Development process of an affective bi-modal Intelligent Tutoring System

Efthymios Alepis\textsuperscript{a,}\textsuperscript{*}, Maria Virvou\textsuperscript{a} and Katerina Kabassi\textsuperscript{b}

\textsuperscript{a}Department of Informatics, University of Piraeus, 80 Karaoli & Dimitriou St., 18534 Piraeus, Greece
\textsuperscript{b}Department of Ecology and the Environment, Technological Educational Institute of the Ionian Islands, 2 Kalvou Sq., 29100 Zakynthos, Greece

Abstract. This paper presents the development process of an Intelligent Tutoring System that performs emotion recognition on the basis of two modalities: keyboard and microphone. The system uses a multi-criteria theory to combine information about the user emotions through the two modalities. In this paper we focus on presenting and discussing the experimental studies that were conducted for the purposes of the design of the system. The experimental studies yield results concerning the way that human observers perform emotion recognition on other humans and then we show how we use these results to design the reasoning mechanism of the Intelligent Tutoring system that performs emotion recognition.

Keywords: Affective modelling, bi-modal interaction, multi-criteria decision making, software life-cycle

1. Introduction

There is an increasing interest within the HCI community in designing affective engagement with interfaces [11]. This is especially the case of computer-based educational applications that are targeted to students who are in the process of learning. Learning is a complex cognitive process and it is argued that how people feel may play an important role on their cognitive processes as well [8].

A way of improving interaction and thus learning is recognizing the users’ emotions by observing them during their engagement with the educational software and then adapting its interaction to their emotional state. Indeed, research in psychology and neurology shows that both body and mind are involved in emotions’ experiences [2,3] and emotions influence people’s body movements [4]. Therefore, observing users may provide a system with adequate information for recognising users’ emotions. Picard [17], on the other hand, argues that people’s expression of emotion is so idiosyncratic and variable, that there is little hope of accurately recognising an individual’s emotional state from the available data. Therefore, many researchers have pointed out that there is a need for combining evidence from many modes of interaction so that a computer system can generate as valid hypotheses as possible about users’ emotions (e.g. [15,16]).

In view of the above, we discuss the software engineering aspects for combining two modes of interaction, namely keyboard and microphone, through a multi-criteria decision making method called Simple Additive Weighting (SAW) [7,10]. More specifically, SAW is used for evaluating different emotions, taking into account the input of the two different modes, and selects the one that seems more likely to have been felt by the user.

The proposed approach has been incorporated in an educational application called Edu-Affe-Mikey that is targeted to first-year medical students for adapting its interaction to each user’s emotional state. For requirements analysis and the effective application of the particular approach two different experimental studies have been conducted. The experimental studies involved real end users as well as human experts. In this way the application of the multi-criteria model in the
The main body of this paper is organised as follows. In Section 2 we present the multi-criteria decision making method called SAW. In Section 3 we present the first empirical study during requirements specification and analysis. In Section 4 we present the second empirical study for defining the criteria that are taken into account while performing emotion recognition. In Section 5 we give an overall description of the system and in Section 6 we give information of how the decision making method has been applied in the system for combining evidence from two different modes and select the user’s emotion. In Section 7 the system’s capability in performing emotion recognition and providing affective interaction is evaluated. Finally, in Section 8 we give the conclusions drawn from this work and discuss ongoing work.

2. Decision making aspects

Decision processes with multiple attributes aim at modelling human decision making and judgment that takes into account several criteria and a result it gives a ranking among conflicting hypotheses. In our view, a system that performs affect recognition is an excellent candidate for embodying this kind of reasoning that is involved in decision making. An affect recognizer should be able to reproduce a human observer’s reasoning that takes into account many criteria concerning the emotional state of a fellow human and as a result recognizes the emotion of the fellow human. For this purpose, decision making theories seem very promising although they have not been used yet for this purpose.

So far the problem of emotion recognition through multiple modalities in human-computer interaction has been approached by other mathematical methods. A lot of them have been described in a comprehensive review of the field made in [12]. Such methods include rule-based systems, discriminate analysis, fuzzy rules, case-based and instance-based learning, linear and nonlinear regression, neural networks, Bayesian learning, Hidden Markov Models, Bayesian networks etc. However, multi-criteria decision making methods have not been used yet in the problem of affect recognition through multiple modalities.

The decision making theories that seem to be more appropriate for this kind of problem are the multi-attribute ones (MADM). In affect recognition that is based on multiple modalities, the criteria that a human observer may use in order to recognize the emotions of a fellow human interactant may be regarded as criteria that may be used by the human observer to select the emotion that the fellow interactant is most likely to have.

