EMODA: a tutor oriented multimodal and contextual emotional dashboard

Mohamed Ez-Zaouia, Elise Lavoué

EMODA: a Tutor Oriented Multimodal and Contextual Emotional Dashboard

Mohamed Ez-zaouia
Université Blaise Pascal, Clermont-Ferrand
SpeakPlus, 56, bd Niels Bohr CS 5213269603
Villeurbanne, France
ezzaouia.mohamed@gmail.com

Elise Lavoué
IAE Lyon, Université Jean Moulin Lyon 3
LIRIS, UMR5205, CNRS
6 cours Albert Thomas, 69008 Lyon, France
elise.lavoue@univ-lyon3.fr

ABSTRACT
Learners' emotional state has proven to be a key factor for successful learning. Visualizing learners' emotions during synchronous on-line learning activities can help tutors in creating and maintaining socio-affective relationships with their learners. However, few dashboards offer emotional information on the learning activity. The current study focuses on synchronous interactions via a videoconferencing tool dedicated to foreign language training. We collected data on learners' emotions in real conditions during ten sessions (five sessions for two learners). We propose to adopt and combine different models of emotions (discrete and dimensional) and to use heterogeneous APIs for measuring learners' emotions from different data sources (audio, video, self-reporting and interaction traces). Based on a thorough data analysis, we propose an approach to combine different cues to infer information on learners' emotional states. We finally present the EMODA dashboard, an affective multimodal and contextual visual analytics dashboard, which allows the tutor to monitor learners' emotions and better understand their evolution during the synchronous learning activity.

CCS Concepts
• **Human-centered computing** → **Visualization** → Visualization application domains → Visual analytics • **Applied computing**→ **Education** → Interactive learning environments;

Keywords
Emotions; Interactive visualizations; Learner monitoring; Tutor dashboard; Language training; Multimodal data.

1. INTRODUCTION
Learning analytics (LA) is a fast-growing field with a strong potential of making on-line learning environments competitive, attractive and, more importantly, efficient and fit for purpose [1]. Several studies position the learner at the center of interest in measuring performance in the learning environment ecosystem [2, 3]. At the same time, teachers lack visual insights that can help them monitor learners during on-line learning activities, especially with indicators on their emotions.

Several recent studies demonstrate that emotions are an essential factor in the learning process [4, 5] and, more precisely, that positive emotions play a crucial role in fostering successful learning [6, 7]. Therefore, considering learners' emotions can have a huge impact on learning and achievement. Research into the emotional dimension of learning environments has thus started to attract more interest in the community (e.g. [8, 9, 10, 11]). Consequently, to optimize the learning experience, it is important for tutors to be aware of and understand the emotional characteristics experienced by the learners.

The majority of the proposed dashboards considering learners' emotions are powered only by learners' interaction traces stored in the learning environment [12, 13]. Other dashboards are based on a single channel [10, 11] or on self-reported data only [9]. However, several studies concerning the affective computing domain, particularly emotion recognition, show that user behavior has to be considered in a global and multimodal fashion (voice, image, interaction, self-report, etc.). In this way, affective multimodal systems are more accurate than unimodal approaches [14, 15]. By considering multimodal data, the visual insights proposed on dashboards can go beyond simple reporting visual displays to attain a multimodal visual analytics tool combining different cues with their associated contextual information. Such rich visualizations can facilitate actionable feedback and decision-making for tutors in the online learning context.

The aim of our paper is to propose a multimodal and contextual emotional dashboard called EMODA. The design of this dashboard is based on a thorough analysis of learners' emotional data collected in real conditions. We more precisely answer the following research questions:
1) How to collect data on learners' emotions during online learning sessions?
2) How to exploit and combine different cues for inferring learners' emotions?
3) How can learners' inferred emotions be visualized by tutors to facilitate actionable feedback?

We conducted an experimental study using a real-time web videoconferencing learning platform called SpeakPlus. This solution aims at training learners in foreign language speaking, based on online training sessions between a tutor and a learner. We applied both dimensional and discrete models for measuring learners' emotions. We used heterogeneous APIs for emotion classification based on different cues and in non-intrusive settings. We provided an exploratory analysis comparing learners' inferred emotions and an approach for unifying different models of emotions. Based on the results, we designed an affective multimodal dashboard that presents the measured emotions with their associated contextual information. These visualizations can help tutors better monitor and understand the emotional

characteristics of the learners and, consequently, improve the socio-affective relationship between them and their learners.

The paper is organized as follows. Section 2 presents the theoretical and technological background. Section 3 details the context and the experimental settings of our study. Section 4 focuses on emotional data collection and analysis. Section 5 presents the EMODA dashboard, a multimodal and contextual emotional dashboard. The last section presents our conclusions and future directions.

2. THEORETICAL AND TECHNOLOGICAL BACKGROUND

This section gives an overview of research in learning analytics and emotions. In the first part we present some recent affective learning systems. The second part is dedicated to emotion theories, emotion recognition and measurement methods.

2.1 Emotion and Learning

Over the last decade, researchers in the educational field have showed huge interest in emotions. Recent studies underscore their key role in learning, regulation processes and strong coupling to motivation and achievement [5, 6, 8, 16]. Learning is more likely to be successful if tutors, as well as learning platforms, help minimize negative emotions (e.g. fear, frustration, stress) and facilitate positive ones (e.g. happiness, enjoyment). This has led to the increased importance of emotional data in online learning and teaching environments. Designing learning environments that are enjoyable, motivating and inspiring for learners is a key issue for the LA community [17].

In a traditional learning environment, tutors can, more or less, easily monitor and evaluate the emotions experienced by their learners. However, in computer-mediated learning settings, this might not be an easy task for tutors. Therefore, emotional analytics can provide tutors with the opportunity to monitor and better understand their learners' emotions (e.g. [9]). The development of such affective monitoring systems depends on the measurement of learners' emotional state. Methods for measuring emotions can be either Objective/Subjective, Snapshot (before and after learning)/Continuous and Qualitative/Quantitative [18].

Recent research has emerged on this topic by tracking learners' emotions to promote awareness, decision-making and information visualization. *Ruiz et al*. [9] propose an emotional dashboard to support students' self-reflection. The authors opted particularly for subjective methods to measure students' emotions using a questionnaire addressed to learners (before and after the learning activity – snapshot type). Their results show that the students' emotions are correlated to their performance in class. *GhasemAghaei et al.* [10] propose a dashboard to help and support tutors in reflecting on students' emotions experienced during learning. The authors used objective and continuous data, mainly images extracted from students' webcams during the learning activity, to infer students' emotions using a dedicated software. *Happy et al*. [11] propose a system that can estimate both alertness and attention levels from ocular parameters and the emotional state using objective data, mainly visual cues (video).

