D. Ververidis, I. Kotsia, C. Kotropoulos, and Ioannis Pitas, "Multi-modal emotion-related data collection within a virtual earthquake emulator," in Proc. Inter. Conf. Language Resources and Evaluation (LREC), Morocco, 2008.

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, ekotsia, costas, pitas}@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 the pupils attempt to evacuate a school during an earthquake, is addressed in this paper. We assess whether pupils' spontaneous emotional state can be accurately recognized, using classifiers trained on elicited emotional speech and videos.

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 biosignals were recorded, while the pupils were immersed in the virtual earthquake environment. The data recorded were face videos, speech utterances, galvanic skin response for sweat indication, and blood pressure. Emotion recognition from facial videos (Kotsia and Pitas, 2007) as well as from speech (Ververidis and Kotropoulos, 2006b) have been thoroughly studied the past years. 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 subjects' facial expressions, emotionally colored speech utterances, sweat indication, and the heart beat rate. An algorithm that recognizes the emotional state of a subject from speech is briefly discussed. To train the classifier, training data are used, whose elicited emotional state is known. Accordingly, 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 the paper is as follows. Data collection is described in Section 2. The classification of the collected facial videos is presented in Section 3. The classification of utterances into emotional state is accomplished via the Bayes classifier, which is described 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.

2 Recording scenario

The VR environment emulates an earthquake occurring while the pupil is in a school classroom. Each pupil wears virtual reality glasses and sweat indicator and heart beat recording sensors prior to his immersion in the VR environment. Two microphones and a joystick, which is used for navigation within the VR environment, are placed nearby. A high resolution (near-field) camera is placed in front of the pupil in order to capture the head with high quality. Next to this camera a laptop exists that displays the VR environment the pupil sees. A second (distant) camera captures the entire experimental setup recording both the pupil and the virtual environment he/she is immersed in, in order to enable the annotation of the experimental recordings by psychologists and the synchronization of all input signals (video, audio, sweat indication, and heart beat signals). The experimental setup is depicted in Figure 1.

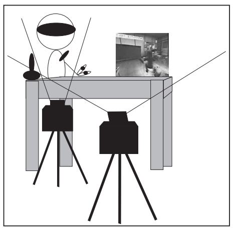

Figure 1: Experimental setup.

Regarding the experimental procedure, 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 by a certain emo-

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.

Figure 2: Recording examples depicting (a) anger; (b) disgust.

tion. By doing so, the pupils are familiarized with their role in the experiments. Two examples are depicted in Figure 2, where two pupil expressions are shown.

In detail, 14 pupils (5 boys and 9 girls) of age between 9 and 17 years were asked to express 13 utterances under 7 elicited emotional states. This utterances are used to train the speech emotion recognition algorithm. The 7 emotional states of both 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.

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

1	*Αυλή* (Yard)
2	*Διάδρομος* (Corridor)
3	*῾Εξοδος* (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 out to the yard slowly)
12	*Μπαίνω κάτω από το θρανίο* (I get underneath the desk)
13	*Περιμένω να σταματήσει ο σεισμός* (I wait until the earthquake stops)

The utterances collected in the first phase are 1396 (i.e. 14 (subjects) × 7 (states) × 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 (the first coming from the camera microphone and the second from a lavaliere microphone) were collected.

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 pupils’emotional states within the VR environment. The following recordings were collected in the second phase: (i) 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); (ii) 3 speech recordings (2 recordings stem from the two cameras microphone and the third comes from a lavaliere microphone); (iii) 1 sweat indicator signal; (iv) 1 blood pressure signal.

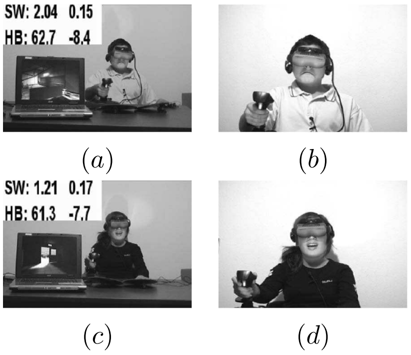

Figure 3: Captured snapshots from the distant camera for two subjects displaying sadness and fear (a) and (c) and the simultaneous high-resolution recordings of the face by the near camera (b) and (d).

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 3. In Figures 3(a) and 3(c), frames captured by the distant camera are shown. Sweat (SW) indication and heart beat (HB) rate at a certain time instant are overlaid in Figures 3(a) and 3(c). A laptop PC that shows exactly what the pupil sees was positioned near the pupil, so that the distant camera captures both the pupil and the VR scenes. Simultaneously, the second video camera records pupil’s face, as shown in Figures 3(b) and 3(d). Technical details of the equipment are briefly summarized 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 peripherals were 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 Classification of videos into emotional states during the immersion in VR

The method used to classify the facial expressions was the one proposed in (Kotsia and Pitas, 2007). However, due to the nature of the VR environment, only the facial expressions related to fear, sadness, and happiness were studied in the second (test) phase. Thus, the geometrical displacements of the deformed Candide grids were used as an input to a three-class Support Vector Machine (SVM) for facial expression recognition.