Such reasoning is not easy to model. Furthermore, as it was shown in Triantaphyllou and Mann [21] MADM methods can give different answers to the same problem. Since the truly best alternative is the same, regardless of the method chosen, an estimation of the accuracy of each method is highly desirable. The adaptation of a multi-criteria theory involves specifying the criteria
that are usually taken into account by a human decision maker and calculating the importance of each criterion in the decision maker’s reasoning process. Moreover, it involves adapting the theory into the software. Therefore, the process of the application of a decision making theory requires conducting experiments which aim at acquiring knowledge from human experts. The experiments play a crucial role in the resulting reasoning of a system. Indeed, if the experiments are not carefully designed and implemented, then there is a possibility that useful pieces of knowledge are missed out and the application of the decision making theory fails in the end.

So far, in the literature of human-computer interaction MADM methods have been used for several purposes, such as selecting the best information source when a user submits a query [14], modelling user preferences in recommender systems, selecting the best route in mobile guides or individualising e-commerce web pages. However, MADM methods have not been used for affect recognition by providing an integrating mechanism of different modalities. A multi-attribute decision problem is a situation in which, having defined a set A of actions and a consistent family F of n attributes g1, g2, . . . , gn (n ≥ 3) on A, one wishes to rank the actions of A from best to worst and determine a subset of actions considered to be the best with respect to F [22]. According to Triantaphyllou and Mann [21] there are three steps in utilising any decision making technique involving numerical analysis of alternatives:

1. Determining the relevant attributes and alternatives.
2. Attaching numerical measures to the relative importance of the attributes and to the impacts of the alternatives on these attributes.
3. Processing the numerical values to determine a ranking of each alternative.

The determination of the relevant attributes and their relative importance is made at the early stages of the software’s life-cycle and is performed by the developer or is based on an empirical study which may involve experts in the domain. However, decision making techniques mainly focus on step 3.

The Simple Additive Weighting (SAW) [7, 10] method is probably the best known and most widely used decision making method. SAW consists of two basic steps:

1. Scale the values of the n attributes to make them comparable. There are cases where the values of some attributes take their values in [0, 1] whereas there are others that take their values in [0, 1000]. Such values are not easily comparable. A solution to this problem is given by transforming the values of attributes in such a way that they are in the same interval.
2. Sum up the values of the n attributes for each alternative. As soon as the weights and the values of the n attributes have been defined, the value of a multi-attribute function is calculated for each alternative as a linear combination of the values of the n attributes.

Scaling the values of the n attributes to make them comparable is the first step in applying the SAW method. However, the way that the values of the attributes should be scaled up is not defined by SAW and, therefore, different approaches have been proposed in order to overcome this problem. The one proposed by Naumann [14] is a very good one as it transforms all values so that they take their values in the interval [0, 1] and has been applied in solving decision making problems in computer science. More specifically, Naumann [14] uses the scaling factor of Eq. (1) in order to normalise the values of attributes such as understandability, extent, availability, time and price.

\[ x_{ij} = \frac{d_{ij} - d_{j}^{\min}}{d_{j}^{\max} - d_{j}^{\min}} \]  

where \( d_{ij} \) is the old value and \( x_{ij} \) is the transformed value of the \( j \) attribute for the \( i \) alternative and \( d_{j}^{\max}, d_{j}^{\min} \) are the maximum and minimum values of the \( j \) attribute for all the alternatives. Using this scaling the values of all attributes are in the interval [0, 1].

The SAW approach consists of translating a decision problem into the optimisation of some multi-attribute utility function \( U \) defined on \( A \). The decision maker estimates the value of function \( U(X_j) \) for every alternative \( X_j \) and selects the one with the highest value. The multi-attribute utility function \( U \) can be calculated in the SAW method as a linear combination of the values of the \( n \) attributes:

\[ U(X_j) = \sum_{i=1}^{n} w_i x_{ij}, \]  

where \( X_j \) is one alternative and \( x_{ij} \) is the value of the \( i \) attribute for the \( X_j \) alternative.

3. Requirements analysis for affective bi-modal interaction

Requirement specification and analysis in the affective bi-modal intelligent tutoring system resulted from
an empirical study. The main aim of this study was to find out how users express their emotions through a bi-modal interface that combines voice recognition and input from keyboard. This empirical study involved 50 users (male and female), of the age range 17–19 and at the novice level of computer experience. The particular users were selected because such a profile describes the majority of first year medical students in a Greek university, which the educational application is targeted to. They are usually between the age of 17 and 19 and usually have only limited computing experience, since the background knowledge required for medical studies does not include advanced computer skills.

In the first phase of the empirical study these users were given questionnaires concerning their emotional reactions to several situations of computer use in terms of their actions using the keyboard and what they say. Participants were asked to determine what their possible reactions would be when they are at certain emotional states during their interaction. Our aim was to recognize the possible changes in the users’ behaviour and then to associate these changes with emotional states like anger, happiness, boredom, etc.

