Multi-modal emotion-related data collection within a virtual earthquake emulator

Dimitrios Ververidis, Irene Kotsia, Constantine Kotropoulos and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
Box 451, Thessaloniki 541 24, Greece
E-mail: {jimver, costas}@aiia.csd.auth.gr

Abstract

The collection of emotion-related signals, such as face video sequences, speech utterances, galvanic skin response, and blood pressure from pupils in a virtual reality environment, when they attempt to evacuate a school during an earthquake, is addressed in this paper. We assess whether pupils' emotional state can be accurately recognized.

1 Introduction

A great expectation in human-centered computer interaction has been to exploit user's emotional state recognition as a feedback mechanism in order to adapt computer's response to user needs or preferences (Picard, 2000; Scherer, 2003; Ververidis and Kotropoulos, 2006a). In this paper, we report on an application-driven multi-modal emotion-related corpus collected in a virtual reality (VR) scenario, when pupils attempt to evacuate a school during an emulated earthquake. Several emotion-related bio-signals were recorded, while the pupils were immersed in the virtual earthquake environment. The objective was to assess how immersed is the pupil and how it feels. The data recorded were face videos, speech utterances, galvanic skin response for sweat indication, and blood pressure. Sweat indicator and blood pressure signals have not been adequately studied yet, though related publications have been appeared and patents have been granted for their measurement. A wearable signal sampling unit with sensors mounted on the hand and the foot was developed in (Picard, 2000). Patents for sensors integrated with mouse, keyboard, and joystick have also been granted (Ark and Dryer, 2001).

The entire experiment was designed so that it provides objective evidence in order to evaluate the VR environment developed for training the pupils to cope with earthquakes, which frequently occur in Greece. In this paper, an assessment of the VR environment is presented, that is based on the recorded speech utterances, the sweat indication, and the heart beat rate of the subjects. An algorithm that recognizes the emotional state of a subject from the speech is briefly presented. The algorithm should be trained on speech utterances, whose emotional state is known. Therefore, the experiments are divided into two phases. In the first phase, the pupils learn how to express their emotions. In the second (or evaluation) phase, the pupils express their emotions during the emulated earthquake situation.

The outline of this paper is as follows. The collection of bio-signals is described in Section 2. The classification of utterances into emotional state is accomplished via a Bayes classifier, which is described in Section 3. The assessment of the VR environment is addressed in Section 4. In Section 5, the galvanic skin response signal and the heart beat rate are analyzed. Finally, conclusions are drawn in Section 6.

(a)

(b)

Figure 1: Expression examples.

2 Recording scenario

During the first phase, the pupils learn how to express their emotions. Episodes from several Greek movies were presented to the pupils. Each movie episode contains facial and speech expressions from an actor/actress colored with a certain emotion. By doing so, the pupils are familiarized with their role in the experiments. Two examples are depicted in Figure 1, where two expressions from two different kids are shown.

In detail, 14 pupils were asked to express 13 utterances under 7 emotional states. These utterances are used to train the speech emotion recognition algorithm, since the emotional state for each utterance is known. The pupils who participated in the experiments were of age between 9 and 17 years. The 7 emotional states of the facial expressions and speech utterances are {anger, disgust, fear, happiness, neutral, sadness, and surprise}. The linguistic content of the utterances recorded during the first phase is described in Table 1.

The utterances collected in the first phase are 1396 (i.e. 14 (subjects) $\times$ 7 (states) $\times$ 13 (repetitions) plus some duplicates). In addition, a video capturing the facial expressions for each emotional state was recorded without any utterance by the pupil. During the first phase, 1 video sequence and two speech recordings (one coming from the camera microphone and a another one from a lavaliere microphone) were collected.

This work has been supported by project 01ED312 co-funded by the European Union and the Greek Secretariat of Research and Technology (Hellenic Ministry of Development) of the Operational Program for Competitiveness within the 3rd Community Support Framework.

Table 1: Linguistic content of utterances in Greek and their translation in English appearing inside parentheses.