Most of the proposed dashboards in the educational domain are rather unimodal-based (e.g. [9, 10, 19]). Some tools collect subjective data from questionnaires that are easy to handle but cannot detect emotional reactions at given times during the learning session. Furthermore, unimodal data and subjective questionnaires are not as accurate as global multimodal approaches as explained in the next section. The latter introduces emotion theories and the main tools for automatic emotion recognition.

2.2 Emotion Recognition Theories

Emotions represent a complex psychological state of the person. Researchers examining the development of our internal emotional aspects have come to diverse opinions. This diversity leads to different perspectives for emotions.

Discrete and *dimensional* are the two most important and widely adopted emotion theories. *Discrete* emotions are a small set of distinct emotions (e.g. anger, disgust, fear, happiness, sadness, surprise), called universal or basic emotions, which constitute the core of all humans' emotions regardless of the socio-cultural factors of the person [20]. *Dimensional* emotions are rather structured in a dimensional space varying along certain dimensions, such as the degree of the *valence* representing the positiveness of the emotion felt (varying between positive and negative) and the *arousal* that represents the physical response and the intensity of that emotional manifestation (varying between low and high) [21]. It should be noted that, while other dimensions have been proposed (e.g. *'control'* [22]), valence and arousal are still the most widely adopted dimensions. At the same time, different studies have been derived, mapping one theory into the other (e.g. [22]).

The interest in emotions has rapidly increased in human-computer interaction applications, more commonly known as *affective computing* [23]. Research in psychology, neuroscience and computer science, especially Artificial Intelligence (AI), has been combined to "teach" computers how to infer humans' emotions. Researchers rely on the intensity "emitted" (internally or externally) by the emotion felt in order to infer the person's emotions. *Subjective emotions* of a person can be collected through self-reporting. Two other approaches are used to collect objective measurements: *perception-based* measurements include all the manifestations expressed by the person (facial, vocal, gesture, textual, etc.) while *physiological* measurements include all human body responses (heart-beat, blood pressure, brain activity, etc.) [24].

The automatic emotion recognition (AER) process can be divided into three main stages: *feature extraction* aims to filter, extract and normalize the suitable features, *feature reduction* aims to reduce the high dimensionality of the extracted features, and *classification* aims to classify emotions using machine learning techniques [25]. Our research follows these three AER steps to feed the visualizations offered in our dashboard.

Three main features are used for automatic emotion recognition in non-intrusive settings. *Facial expression recognition* can be based on two techniques: a geometric model-based technique that relies on distinctive facial features such as the position of nose, eyes, mouth, and an appearance model-based technique that considers the face as an array of intensity values (pixels). Thus, facial expressions can be mapped to emotions through theories such as those proposed by [20]. In terms of tools we can cite for example FaceReader[1] and Microsoft Emotion Service API[2] that we adopted in our study. Other automatic emotion recognition features are based on *voice (sound)* and are attracting increasing interest with researchers. The most commonly used features are the prosodic (pitch, intensity and first formant frequency) [26] and the spectral features, (Mel-frequency spectral coefficients (MFCC) [26]. In

[1]http://www.noldus.com

[2]https://www.microsoft.com/cognitive-services/en-us/emotion-api

terms of tools, we can cite for example OpenEAR [28] and Beyond Verbal Emotion Service API[3] that we chose for our study. Textual cues are also considered. However, we do not consider them since the learning activity we observe is not textual-oriented.

The most important current trend in emotion recognition is the combination of different cues (visual, vocal, textual, etc.) for inferring emotions. In fact, many studies (e.g. [14, 15]) show that the multimodal approach for measuring people's emotions is more accurate than the unimodal approach. Different strategies exist for data fusion. The first one, called *feature level fusion,* relies on feature vectors for the different cues that are stacked together and used as input to the classifier. The second strategy is called *decision level fusion*; each cue is classified separately with its classifier, and the result of each classifier is combined and used as input for a final classifier. We opted for this second strategy.

In our study, the means of communication between the tutor and the learner is a videoconferencing: audio-visual cues are thus the main data channel. We propose to apply both dimensional and discrete emotion theories using several APIs for emotion recognition. We combine these data with self-reported data and interaction traces. We thus aim to acquire rich data on learners' emotions to produce a multimodal and contextual emotional dashboard for tutors.

3. CONTEXT AND METHODOLOGY

This section is dedicated to the description of the context and the procedure of our study, mainly the learning environment and the experimental settings.

3.1 SpeakPlus: a Web-based Learning Environment for Foreign Language Training

SpeakPlus is a commercial Software as a Service (SaaS), a web-based environment dedicated to improving oral communication skills in a foreign language. English, French and Spanish are supported by the platform. The platform connects qualified tutors and learners around the world in a way that the learning activity focuses mainly on the goals set by the learner him/herself (e.g. improving oral presentation skills, preparing for travel abroad). The platform is designed in a modular fashion to be used as a research tool by SpeakPlus researchers' partners, mainly for testing and evaluating new approaches and didactic models in the educational domain.

The learning activity in SpeakPlus can be divided into two main steps. The first step takes place before the learning session: at this level, the tutor can create or customize learning materials (activities) for each learning session by creating new materials or reusing old ones from his/her library. By combining the materials, the tutor can create a learning plan for each session. The learning activity has a duration and can be associated with documents, instructions for the learner or personal notes for the tutor. After preparing the learning materials, the tutor can conduct a synchronous learning session with a learner (see Figure 1). The tutor and the learner communicate in real-time by means of a dedicated videoconferencing technology optimized for online learning. The tutor has the possibility of sharing the prepared materials with the learner (e.g. pdf, word, image, audio, and video) or of communicating via the chat (mainly to point out some expressions, words, etc.).

3.2 Experiment Procedure and Participants

Our study is based on authentic and non-intrusive settings. In fact, the data were collected using sensors already used in the platform for the learning activity, namely a webcam and a microphone for video and audio recording, respectively. The platform has been slightly augmented to collect and process data (see section 4).

Two graduate female students (in their twenties) agreed to participate in our pilot study. They were contacted after their enrollment in the platform to train in French language speaking for professional goals, more precisely for preparing a job interview. One learner is from Latin-America and the other from Asia. Each learner trained for five learning sessions with the same tutor. The learners had one learning session every week. Each session lasted 45 minutes. Data collection lasted around eight (8) weeks as the students did not start their learning sessions at the same time. We collected a total of 7 hours and 30 minutes of audio/video data.

