in the speech emotion field. Emotion identification is a very complex task because it is dependent on, among others factors, culture, language, gender and the age of the subject. The consulted literature mentions a few databases and data collections of emotional speech in different languages but in many cases this information is not open to the community and not available for research. There is not an emotional voice data set recognized for the research community as a basic test bench, which makes a real progress in the field very complicated, due to the difficulty in evaluating the quality of new proposals in parameters for characterization and in the classification algorithms obtained using the same input data. To achieve this aim, we propose the design of a new protocol or methodology which should include some features such as *novelty*, *reproducibility* and *ubiquity*.

Typically, emotional databases have been recorded by actors simulating emotional speech. These actors read the same sentence with different tones. In our research, we have requested the collaboration of different volunteers with different ages and gender. Most of volunteers were students who donated their voices. First of all, they had to give their consent in order to participate in our experiment. Therefore, we show a **novel** way of obtaining new speech recordings. The next key feature for this task is **reproducibility**, where each experiment should provide spontaneity, although the exercise was repeated a lot of times. This characteristic is the most important drawback we have found in the literature. Most of the time, when it carries out an experiment, this user is discarded immediately, because he/she knows perfectly the guideline of the exercise, for this reason the spontaneity is deleted. The third feature is, **ubiquity**. When we speak about this concept, we refer to carrying out the exercise in every part of the world, but that does not mean that we cannot use the same location or devices. Nowadays, new technologies such as smart-phones, tablets and the like are necessary allies in this aspect.

In view of all the above, multiplayer videogames are the perfect way to achieve the last three premises. Each game session or scenario can be different. Moreover, we can play at home, in a laboratory or anywhere. Thanks to the Internet, we can find different players or rivals around the world who speak other languages and have other cultures, etcetera. Each videogame has its own rules, thus each player knows the game system and they follow these rules if they want to participate. For this reason we find the standardization feature intrinsic in the videogames. Therefore, we conclude videogames are the perfect tool in order to elicit spontaneous emotions.

This research has two main stages; they consist of capturing emotions through a videogame, more specifically League of Legends (aka LoL), and identify the captured emotions through the new cooperative framework developed by our team. To assess the viability of our protocol, we have developed a controlled experiment in our laboratory. In subsequent sections, it will be explained in detail.

The contribution of this work is to establish a community to cooperate in collecting, developing and analyzing emotional speech data and define a standard corpus in different languages where the main source of samples will be emotional speeches captured through videogame rounds. In this sense, this paper is a first step in proposing the design and development of an online cooperative framework for multilingual data acquisition of emotional speech. This paper is organized as follows. In the next section, we introduce some emotional databases of speech and foundations for modeling emotions in games. In section III, we introduce the proposed experiment. And finally, we conclude with the summary and future works.

2 Previous Works

Below, we present previous works carried out by different researchers who have focused their attention in emotional areas. Some of them have elaborated emotional databases, others have developed affective models to improve the realism of NPCs (Non Player Character) or have attempted to verbalize certain situations that happen for game rounds, etc. We are going to attempt to find a common ground between using videogames and the design of a protocol to capture emotions through voice.

2.1 Emotions in Videogames

According to [2] there exists a lack a common, shared vocabulary that allows us to verbalize the intricacies of game experience. For any field of science to progress, there needs to be a basic agreement on the definition of terms. This concept is similar to the lack of agreement for the relevant features in order to characterize and classify speech emotions. They define two concepts, *flow* and *immersion*. Flow can be explained as an optimal state of enjoyment where people are completely absorbed in the activity. This experience was similar for everyone, independent of culture, social class, age or gender. This last assertion is a key point for us, because we are searching for the most suitable method or protocol for anyone. Immersion is mostly used to refer to the degree of involvement or engagement one experiences with a game. Regarding *arousal* and *valence*, *Lottridge* has developed a novel, continuous, quantitative self-report tool, based on the model of valence and arousal which measures emotional responses to user interfaces and interactions [3]. Moreover, *Sykes and Brown* show the hypothesis that the player's state of arousal will correspond with the pressure used to press buttons on a gamepad [4].

Concerning elicit emotion and emotional responses to videogames, in [5] presents a comprehensive model of emotional response to the single-player game based on two roles players occupy during gameplay and four different types of emotion. The emotional types are based on different ways players can interact with a videogame: as a simulation, as a narrative, as a game, and as a crafted piece of art.