1	*Αυλή* (Yard)
2	*Διάδρομος* (Corridor)
3	*Έξοδος* (Exit)
4	*Θρανίο* (Desk)
5	*Παιδιά* (Pupils)
6	*Παράθυρο* (Window)
7	*Πόρτα* (Door)
8	*Σχολείο* (School)
9	*Σεισμός* (Earthquake)
10	*Τάξη* (Classroom)
11	*Θα βγούμε στην αυλή αργά* (We shall go slowly out to the yard)
12	*Μπαίνω κάτω από το θρανίο* (I get underneath the desk)
13	*Περιμένω να σταματήσει ο σεισμός* (I wait until the earthquake stops)

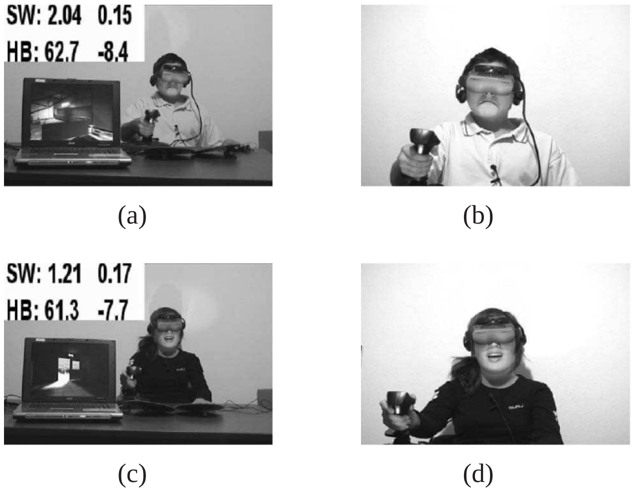

Figure 2: Captured snapshots from distant cameras for two subjects.

In the second phase, the pupils are immersed in a VR earthquake environment that consists of VR glasses and a joystick. The VR environment was developed on the top of the engine of the "Quake" game (Tarnanas et al., 2003). During the earthquake immersion, a virtual teacher (avatar) is giving instructions on how to cope with the situation, e.g. "Wait for the earthquake to stop" or "Proceed carefully to the exit". The objective is to assess how realistic the VR environment is by recording bio-signals recorded from each pupil. The following bio-signals were collected in the second phase:

- 2 video sequences (one sequence capturing the facial expressions and another one recording the VR environment simultaneously with the pupils' expressions so that psychologists could evaluate the pupils' reactions);
- 3 speech recordings (the 2 recordings stem from the camera microphones and the third comes from a lavaliere microphone);
- 1 sweat indicator signal;
- 1 blood pressure signal;

The sweat indicator signal is the electrical conductivity between fingers (Galvanic Skin Response, GSR) when a small electric current is applied. The blood pressure is measured by a plethysmograph, that is a pressure sensor positioned on a finger with a velcro strap. Through the blood pressure, one is able to measure heart beat rate by peak picking. Snapshots from the second phase are shown in Figure 2. In Figures 2(a) and 2(c) frames captured by the distant camera are shown. Sweat (SW) and heart beat (HB) rate (HB) indications at the certain time instant are displayed overlaid in Figures 2(a) and 2(c). A portable PC that shows exactly what the pupil sees was positioned near the kid, so that the distant camera captures both the pupil and the VR scenes. Technical details about the equipment are briefly provided next. 2 PCs were used. The first PC was used to record the sweat signal, the blood pressure, and the speech from the lavaliere microphone. The second PC was running the VR environment. The following peripheral equipment was used: 2 video cameras, 1 data recorder (IWorx-114), 1 pressure sensor (PT-100), 2 electrodes for measuring GSR (GSR-200), 1 sound sampling console (Behringer UB802), 1 condense microphone (AKG C417III), 1 joystick with force feedback, and 1 pair of VR glasses.