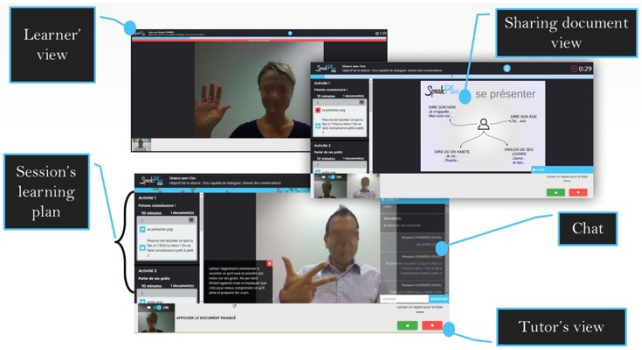

Figure 1. Live learning session on the SpeakPlus platform

4. EMOTIONAL DATA ACQUISITION AND ANALYSIS

This section presents the first three stages of our research: learners' emotions data acquisition, data exploration, and data analysis, that we applied to produce visualizations on the proposed emotional dashboard for tutors.

4.1 Emotional Data Acquisition

Throughout this section we aim to respond to research questions *1)* and *2)* identified in the introduction. Therefore, different trackable cues have been chosen to investigate how they can be used to track learners' emotions in the SpeakPlus learning environment. We adopted a multimodal approach and identified four data sources: audio, video, self-report, and interaction traces. Audio and video are the primary communication channels in the platform as the learning activity aims at improving oral skills via videoconferencing technology. Self-reported data inform on learners' subjective emotions before and after the learning session. Finally, interaction traces have been used to contextualize the measured emotions and investigate into whether or not there are any particular triggers behind learners' emotions.

According to our multimodal approach, we used heterogeneous cloud APIs for measuring emotions. These APIs were chosen to apply both dimensional and discrete emotions models. The Microsoft emotion recognition Service API (MS API) classifies emotions based on facial expressions (video) according to a discrete emotion model. The Beyond Verbal service API (BV API) classifies emotions based on voice (audio), according to a dimensional emotion model (arousal and valence).

Figure 2 shows the global architecture of the system we have built to collect emotional data during learning sessions. The learning sessions between the tutor and the learner are recorded

[3]http://www.beyondverbal.com/api

(audiovisual). Records are stored along with the session in the cloud (AWS S3). At the end of the session, audiovisual records are downloaded from S3 and processed via the BV API and the MS API. The resulting classified emotions are then stored in a database, together with the interaction traces and self-reported data, via a developed API for data collection. The stored data are then processed and aggregated to produce visualizations. The following subsections describe for each data source how the data are collected and stored and how emotions are classified.

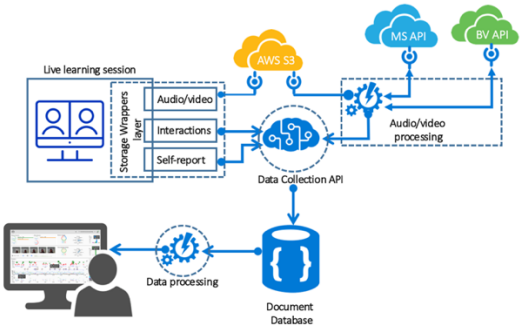

Figure 2. EMODA global architecture

4.1.1 Self-Reported Emotion Collection

We developed an interactive interface that enables learners to express their emotion(s) before and after the learning session, according to both dimensional and discrete emotions. The dimensional interface is inspired from the MoodMap tool proposed in [29]. Basically, it proposes a two-dimension (valence and arousal) squared area (Figure 3-a). The learner can click on it to express his/her valence and arousal levels. Four colors were combined in a conic gradient form to represent the four levels: positive and negative valence by green and red colors respectively, and positive and negative arousal by blue and yellow colors respectively. A mouse hover event has been added to track and display the position of the mouse. This information is reflected properly in a mirror bar chart at the bottom of the squared area to guide the learner when reporting his/her two-dimensional emotion. Figure 3-b presents the interface that enables learners to report their emotions in a discrete manner. A slider for each emotion (neutral, happiness, surprise, sadness, disgust, contempt, anger) is used to report its intensity on a continuous scale from 1 to 100. An emoticon and a tooltip displaying the emotion name (when the mouse is hovering the slider) are used to distinguish between each emotion. The value of the slider is tracked and displayed in a marker associated with the slider and inside a bubble next to it.

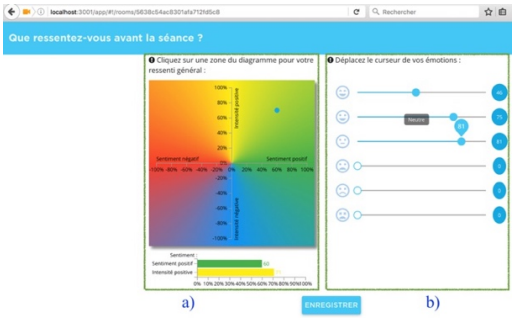

Figure 3. Interface for self-reporting of emotions

In our study, learners were asked to express their emotions both at the start and at the end of the learning session. This emotional information is complementary to the data automatically collected during the learning session. The collected self-reported data are stored in a database as illustrated in Figure 2.

4.1.2 Facial Expression Acquisition and Analysis

The learning sessions are recorded in the SpeakPlus platform. Each session may have one or more archive folder(s) depending on whether or not there were any interruptions during the session. Each archive folder contains two files: one for the learner recorded stream and the other for the tutor recorded stream. An algorithm based on the *FFmpeg* library (ffmpeg.org) was used to concatenate the video streams from the archive folders when needed. Once the video sequence has been constructed, emotions are classified and stored using a simple algorithm that first uses *FFmpeg* to extract frames (one frame per second) from the video stream and then calls the MS API to classify the emotions identified for each frame, before storing the result in a database. More precisely, the MS API returns a face entry and its associated emotion scores. An empty response indicates that no face was detected. An emotion entry contains the fields listed in Table 1.

Table 1. Description of MS API response (& JSON example)

Field	Description
faceRectangle	Rectangle location of the face in the image.
scores	Emotion scores for each face in the image; (neutral, happiness, surprise, sadness, disgust, contempt, fear, anger).
application/JSON [{ "faceRectangle": { "left": 68, "top": 97, "width": 64, "height": 97 }, "scores": { "anger": 0.00300731952, "contempt": 5.14648448E-08, "disgust": 9.180124E-06, "fear": 0.0001912825, "happiness": 0.9875571, "neutral": 0.0009861537, "sadness": 1.889955E-05, "surprise": 0.008229999 } }]	

4.1.3 Audio Data Collection and Analysis

The same algorithm as for video was used to construct the sound sequence from archive folders (only the learner channel). A call of BV API then returns an array of time interval entries and their associated emotion scores. An empty response indicates that no emotions were detected in the audio file. An emotion entry contains the fields listed in Table 2. The classified emotions are then stored in a database as illustrated in Figure 2.