On the other hand, some researchers focus their attention on psychophysiological methods in game research. *Kivikangas et al.* carry out a complete review of some works in relation with these kind of methods. They present the most useful measurements and their effects in different research areas such as game effects, game events, game design and elicited emotions. Electromyography (EMG), Electrodermal activity (EDA), Heart Rate (HR), among others, are some of these measurements [6].

Another initiative was developed by [7], designing requirements engineering techniques to emotions in videogame design, where they introduced emotional

terrain maps, emotional intensity maps, and emotional timelines as in-context visual mechanisms for capturing and expressing emotional requirements.

Regarding emotion modeling in game characters, *Hudlicka and Broekens* present theoretical background and some practical guidelines for developing models of emotional effects on cognition in NPCs [8]. In [9] have developed a toolkit called the *Intelligent Gaming System (IGS)* that is based on *Command* (Atari, 1980). The aim was to keep engagement as measured by changing heartbeat rate, within an optimum range. They use a small test group, 8 people, whose experience was documented and thanks to their conclusions, they could design a theory of modes of affective gaming.

2.2 Emotional Databases

Most databases have been recorded by actors simulating emotional discourses and there are a very few of them of spontaneous speech [10], [11]. The emotions are validated and labeled by a panel of experts or by a voting system. Most of the databases include few speakers and sometimes they are gender [12] unbalanced, and most of recorded data do not consider age. Then they restrict to carrying out research related with subject age range [13]. It can be noticed in several publications that the data are produced just for specific research, and the data are not available for the community. Some databases related to our research are briefly mentioned next.

Two of the well-known emotional databases for speech are the Danish Emotional Speech Database (DES) [14] and the Berlin Emotional Speech Database (BES) [15] in German. BES database, also known as Emo-Database, is spoken by 10 professional native German actors, 5 female and 5 male. It includes the emotions of neutral, anger, joy, sadness, fear, disgust and boredom. The basic information is 10 utterances, 5 short and 5 longer sentences, which could be used in daily communication and are interpretable in all applied emotions. The recorded speech material was around 800 sentences. All sentences have been evaluated by 20-30 judges. Those utterances for which the emotion was recognized by at least 80% of the listeners will be used for further analysis. DES database is spoken by four actors, 2 male and 2 female. It contains the emotions of neutral, surprise, happiness, sadness and anger. Records are divided into simple words, sentences and passages of fluent speech.

Concerning stress in speech, the SUSAS English database (Speech Under Simulated and Actual Stress) is public and widely used [16]. It contains a set of 35 aircraft communication words, which are spoken spontaneously by aircraft pilots during a flight, and also contains other samples of non-spontaneous speech.

Finally, the work closest to our approach that we have found in literature, has been in [17]. They have developed an annotated database of spontaneous, multimodal, emotional expressions. Recordings were made of facial and vocal

expressions of emotions while participants were playing a multiplayer first-person shooter (fps) computer game. During a replay session, participants scored their own emotions by assigning values to them on an arousal and a valence scale, and by selecting emotional category labels.

3 Affective Data Acquisition

As mentioned before, we used a videogame-like source of elicited emotions. The chosen game was *League of Legends* (aka LoL). To carry out this task, it was necessary to organize a little tournament, where the team with the best score at the end of the tournament, obtained a check for the amount of 20 € per person as well as a diploma. With the obtained samples, we attempt to find correlates between acoustics, glotals or biomechanics parameters and the elicited emotions. To extract these parameters, we have used [18].

3.1 The Subjects

The subjects are students of Computer Science at the Universidad Politécnica of Madrid. Apparently, students had not got any disease in their voices and they gave their explicit consent in order to participate in the experiment.

3.2 The Game

LoL is a multiplayer online battle arena (MOBA) video game developed and published by Riot Games. It is a free-to-play game supported by micro-transactions and inspired by the mod Defense of the Ancients for the video game Warcraft III: The Frozen Throne. League of Legends was generally well received at release, and it has grown in popularity in the years since. By July 2012, League of Legends was the most played PC game in North America and Europe in terms of the number of hours played [19]. As of January 2014, over 67 million people play League of Legends per month, 27 million per day, and over 7.5 million concurrently during peak hours [20].

3.3 The Environment

Each of the game sessions are carried out in our Laboratory, Neuromorphic Speech Processing Laboratory, which belongs to R+D group Centro de Tecnología Biomédica. In this laboratory, we possess a quasi-anechoic chamber that is designed in order to entirely absorb the acoustic waves without echoing off any surface of the chamber such the floor, roof, walls and the like. The chamber has a personal computer inside, where a player remains throughout the game round. Outside of the chamber, there will be another four personal computers at the disposal of four members of the rest of the team. Concerning the choice of the anechoic chamber, it is easy to understand that we are looking for ideal conditions in order to develop of following stages such as characterization and extraction of parameters, selection of parameters, classification and finally, detection of emotions in speech [21].