3 Classifier training and testing during the first phase

The utterances collected during the first phase are used to train a speech emotion classifier. Speech from three emotional classes was used, namely, fear, happiness, and neutral. A set of 113 statistics of short-term pitch, energy, frequency contours is extracted, as in (Ververidis and Kotropoulos, 2006b). Each class-conditional probability density function of the extracted acoustic features is modeled by a multivariate Gaussian. Based on the aforementioned assumption, the Bayes classifier was designed. In order to find an unbiased estimate of the correct classification rate (CCR) admitted by the Bayes classifier, cross-validation is used, where 90% of the available utterances are exploited to train the classifier and the remaining 10% is used for classifier testing. The average CCR for several cross-validation repetitions is an unbiased estimate of CCR. The number of cross-validation repetitions for an accurate estimate of the average CCR is about 200 (Ververidis and Kotropoulos, 2006b). In order to avoid CCR deterioration, the Sequential Floating Forward Selection (SFFS) algorithm (Pudil et al., 1994) is used to select the feature subset that optimizes the CCR.

Two classification schemes are used, namely, the single-level scheme and the two-level one. In the single-level scheme, classification is performed in three classes. In the two-level scheme, two classifiers were employed. The first classifier is optimized by SFFS for separating $\{$fear,happiness$\}$ vs. $\{$neutral$\}$ states, and the second one is used for separating $\{$fear$\}$ vs. $\{$happiness$\}$. The main idea behind the two-level scheme is that the acoustic features selected by SFFS in the first level are different than those selected by SFFS in the second level.

The CCR achieved by the Bayes classifier with SFFS in the single-level scheme is 61.7%, when the random classification is 1/3 $\approx$ 0.33%. In Table 2, the classification rates among the three states for each stimulus are shown. From the inspection of Table 2, it is deduced that the utterances expressed under the neutral state are easily recognized with a rate of 73.8%, whereas utterances colored by fear and happiness are recognized with a rates 58.7% and 52.5%, respectively. Anger and fear are often confused, due to their high arousal.

Table 2: Confusion matrix for the single-level scheme.

Stimuli/Response	Fear	Happiness	Neutral
Fear	58.7	19.9	21.4
Happiness	20.5	52.5	27.0
Neutral	14.3	11.9	73.8

The two-level classification scheme is depicted in Figure 3. The CCR achieved in the two-level scheme is 64.1%, i.e. there is an improvement of 2.4% against the single-level scheme. The confusion matrix of the two-level scheme is presented in Table 3. From the comparison between the confusion matrices in Tables 2 and 3, it is seen that the CCRs for fear and happiness are improved by 5%, whereas the CCR for the neutral state is reduced by 3%.

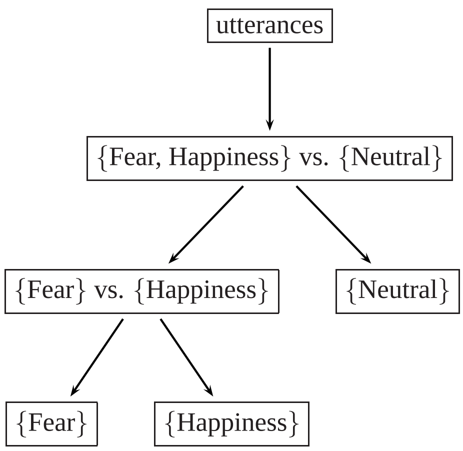

Figure 3: Classifying an utterance with the proposed wo-level scheme.

Table 3: Confusion matrix for the two-level classification scheme

Stimuli/Response	Fear	Happiness	Neutral
Fear	64.4	21.4	14.2
Happiness	25.7	57	17.3
Neutral	16.9	12.4	70.7

4 Classification of utterances into emotional states during the immersion in VR

The Bayes classifier with the two-level classification scheme is used to classify another 155 utterances (disjoint to those used during the first phase) into emotional states, which were expressed by the pupils during the VR immersion. The classification results are summarized in Table 4. From the 155 utterances, 91 utterances are classified into fear, 15 into happiness, and 49 into neutral state. Accordingly, it is deduced that the pupils faced mostly fear during the VR immersion. This is an objective evidence demonstrating that the VR immersion level of pupils is large enough.