Table 2. Description of BV API response (& JSON example)

Field	Description
analysisSegments	The array containing analysis segments.
Offset	The offset of the segment from the beginning of the sample being analyzed (in ms).
duration	Duration of the analysis segment in the sample being analyzed.
analysis	Analysis object. Contains analysis values for the segment. The content of the object is provided as an example.
valence	Valence object score (has value and group): - Value: a value between 0 and 100. - Group: positive, neutral or negative.
arousal	Arousal object score (has value and group). - Value: between 0 and 100. - Group: low, neutral or high
application/JSON "analysisSegments": [{ "analysis": { "Arousal": { "Group": "low", "Value": "4.35" }, "Valence": { "Group": "positive", "Value": "82.28" }	

```
"duration": 37410,
"offset": 4576 } ]
```

4.1.4 Interaction Trace Collection

We decided to track all actions performed by the tutor and the learner on the SpeakPlus environment during the learning session. This information can be used to contextualize the collected emotional data. We list in Table 3 the events tracked (in real-time) and stored in database using WebSocket.

Table 3. List and description of events collected in interaction traces with the SpeakPlus environment

Field	Description
action_name	The name of the action (event) triggered between the actors, it can be: - SHARING_DOC: the tutor starts sharing a document - STOP_SHARING_DOC: the tutor stops sharing a document - ACTIVITY_TASK: an action sent to the learner by the tutor - FREE_TEXT: message chatting - SHOW_DOC: the learner enables display of the shared document - HIDE_DOC: the learner disables display of the shared document
action_content_type	The type of event: - PDF: Pdf document sharing - AUDIO: Audio file sharing - VIDEO: Video file sharing - IMAGE: Image file sharing - TEXT: Text action sent to the learner by the tutor - SHOW_HIDE_DOC: for the SHOW_DOC or HIDE_DOC action
activity	Activity Id
item	Item Id in the activity (an activity might have one or more items)
document	Document Id (pdf, audio, video, etc.)
chat_message	Message-Id

4.2 Emotional Data Exploration and Analysis

This part presents the data exploration and analysis stage of our study. It addresses mainly research question *2)*, i.e. investigating whether or not audio and video data are correlated. We do not include self-reported information in the analysis because it is not collected at the same time (before and after the session). Throughout this section we answer more specifically the following sub-questions: a) Is there a correlation between audio and video data? and b) What valuable emotional information should be visualized?

4.2.1 Emotional Data Comparison

Use of two different models for emotions (dimensional based on audio and discrete based on video) was the most challenging part of this comparison. Thus, to be able to compare these cues both models have to be unified. Several studies have been conducted in this context to convert discrete emotions into a dimensional model (valence and arousal). We used the valence coordinates of the discrete emotions (neutral, happiness, surprise, sadness, disgust, contempt, fear, anger) as proposed in [22]. The coordinates provided are on a scale from -100 to 100, from negative valence to positive valence (neutral emotion has a valence equal to zero).

As described in section 3.1.3, audio valence is on a scale from 0 to 100 (as returned by the BV API). For instance, neutral valence is equal to 50. Therefore, scale unification was needed between the coordinates from [23] and the returned result from the VB API. To this end, we used the formula $(f1)$ that maps a *domain* [a, b] interval to a *range* interval [c, d] to scale VB interval from [0, 100] to [-100, 100].

$$f(x) = \frac{(d-c)\,(x\ -\ a)}{b-a} + c \qquad (f1)$$

As mentioned in 3.1.2, a successful call to the MS API returns the emotion scores of the frames (images). By combining the coordinates from [23] and the returned scores, we were able to compute for each frame the corresponding value of the valence. The weighted mean was used for this purpose, as illustrated in the formula $(f2)$:

$$valence_{avg}^{frame^n} = \sum_{\substack{i \in \{\text{neutral,happiness,}\\ \text{surprise,sadness,}\\ \text{disgust,contempt,}\\ \text{fear,anger}\}}} score^{frame^{n^i}} * valence^i \qquad (f2)$$

- $\text{score}^{\text{frame}^{\text{n}^{\text{i}}}}$: is the score of the emotion i ∈ { neutral, happiness, surprise, sadness, disgust, contempt, fear, anger } in the result returned by the MS API for the frame n (frame^{n})
- $\text{valence}^{\text{i}}$: the valence corresponding to the emotion i.
- $\text{valence}_{\text{avg}}^{\text{frame}^{\text{n}}}$: the weighted valence average (weighted mean) of the frame n.

As described in 3.1.2, a call to the BV API returns an array of time segment entries with their associated emotion scores (valence and arousal). Thus, a simple algorithm has been implemented to map the images extracted from video, every second, to its corresponding audio time segments. Once the images have been grouped, the valence is computed first for each image (with $(f2)$). The average over images belonging to each time segment is then computed.

4.2.2 Emotional Data Correlation

This section responds to the sub-question a). Once the two data models were unified, the aim was to investigate whether there is any correlation between both variables: audio and video valence.

There are several coefficients for measuring dependencies between variables. Spearman's and Pearson's coefficients are those most used in the literature. However, they can detect only linear or non-linear monotonic correlations as stated in [30]. Conversely, the Maximal Information Coefficient (MIC) may be more powerful as it can detect several associations between variables, such as linearity, nonlinearity, asymmetry and even non-functionality [30]. Table 4 presents the correlation coefficients (MIC, Spearman, and Pearson coefficients) of audio and video valence.

Table 4. Correlation coefficients between Audio / Video valence - MIC vs. Pearson vs. Spearman

	Coefficients		
	MIC	**PEARSON-R**	**SPEARMAN-R**
Valence Audio Vs Video	0.787	0.052	0.172

Regarding the MIC coefficient, the variables (audio and video valence) are correlated (contrary to Pearson and Spearman). We deduce that there is no clear linear correlation between the video and audio variables, but there must be a correlation between the peaks (either positive or negative) of both variables, which is a non-functional correlation that the MIC can detect. At this stage, an exploratory analysis using audio-video records was necessary to further investigate and understand this non-functional dependency between both cues.

4.2.3 *Emotional Data Exploration*

To explore audio and video valences, we draw a line chart for the ten learning sessions. The result reveals an explicit dependency between both valences. Figure 4 and Figure 5 are examples of one learning session valence timeline. The x-axis represents the time-segment indices, while the y-axis represents the corresponding valence score between -100 and 100. As annotated with blue color rectangles, many apparent *similarities* and *dissimilarities* exist between the peaks in audio and video valences.