3.4 Hardware Features

Each player has used a SENNHEISER PC 131 headset with wire connector, 30 - 18000 Hz of headphone frequency, 80 - 15000 Hz of microphone frequency, 2000 Ohms of output impedance and 38 Db of sensitivity. On the other hand, computers used in the experiment had the following features. Intel Core 2 Quad - CPU Q6600, 4 cores up to 2.40 Ghz, 4 GB of RAM and Nvidia GForce 8600GT with 512 MB Graphics Card. Moreover, it has been connected to a webcam in order to record faces and gestures for each game session. Video recordings could be crucial in order to recognize in the following stages the saved emotion.

3.5 The Experiment

During two weeks, we will convene ten subjects in two shift sessions. The will be one shift in the morning and another one in the afternoon. Each session will have five players, where each player will log in to LoL's website with his/her official account. Four of five players remain together in the same room, whereas the remaining player will be isolated inside of anechoic chamber. Approximately, each session will be limited to three hours and a half, because the average length of a game is 40 minutes. Once a game is over, the player who stays inside the chamber, will come out the chamber and the following mate takes his place. This continues until five rounds have been completed. Therefore, each day of tournament, we will have recorded 10 different players for 40 minutes. These records will be saved in our raw speech recordings database. Each team will play a maximum of 3 games over the two weeks of tournament. The process of the experiment is depicted in the Fig. 1. This experiment and others, which although they are not the objective of this paper to analyze them deeply, have been developed in order to elicit spontaneous emotions by our team. These experiments are incorporated inside the framework *Emotions Portal*.

3.6 The Framework

The aim of this experiment is the capturing of speech recordings through a cooperative online videogame. The idea is that our server collects spontaneous emotional voice in different languages, with different accents and origins, etc. This framework can be defined as cooperative, scalable or modular and not subjective. According to Fig. 2, we divide our online framework into four stages: *User identification, Start of the recording, Play the game and End of the recording and save the speech and video Recording*. At the top of the picture is depicted the player who is inside of the anechoic chamber. He/she is connected to our platform which is deployed in our server. The user identification step consists of a sign up process through a web form. In this web form, users provide their personal data, for instance, their name, native language, country, gender, age and email. The last requirement, email address, is convenient in order to have the chance to keep in touch with the user and to give him/her information about the progress of the project. Once logged in to our platform, the user can choose

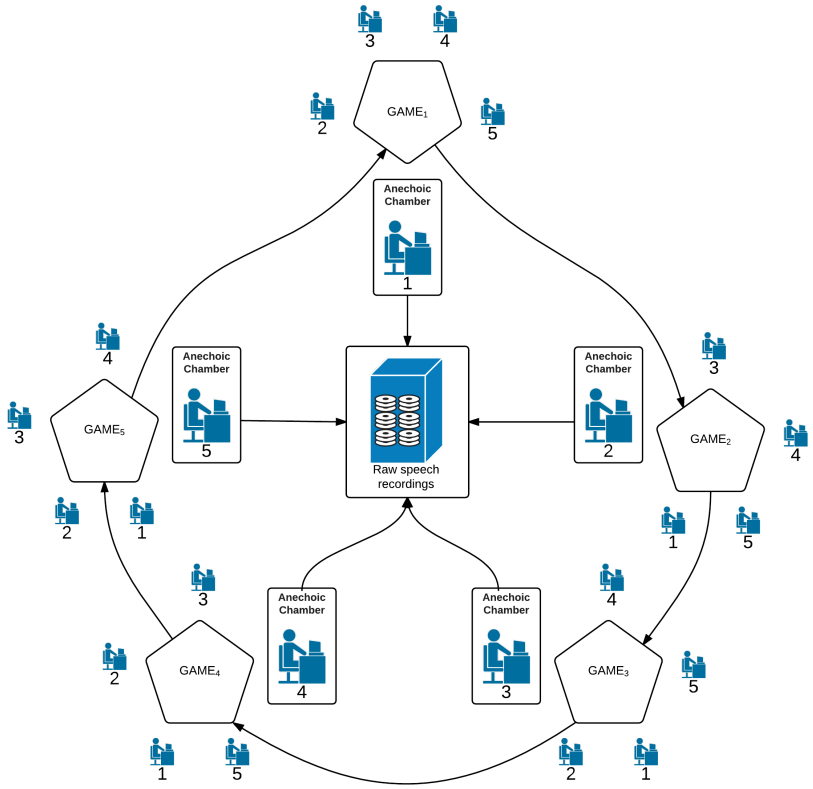

Fig. 1. Scenario of Experiment.

the different kind of experiments which are available. The actions mentioned before are explained as follows.