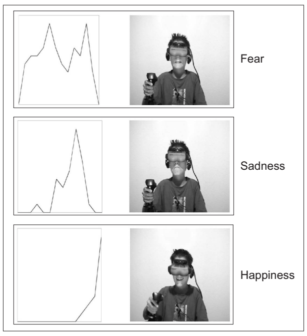

Figure 4: Evolution of anger, sadness, and surprise in time.

Fear is expected to appear most often during the experiments as it is the most common facial expression in a case of an earthquake. Sadness is also expected as the pupil tends to be disappointed when obstacles prevent him/her to go out to the yard. When the pupil finally manages to get out (at the end of the immersion) he displays happiness.

These observations agree with the results collected when the system proposed in (Kotsia and Pitas, 2007) was used. An example of the evolution of the three aforementioned facial expressions in time is shown in Figure 4. As can be seen, fear is present during the entire video recording, apart from the beginning (when the pupil is in a neutral state) and the end (when the pupil is happy). In between, the intensity of fear decreases as sadness appears due to the several obstacles preventing the pupil's exit. Happiness appears only at the end of the recording, when the pupil manages to get out.

4 Classification of utterances into emotional states

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. Cross-validation is performed in a subject-independent manner. The average CCR for several cross-validation repetitions is used to estimate the 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 ≈ 0.33%. In Table 2, the classification rates among the three elicited states for each stimulus during training 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 5. 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%.

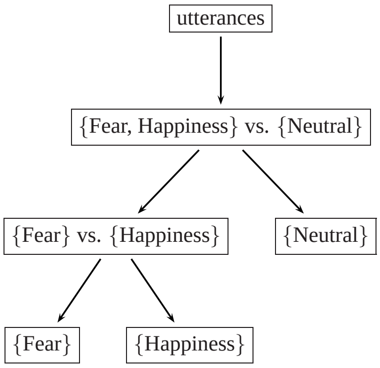

Figure 5: Classifying an utterance with the proposed two-level scheme.

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

In the second phase, 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 in the second phase.

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 of 3 sample pupils among the 14 participated in the experiment is plotted in Figure 6. 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.

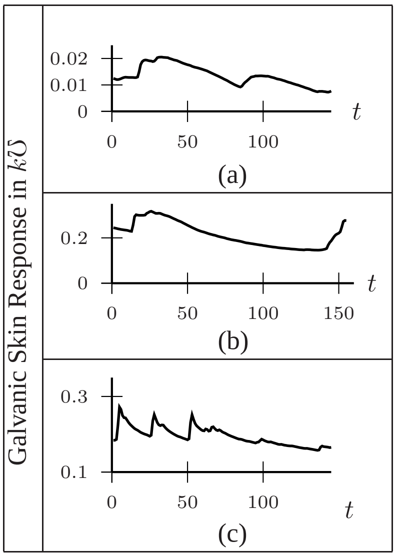

Figure 6: Sweat indication (GSR, electrical conductivity) plotted vs. time.

The heart beat rate signal of 3 sample pupils is plotted in Figure 7. From the inspection of these signals, a certain pattern can not be deduced. In Figure 7(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 7(b), the pupil has approximately 100 pulses per minute without the HB rate function attaining any increasing or decreasing slopes. In Figure 7(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.

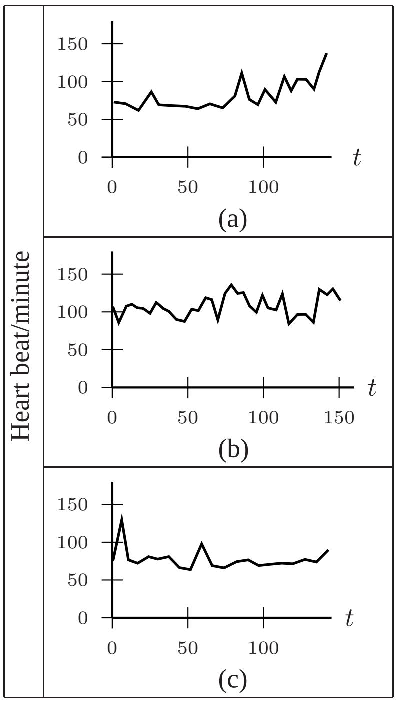

Figure 7: Heart beat rate as a function of time.

6 Conclusions

Emotion-related data have been recorded in the context of a VR earthquake scenario including facial video, emotional speech, and physiological signals. First results demonstrating the use of emotion recognition to assess the emotional state of pupils within the VR environment have been presented.

7 References

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

I. Kotsia and I. Pitas. 2007. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions Image Processing*, 16(1):172–187.

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)*.