Figure 4 presents some examples of *dissimilarities* comparing raw audio and video data for one learning session. Table 5 presents the associated interpretations and comments for the first four segments.

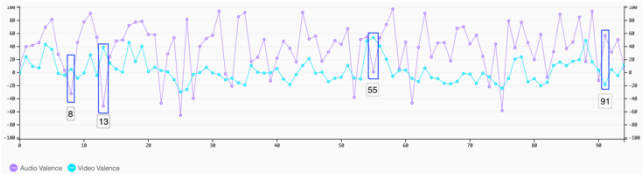

Figure 4. Audio and video valence - examples of dissimilarities

Table 5. Interpretation of valence dissimilarities (audio/video)

Index	Audio	Video	Comments
8	The correction of a mistake by the tutor triggers a regret reaction in the learner. Thus valence decreased.	Almost half of the images have been classified as 'happiness'. Thus valence increased.	Even if the learner's voice showed regret, there was no sign of this on his/her face.
13	A kind of 'aaaah' expressed by the learner. Thus valence decreased.	Neutral emotion	Audio might be a better indicator for some specific emotional vocal expressions.
55	The learner's voice was calm (neutral).	Many images have been classified as 'happiness'. Thus valence increased.	The learner was smiling, which was a sign of 'happiness', but a calm voice was a sign of 'neutral' for audio.
91	The learner was speaking loudly and confidently. Thus valence increased.	The majority of images have been classified as 'neutral'. Thus valence decreased.	Neutral was the dominant emotion for the video while the audio indicated a high level of valence which is not reflected on the face.

Similarly, Figure 5 presents some similarity examples comparing raw audio and video data for the same learning session as above. Table 6 presents the associated interpretations and comments for the first two segments in the figure. The remainder of the segments is more or less similar to the commented segments.

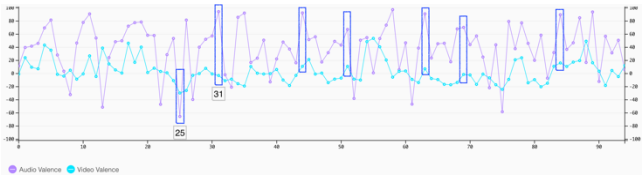

Figure 5. Audio and video valence - examples of similarities

The comparative analysis conducted in this section reveals some interesting points. First, by adopting different models of emotions and using heterogeneous APIs for emotion recognition, we were able to identify many similarities and dissimilarities between the measured emotions from both audio and video data. This result underscores a dependency between these cues. Second, as described in Table 5 and Table 6, audio data can reveal some particular levels of valence that video data fail to detect and vice versa. We deduce that both cues should be considered for inferring learners' emotions.

Table 6. Interpretation of valence similarities (audio/video)

Index	Audio	Video	Comments
25	The learner was obliged to interrupt the session in order to talk to a member of his/her family.	'Neutral' emotion	This proves that the result might be more accurate when both cues are combined.
31	There was a joke in this segment that leads to positive valence, as well as a peak in the audio valence.	Only one image has been classified as 'happiness'.	This also proves that the result might be more accurate when both cues are combined.

4.2.4 *Emotional Data Integration*

This section responds to sub-question b) and tends to aggregate data to create information for the visualizations offered in the dashboard. The amount of data collected for each session is huge, and only a subset of these data is likely to constitute an effective emotional state. Hence, both dimensional and discrete emotions need to be filtered and synthetized to keep only valuable emotional reactions.

Concerning dimensional emotions (based on audio), we use arousal (intensity of the emotion) to filter the valence of audio data. A list of arousal thresholds (10, 20, 30, 40, 50) was defined. When a threshold is selected, only the time-segment entries with the highest corresponding arousal are considered of interest during a learning session.

Concerning discrete emotions (based on video), time series analysis was used to synthetize the collected emotions. First, the set of emotions (neutral, happiness, surprise, sadness, disgust, contempt, fear, anger) is divided into three categories: neutral, positive (happiness, surprise) and negative (sadness, disgust, contempt, fear, anger). Second, the scores of these emotions are computed as the sum of the corresponding emotions for each category and so, for each image (frame), (see $(f3)$ $(f4)$ $(f5)$). Third, once the scores have been computed for each image, we identify the dominant type of emotion (neutral, positive, and negative) as the most frequent over all the frames of the segment.

$$score_{Neutral}^{frame^n} = score^{frame^{n^{neutral}}} \quad (f3)$$

$$score_{Positive}^{frame^n} = \sum_{i \in \{happiness,\ surprise\}} score^{frame^{n^i}} \quad (f4)$$

$$score_{Negative}^{frame^n} = \sum_{i \in \{sadness,\ disgust,contempt,fear,anger\}} score^{frame^{n^i}} \quad (f5)$$

- $score^{frame^{n^i}}$: is the score of the emotion $i \in$ { neutral, happiness, surprise, sadness, disgust, contempt, fear, anger } in the result returned by the MS API for the frame n.

We then identified two patterns: the positive moments and the negative moments (neutral is dropped as it may not be interesting

for learner monitoring). Thus, two measurements have been defined, *'True positive'* and *'True negative',* to track these patterns. *'True positive'* detects the moments where the average of positive points is highest or equal to a fixed threshold. *'True negative'* is defined to detect moments where negative emotions are defined as significant, i.e. the gap between the average of negative points and the maximum score of the remaining emotions is less than or equal to a fixed threshold. The detected positive/negative moments for both dimensional and discrete emotions are considered as the most important time-segments that are more likely to represent effective emotional reactions during the learning session, hence, they will be used to power our visualizations described in the next section.

5. EMOTIONAL DATA VISUALIZATION

This section addresses our research question *3)*. We defined four design principles (DPs) to build easily readable and understandable visualizations: *dp1) Unified design*, *dp2) Self-explanatory*, *dp3) Easy to use and dp4) Interactive design*. By adopting these principles, we aim to make the dashboard interface intuitive and thus increase the learning curve.

5.1 Dashboard Overview

The emotional dashboard we propose is illustrated in Figure 9. It is a tutor oriented web-based application. The tutor connects to the dashboard by selecting a learner from his/her learners' list. The dashboard is then powered by the data for the selected learner. The dashboard is designed so as to present information from general-to-specific at different levels of abstraction: 1) *Global information* about the learning session reflects in a summarized manner the emotional state of the learner during the learning session, combining both subjective and objective emotions; 2) *Timeline information* integrates contextual information to illustrate the learning session with the emotions experienced by the learner along the learning activity; 3) *Time-segment information* is designed to provide more details about each positive or negative emotion identified in the timelines.