Table 4: Classification of utterances of part B

Emotional state	Fear	Happiness	Neutral
Number of utterances	91	15	49
Percentage (%)	58.7	9.7	31.6

5 Sweat indication and heart beat rate signals

The sweat indication signal for 3 pupils is plotted in Figure 4. It is seen that the signal has many peaks and intense slopes in the first 50 sec, whereas a downward slope appears for the remaining 100 sec. This is due to the fact the virtual earthquake happens in the first 50 sec, and therefore kids become nervous. In the remaining 100 sec, kids are mostly focused on how to find the main exit of the virtual school, and accordingly they are more distracted and relaxed.

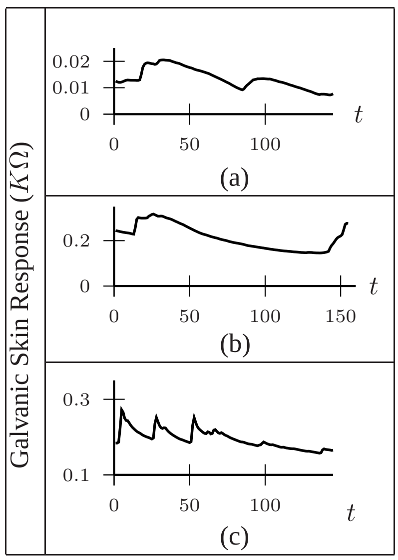

Figure 4: Sweat indication (GSR) plotted vs. time.

The heart beat rate signal of 3 pupils is plotted in Figure 5. From the inspection of these signals, a certain pattern can not be deduced. In Figure 5(a), an increasing slope of HB rate vs. time appears in the last 50 sec, when the pupil tries to find the school exit. In Figure 5(b), the pupil has approximately 100 pulses per minute without the HB rate function attaining any increasing or decreasing slopes. In Figure 5(c), the pupil's HB rate exhibits some peaks during the first 50 sec, and the HB rate function remains constant vs. time in the remaining 100 sec. The HB rate signal measured by the finger blood pressure is not so reliable as the sweat signal, because the pressure sensor is sensitive to the small movements of the finger.

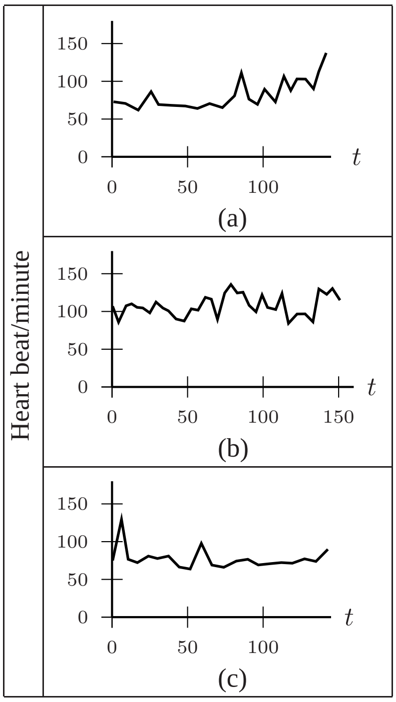

Figure 5: Heart beat rate as a function of time.

6 Conclusions

Emotion recognition from bio-signals was addressed in this paper within the context of a VR earthquake scenario. Emotion-related date were recorded in this context and first results demonstrating the use of emotion recognition to assess the immersion of a VR environment were presented.

7 References

W. Ark and C. Dryer. 2001. Computer input device with biosensors for sensing user emotions. *US patent 6190314*.

R. W. Picard. 2000. *Affective Computing*. Cambridge: The MIT Press.

P. Pudil, J. Novovicova, and J. Kittler. 1994. Floating search methods in feature selection. *Pattern Rec. Letters*, 15:1119–1125.

K. R. Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227–256.

I. Tarnanas, I. Tsoukalas, and A. Stogiannidou. 2003. *Virtual Reality as a Psychosocial Coping Environment*. CA: Interactive Media Institute.

D. Ververidis and C. Kotropoulos. 2006a. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181.

D. Ververidis and C. Kotropoulos. 2006b. Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections. In *Proc. European Signal Processing Conf. (EUSIPCO)*.