First of all, users answer if they suffer from any type of disease in their voice at the moment of performing the test. It is crucial to know if the user suffers from any organic or functional dysphonia, diseases that can affect voice and prevent the use of biometric techniques. Then, the player chooses the language of procedure. When the user is ready to start playing, he/she presses the *Start* button. In this moment, the platform begins recording and throws a call to LoL's application. LoL's platform opens and the user carries out the process of sign in with his/her LoL's account. These steps are depicted through numbers 1 and 2 in Figure 2. Approximately 40 minutes later, the game round is over. The player signs out of the LoL's platform and he/she presses the *Stop* button in our platform. Finally, the speech and video recording are saved in our repository inside of our server. The last steps are depicted through numbers 3 and 4 in Figure 2.

4 Summary

The spontaneous data acquisition is a very complex topic to resolve. However, it is the key point to improve human computer interfaces, robots, video games and the like. As mentioned before, some researchers have developed new models and methodologies to improve NPCs, and others have researched new designs to evoke certain emotions during the game sessions. We have developed a methodology in order to capture spontaneous emotions through a cooperative videogame with high quality audio in a controlled environment. Afterwards, we will be able to design a well-labeled database and continue with our previous work on emotion recognition.

Acknowledgments. This work is being funded by grants TEC2012-38630-C04-01 and TEC2012-38630-C04-04 from Plan Nacional de I+D+i, Ministry of Economic Affairs and Competitiveness of Spain.

References

1. P. Ekman, Handbook of cognition and emotion. Wiley Online Library, 1999, ch. Basic emotions, pp. 45–60.
2. W. IJsselsteijn, Y. De Kort, K. Poels, A. Jurgelionis, and F. Bellotti, "Characterising and measuring user experiences in digital games," in International conference on advances in computer entertainment technology, vol. 2, 2007, p. 27.
3. D. Lottridge, "Emotional response as a measure of human performance," in CHI'08 Extended Abstracts on Human Factors in Computing Systems. ACM, 2008, pp. 2617–2620.
4. J. Sykes and S. Brown, "Affective gaming: measuring emotion through the gamepad," in CHI'03 extended abstracts on Human factors in computing systems. ACM, 2003, pp. 732–733.

5. J. Frome, "Eight ways videogames generate emotion," Obtenido de http://www.digra. org/dl/db/07311.25139. pdf, 2007.
6. J. M. Kivikangas, G. Chanel, B. Cowley, I. Ekman, M. Salminen, S. Järvelä, and N. Ravaja, "A review of the use of psychophysiological methods in game research," Journal of Gaming & Virtual Worlds, vol. 3, no. 3, 2011, pp. 181–199.
7. D. Callele, E. Neufeld, and K. Schneider, "Emotional requirements in video games," in Requirements Engineering, 14th IEEE International Conference. IEEE, 2006, pp. 299–302.
8. E. Hudlicka and J. Broekens, "Foundations for modelling emotions in game characters: Modelling emotion effects on cognition," in Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on. IEEE, 2009, pp. 1–6.
9. K. Gilleade, A. Dix, and J. Allanson, "Affective videogames and modes of affective gaming: assist me, challenge me, emote me," DiGRA - Digital Games Research Association, 2005.
10. S. Ramakrishnan, "Recognition of emotion from speech: a review," Speech Enhancement, Modeling and recognition–algorithms and Applications, 2012, p. 121.
11. D. Ververidis and C. Kotropoulos, "A review of emotional speech databases," in Proc. Panhellenic Conference on Informatics (PCI), 2003, pp. 560–574.
12. V. Rodellar, D. Palacios, P. Gomez, and E. Bartolome, "A methodology for monitoring emotional stress in phonation," in Cognitive Infocommunications (CogInfoCom), 2014 5th IEEE Conference on. IEEE, 2014, pp. 231–236.
13. C. Muñoz-Mulas, R. Martínez-Olalla, P. Gómez-Vilda, E. W. Lang, A. Álvarez-Marquina, L. M. Mazaira-Fernández, and V. Nieto-Lluis, "Kpca vs. pca study for an age classification of speakers," in Advances in Nonlinear Speech Processing. Springer, 2011, pp. 190–198.
14. I. S. Engberg and A. V. Hansen, "Documentation of the danish emotional speech database DES," Internal AAU report, Center for Person Kommunikation, Denmark, 1996.
15. F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech." in Interspeech, vol. 5, 2005, pp. 1517–1520.
16. J. H. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom, "Getting started with SUSAS: a speech under simulated and actual stress database." in Eurospeech, vol. 97, no. 4, 1997, pp. 1743–46.
17. P. Merkx, K. P. Truong, and M. A. Neerincx, "Inducing and measuring emotion through a multiplayer first-person shooter computer game," in Proceedings of the Computer Games Workshop, 2007, pp. 06–07.
18. "BioMetroPhon - Official Webpage," 2008, URL: http://www.glottex.com/ [accessed: 2015-05-04].
19. J. Gaudiosi. Riot games' league of legends officially becomes most played pc game in the world. [Online]. Available: "http://www.forbes.com/sites/johngaudiosi/2012/07/11/riot-games-league-of-legends-officially-becomes-most-played-pc-game-in-the-world/" (2012)
20. I. Sheer. Player tally for league of legends surges. [Online]. Available: "http://blogs.wsj.com/digits/2014/01/27/player-tally-for-league-of-legends-surges/" (2014)
21. V. Rodellar-Biarge, D. Palacios-Alonso, V. Nieto-Lluis, and P. Gómez-Vilda, "Towards the search of detection in speech-relevant features for stress. expert systems," Expert Systems, 2015.