The dashboard consists of a toolbar at the top and a body (see Figure 9). The toolbar provides quick access to the filtering functions. The tutor can select in the drop-down menu a learning session corresponding to the selected learner. The tutor can also select an arousal threshold in the drop-down menu. The toolbar is also used to quickly show insights about learners' emotions during the session by displaying the number of positive and negative emotions in the form of markers. The body of the dashboard groups all the emotional visualizations related to the selected session. When the session changes, all the visualizations in the body are updated accordingly.

The top level of the body is designed to convey emotional insights about the whole learning session. It is divided into two parts: the left part presents the emotions expressed by the learner (self-reported) before and after the learning session, while the right part presents the average of the emotions measured over the session. The middle part of the interface provides details on a selected time-segment. It is divided into two parts: the left part presents contextual information related to the time-segment, in particular learners' facial expressions recorded during the learning session and the audio record, while the right part displays quantified information on associated emotions. The bottom part is composed of three timelines, where the degree of detail increases from top to bottom. On the top, bubbles represent the positive and negative moments. In the middle, we placed contextual and emotional information in a multiline-based chart. At the bottom, a segment-based chart presents the top three emotions for each time-segment.

Color is an important abstraction widely adopted in visual design and particularly in emotional visualizations (e.g. [9, 10]). Hence, we decided to define a color scheme for our visualizations. Figure 6 presents three sets of colors: the first for discrete emotions (neutral, happiness, surprise, sadness, disgust, contempt, fear, anger), the second for what we call markers: positive/negative emotions and events (chatting and sharing documents between tutor and learner) and the third for the audio/video valence.

Figure 6. Color Scheme

5.2 Global Information Level

The left part of the global level concerns the subjective information expressed by the learner. As explained in section 3.1.1., learners can report their emotions immediately before the beginning of the live learning session and just after its end. Figure 7 presents the implemented visualization of both dimensional (valence and arousal) and discrete emotions at both times. This representation facilitates the comparison between the two times (before and after). Two different charts are used to visualize self-reported dimensional emotions and discrete emotions. A pie donut chart is adopted to visualize dimensional emotions, where the corresponding name and value (on a scale from 0 to 100) are placed inside the pie chart. Chart color is green or red, depending on whether the value of the emotion is positive or negative, respectively. As the learner can choose several discrete emotions, a light bar-based chart is adopted to represent them. Each bar represents an emotion, and the width corresponds to the value of the emotion as expressed by the learner on a scale from 0 to 100. The emotion name and value are displayed on the top of each bar.

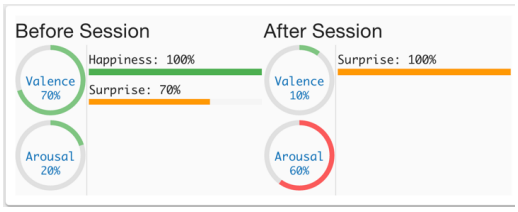

Figure 7. Self-reported emotion visualization

Figure 8. Session detail visualizations

The right part presents the average information of the inferred emotions (using APIs). Three visualizations are proposed (see Figure 8), ranging from the most general (on the left) to the most detailed (on the right). The visualization in Figure 8-(a) is based on a multi-arc pie donut chart. It presents the ratio of each type of discrete emotion, corresponding to the abstraction of positive, negative and neutral emotions as described in section 4.2 (emotional data integration). Each type is represented by a color (red, green and blue); the percentage is also displayed with

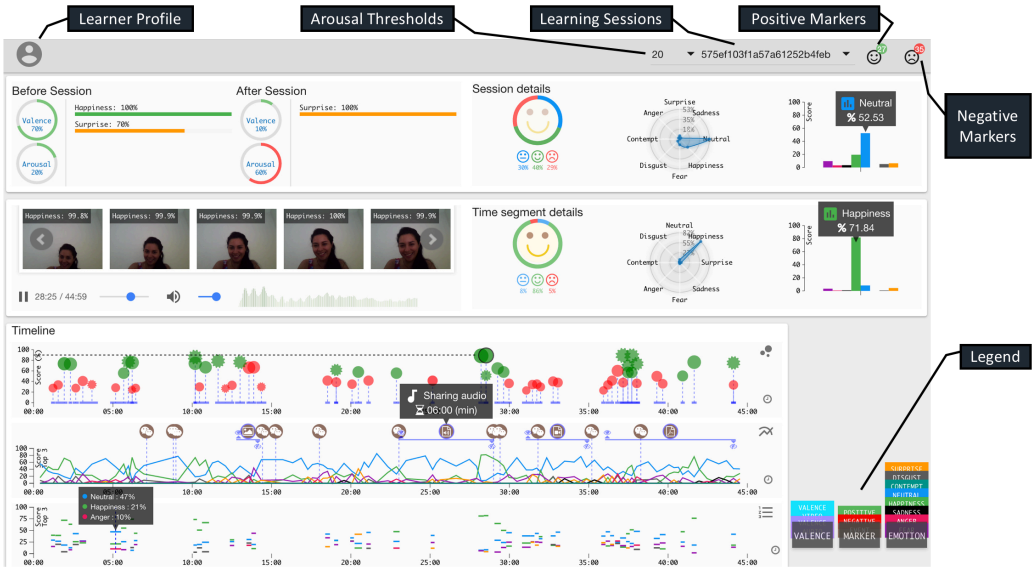

Figure 9. An overview of the EMODA dashboard. A positive marker is selected in the first timeline

three emoticons for positive, negative and neutral emotion. An emoticon representing the dominant emotion is placed inside the chart. This chart thus quickly conveys the dominant emotion set as well as the percentage of each type. The second and third segment detailed visualizations are a radar and bar-based chart (Figure 8-(b) and Figure 8-(c)), respectively. Both visualizations present the discrete emotion average data corresponding to the learning session. Bars are colored according to the color scheme; the y-axis represents the score of the emotion, which is also displayed in a tooltip (*dp4*) when the mouse hovers the bar.

5.3 Timeline Level of the Learning Session

Timeline visualizations offer more explanatory (*dp1*) representations of emotions as they incorporate the *time* notion of the learning session. Our main timeline visualization is based on the analysis described in part 4.2. As illustrated in Figure 10, a specific timeline chart is implemented to combine both discrete and dimensional emotions. We represent positive/negative markers for positive and negative time-segments, respectively. We choose a bubble shape to represent discrete emotion markers and a star shape to represent dimensional emotions. The size of the bubbles and stars depends on the computed score for each emotion which is also represented with the height of the marker.

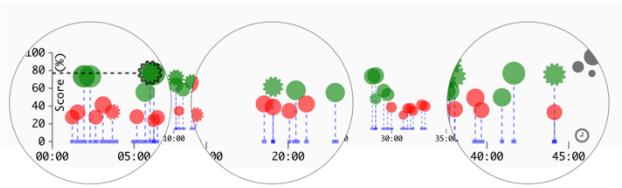

Figure 10. Positive/Negative markers visualization

The duration of each marker (time-segment) is added as a small horizontal bar over the x-axis, which represents learning session time in minutes. As mentioned above, a green color is chosen for positive markers and a red color for negative ones. When the session or arousal threshold changes (using drop-down menus in the toolbar), the visualization is updated accordingly. Integrating both markers (dimensional and discrete) in the same timeline facilitates the comparison between the emotions from both audio and video data sources.

Figure 11 presents a second timeline displaying only the top three discrete emotions associated with each marker. The x-axis represents the time in minutes, and each emotion is represented by a horizontal bar with the associated color. The width of the bar corresponds to the duration of the time-segment, while the position of the bar in the y-axis corresponds to the emotion score on a scale from 1 to 100. This visualization simplifies the comparison between the top three emotions.

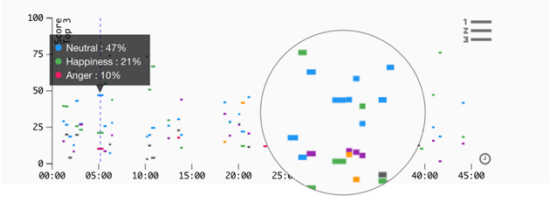

Figure 11. Top 3 discrete emotion visualization

The third timeline visualization (see Figure 12) allows the tutor to monitor the discrete (top three) and the dimensional (audio/video valence) emotions over time with the associated contextual information (*events*). We collected two main events: chatting and document sharing (pdf, audio, video or image). Emotions are represented by lines with different colors. *Events* are added to the top of the chart using bubbles as markers, with different icons to distinguish *events* by their types. The duration of the event is shown by a horizontal bar and two small icons (eye, eye-slash) placed at both ends of the bar (showing the start and end of document sharing). This setting not only allows better monitoring of learners' emotions but also correlation between emotions and the events and activities that may have triggered these emotions. For instance, we can observe on this chart that the learner was greatly surprised at ~ 14 minutes while sharing an image.

By providing tutors with different levels of abstraction of learners' emotions and contextualized information on these emotions, we aim at helping them easily identify specific emotional moments and reactions in the learning sessions (both positive and negative) and explore the reasons (context) of such moments.

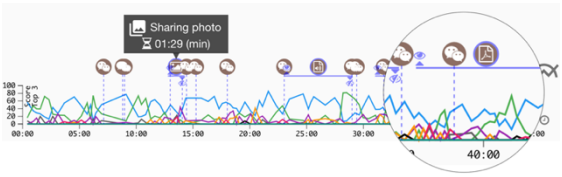

Figure 12. Contextual event and emotion visualization

5.4 Time Segment Detail Visualization

We provide tutors with details on emotional information corresponding to each particular time-segment (marker). These visualizations offer a detailed view of learners' emotions at a given moment selected in the timeline described above. The dashboard is interactive (*dp4*): when the mouse hovers a marker in the timeline. The detailed visualizations are updated accordingly with the data related to the hovered marker.

To compare time-segment information and global information, the same visualization presenting the average of the objective emotional information was used and stacked vertically (see Figure 9). We also help tutors better understand quantified information on learners' emotions by showing either facial expressions and the sound corresponding to each time-segment (see Figure 9). Images are updated accordingly when the mouse hovers the markers (see Figure 10). Likewise, the tutor can listen to the sound of a selected time-segment by clicking on its associated marker. To know whether or not the learner is talking, we implemented a bar-based chart to display audio frequency with green or red colors (depending on whether the selected marker is positive or negative). Audio-visual contextual information is placed at the same level as time-segment details visualizations.

5.5 Pilot Study on Tutors' Perception

We conducted an initial study in order to evaluate the EMODA dashboard and obtain feedback on how tutors perceive its usability and utility as recommended in [31]. This pilot evaluation was conducted with two tutors using the SpeakPlus platform. For this purpose, the evaluation was conducted in three steps. The first step consisted of a case study with 8 associated questions (e.g. *'At what time did the learner feel a peak of fear during the learning session?'*). The second step consisted of 6 questions based on a 7-Likert Scale regarding global perception on the dashboard (e.g. *'How do you find the display and organization of data on the screen?', 'To what extent do you think the dashboard will help you as a tutor to improve the learning experience?*). The third step consisted of 7 grid questions with 7-Likert scales for each visualization (e.g. *'How do you find the learner image carousel displayed in part (II-a)?'*). We added to each question a text area requesting more explanations.

The answers to the questionnaire associated with the case study confirmed that the dashboard is easy to use, as the tutors were able to answer all questions correctly without any help. Regarding global perception, it was clear that our dashboard has a problem in displaying timeline visualizations. Tutors state that *'reading is a little bit difficult'* and *'Conversely, timeline visualizations were not rapidly comprehensible... for me it would be better to keep only one simplified visualization'*. This suggests that one simplified visualization of the three timeline visualizations would be easier to understand, for instance by adding a tooltip with the mouse hover events to the marker timeline (Figure 10). This apart, the perception of the dashboard was rather positive. From the last part of the questionnaire, we note that the idea of combining emotions and events was very interesting for both tutors. In fact, they suggested further "promoting" this visualization: *"should be better valued*", *"the idea of combining events and emotions is interesting, but the chart is not big… it's a good idea but should be simplified"*. Therefore, more focus should be given to the contextualization part as it provides essential understanding of the emotions experienced by learners.

6. CONCLUSION AND FUTURE WORK

Our initial investigations prove the feasibility of using heterogeneous APIs for emotion recognition in online learning environments. By applying both dimensional and discrete emotion models, we showed that we can detect different learners' emotions during such settings. This highlights the need for a multimodal approach to emotion collection and recognition. We believe that providing tutors with an awareness tool to visualize and track their learners' emotions will help them better maintain a sustainable socio-affective relationship as well as adjust learning activity materials accordingly. Furthermore, our study has a key side benefit for the designers of the SpeakPlus platform. The feedback on learners' emotions can help them for instance monitor the performance of the learning sessions and learners' satisfaction.

In this study we opted for a post-processing (off-line) approach for emotion tracking, analysis and visualization to build a first prototype of our dashboard based on real data. Through an iterative design process, the pilot study will help us to make the dashboard evolve according to tutors' feedbacks and needs. In fact, the current architecture should be altered to offer a real-time feedback to tutors in order to optimize the learning session and make it more enjoyable.