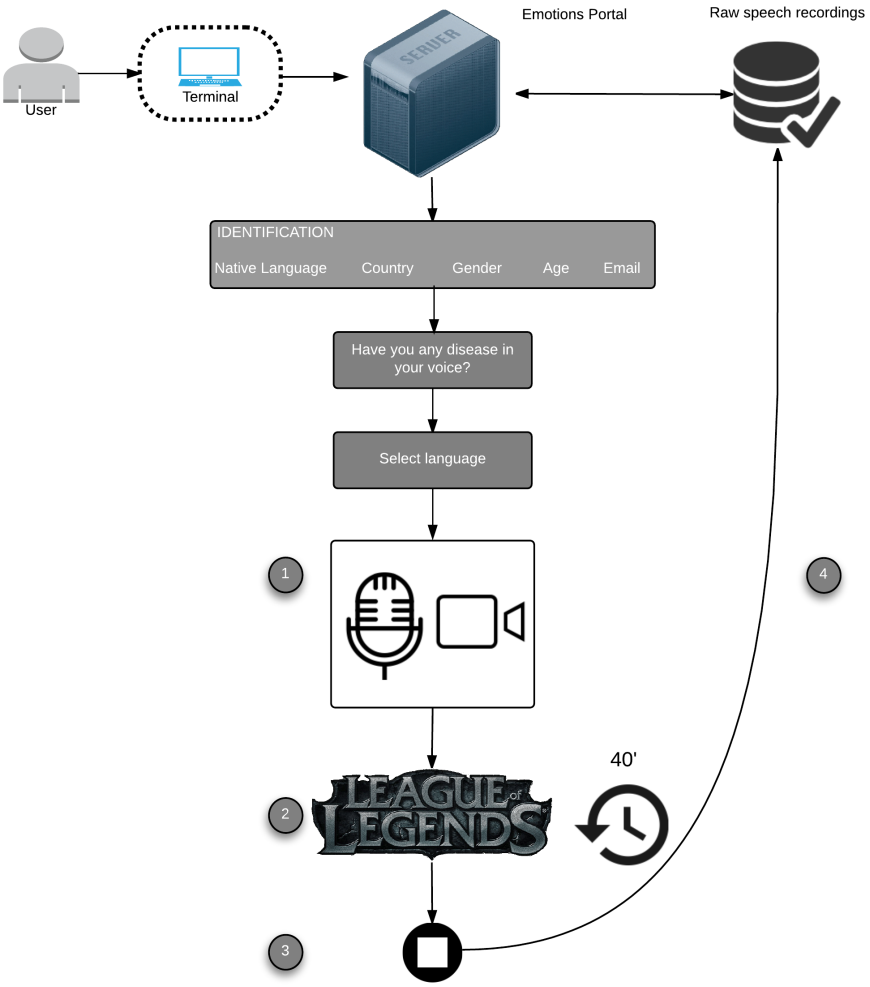

Fig. 2. Scenario of player inside anechoic chamber.