Our study is learner-centric. As such, we also think it is important to investigate whether there are correlations between the emotions felt by the tutor and by the learner. Furthermore, our study focuses on universal discrete emotions and dimensional emotions (arousal and valence). Several recent studies (e.g. [6]) show that the learning context is more concerned with domain-specific emotions (e.g. confused, bored, frustrated, pride). Thus, investigating such emotions or combining them with universal ones might be a promising research issue for future work.

7. Acknowledgements

This research is conducted within the EmoViz project funded by the Région Auvergne-Rhône-Alpes. We thank the SpeakPlus (https://speakplus.fr) company for funding this research work. We also thank the learners and tutors who participated to the study.

8. REFERENCES

[1] Siemens, G. and Long, P. Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, (September/October 2011).

[2] Wu, J. H., Tennyson, R. D. and Hsia, T. L. A study of student satisfaction in a blended e-learning system environment. *Computers and Education*, (August 2010).

[3] Sun, P. C., Tsai, R. J., Finger, G., Chen Y. Y. and Yeh, Dowming. 2008. What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction. *Educational Psychology Review*, (May 2008).

[4] Pekrun, R. The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Computers and Education*, (2006).

[5] Boekaerts, M. The crucial role of motivation and emotions in classroom learning. In: Dumont, H., Istance, D., Benavides, F. (Eds.) The nature of learning: Using research to inspire practice. *OECD*, (2010).

[6] Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P. and Perry, R. P. 2011. Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology*, (January 2011).

[7] D'Mello, S. K., Picard, R. W., and Graesser, A. C. Towards an Affect-Sensitive AutoTutor. *Special issue on Intelligent Educational Systems – IEEE Intelligent Systems*, (July/August 2007).

[8] Mega, C., Ronconi, L., De, B. R. What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Educational Psychology*, (Feb 2014).

[9] Ruiz, S., Charleer, S., Urretavizcaya, M., Klerkx, J., Isabel Fernandez, C. I. and Duval, E. Supporting learning by considering emotions: Tracking and Visualization. *Sixth International Conference on Learning Analytics and Knowledge*. Edinburgh, UK, (April 2016).

[10] GhasemAghaei, R., Arya, A. and Biddle, R. A Dashboard for Affective E-learning: Data Visualization for Monitoring Online Learner Emotions. Proceedings of *EdMedia: World Conference on Educational Media and Technology*. Vancouver, Ca., (Jun 28, 2016).

[11] Happy, S. L., Dasgupta, A., Patnaik and P., Routray. Automated Alertness and Emotion Detection for Empathic Feedback during e-Learning. *IEEE Fifth International Conference on Technology for Education (T4E)*. Edinburgh, United Kingdom, (December 2013).

[12] Mazza, R. and V., Dimitrova. CourseVis: A Graphical Student Monitoring Tool for Supporting Instructors in Web-Based Distance Courses. *Human-Computer Studies*, (2007).

[13] May, M., George, S. and Prévôt, P. TrAVis to Enhance Self-Monitoring in Online Learning Supported by Computer-Mediated Communication Tools. *International Journal of Computer Information Systems and Industrial Management Applications*, (2011).

[14] Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S. and Karpouzis, K. Multimodal emotion recognition from expressive faces, body gestures and speech. *Artificial Intelligence and Innovations from Theory to Applications*, (2007).

[15] Pantic, M., Sebe, N., Cohn, J. and Huang, T. S. Affective Multimodal Human-Computer Interaction. *Proceedings of the 13th annual ACM international conference on Multimedia*, (November 2005).

[16] Lavoué E., Molinari G., Prié Y., Khezami S., Reflection-in-Action Markers for Reflection-on-Action in Computer-Supported Collaborative Learning Settings. *Computers & Education*, (2015).

[17] Bouvier P., Sehaba K., Lavoué E. A trace-based approach to identifying users' engagement and qualifying their engaged-behaviours in interactive systems: Application to a social game. *User Modeling and User-Adaptated Interaction (UMUAI)*, (2014).

[18] Afzal, S.and Robinson, P. Modelling Affect in Learning Environments-Motivation and Methods. *Proceedings of the 10th IEEE International Conference on Advanced Learning Technologies*, (July 2010).

[19] Michel, C., Lavoué, E., George, S., & Ji, M. Supporting Awareness and Self-Regulation In Project-Based Learning through Personalized Dashboards. *International Journal of Technology Enhanced Learning*. To be published, (2016.).

[20] Ekman, P. and Friesen, W. V. The Facial Action Coding System: A Technique for the Measurement of Facial Movement. *Environmental psychology and nonverbal behavior*, (September 1976).

[21] Barrett L. F. and Russell, J. A. 1998. Independence and Bipolarity in the Structure of Current Affect. *Personality and Social Psychology*, (1998).

[22] Scherer, K. R., Fontaine, J. R. J. and Soriano, C. CoreGRID and MiniGRID: Development and validation of two short versions of the GRID instrument. *In Fontaine, J. R. J., Scherer, K. R. and Soriano, C. (Eds.), Components of Emotional Meaning*. A sourcebook2. Oxford, UK: Oxford University Press, (2013).

[23] Picard R. W. Affective Computing. *MIT Media Lab. Perceptual Computing Section*. Technical Report, (1995).

[24] Cernea, D. and Kerren, A. 2015. A survey of technologies on the rise for emotion-enhanced interaction. *Journal of Visual Languages and Computing*, (December 2015).

[25] Konar, A. and Chakraborty, A. Emotion Recognition: A Pattern Analysis Approach. *Wiley*, (2015.).

[26] Lee, C. M. and Narayanan, S. S. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, (2005).

[27] Wu, C.H. and Liang, W. B. 2011. Emotion recognition of affectives peech based on multiple classiers using acoustic prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, (2011).

[28] Eyben, F., Wollmer, M. and Schuller, B. OpenEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. *Affective Computing and Intelligent Interaction and Workshops*. (October 2009).

[29] Mora, S., Pelayo, V. R and Muller, L. Supporting Mood Awareness in Collaborative Settings. *7th International Conference on Collaborative Computing: Networking, Applications and Worksharing, CollaborateCom*. Orlando, FL, USA, (January 2011).

[30] Posnett, D., Devanbu, P. and Filkov, V. MIC Check: A Correlation Tactic for ESE Data. *Proceedings of the 9th IEEE Working Conference on Mining Software Repositories*. IEEE Press, Piscataway, NJ, USA, (2012).

[31] Lam, H., Bertini, E., Isenberg, P., Plaisant, C. and Carpendale, S. Empirical Studies in Information Visualization: Seven Scenarios. *Vis. Comput. Graph*. IEEE Trans, (Sep. 2012).