After collecting and processing the information of the empirical study we came up with results that led to the design of the affective module of the educational application. For this purpose, some common positive and negative feelings were identified.

The results of the empirical study were also used for designing the user stereotypes. In our study user stereotypes where built first by categorizing users by their age, their educational level and by their computer knowledge level. The reason why this was done was that people’s behaviour while doing something may be affected by several factors concerning their personality, age, experience, etc. Indeed, the empirical study revealed many cases of differences among users. For example, experienced computer users may be less frustrated than novice users. Younger computer users are usually more expressive than older users while interacting with an animated agent and we may expect to have more data from the audio mode than by the use of a keyboard. The same case is when a user is less experienced in using a computer than a user with a high computer knowledge level. In all of the above cases, stereotypes were constructed to indicate which specific characteristics in a user’s behaviour should be taken more into account in order make more accurate assumptions about the users’ emotional state.

The empirical study also revealed that the users would favor a system that would automatically adapt its interaction to the users’ emotional state. Therefore, the system could use the evidence of the emotional state of a user collected by a bi-modal interface in order to re-feed the system, adapt the agent’s behaviour to the particular user interacting with the system and as a result make the system more human-like and user-friendly.

4. Specification and analysis of multiple criteria

Any decision making method requires prior to its application the specification of important criteria. Therefore, an empirical study was conducted in order to locate the criteria that human experts take into account while performing emotion recognition. For the empirical study to be efficiently used for the purposes of the specification of criteria and their respective weights, it had to involve a satisfactory number of human experts. Therefore, in the experiment conducted for the application of the multi-criteria theory in the e-learning system, 16 human experts were selected in order to participate in the empirical study. All of the human experts possessed a first and/or higher degree in Computer Science. These experts acted as the human decision makers and were interviewed about the criteria that they took into account when providing individualized advice.

The human experts of the empirical study were asked which input action from the keyboard and the microphone would help them find out what the emotions of the users were. From the input actions that appeared in the experiment, only those proposed by the majority of the human experts were selected. In particular considering the keyboard we have: a) user types normally b) user types quickly (speed higher than the usual speed of the particular user) c) user types slowly (speed lower than the usual speed of the particular user) d) user uses the backspace key often e) user hits unrelated keys on the keyboard f) user does not use the keyboard.

Considering the users’ basic input actions through the microphone we have 7 cases: a) user speaks using strong language b) users uses exclamations c) user speaks with a high voice volume (higher than the average recorded level) d) user speaks with a low voice volume (low than the average recorded level) e) user speaks in a normal voice volume f) user speaks words from a specific list of words showing an emotion g) user does not say anything.

Concerning the combination of the two modes in terms of emotion recognition we came to the conclusion that the two modes are complementary to each other to
a high extent. In many cases the human experts stated that they can generate a hypothesis about the emotional state of the user with a higher degree of certainty if they take into account evidence from the combination of the two modes rather than one mode. Happiness has positive effects and anger and boredom have negative effects that may be measured and processed properly in order to give information used for a human-computer affective interaction. For example, when the rate of typing backspace of a user increases, this may mean that the user makes more mistakes due to a negative feeling. However this hypothesis can be reinforced by evidence from speech if the user says something bad that expresses negative feelings.

5. Overview of the system

In this section, the overall functionality and emotion recognition features of our system, Edu-Affe-Mikey is described. The architecture of Edu-Affe-Mikey consists of the main educational application with the presentation of theory and tests, a programmable human-like animated agent, a monitoring user modelling component and a database.

While using the educational application from a desktop computer, students are being taught a particular medical course. The information is given in text form while at the same time the animated agent reads it out loud using a speech engine. The student can choose a specific part of the human body and all the available information is retrieved from the systems’ database. In particular, the main application is installed either on a public computer where all students have access, or alternatively each student may have a copy on his/her own personal computer. An example of using the main application is illustrated in Fig. 1. The animated agent is present in these modes to make the interaction more human-like.

While the users interact with the main educational application and for the needs of emotion recognition a monitoring component records the actions of users from the keyboard and the microphone. These actions are then processed in conjunction with the multi-criteria model and interpreted in terms of emotions. The basic function of the monitoring component is to capture all
the data inserted by the user either orally or by using the keyboard and the mouse of the computer. The data is recorded to a database and the results are returned to the basic application the user interacts with. Figure 2 illustrates the “monitoring” component that records the user’s input and the exact time of each event.

Instructors have also the ability to manipulate the agents’ behaviour with regard to the agents’ on screen movements and gestures, as well as speech attributes such as speed, volume and pitch. Instructors may programmatically interfere to the agent’s behaviour and the agent’s reactions regarding the agents’ approval or disapproval of a user’s specific actions. This adaptation aims at enhancing the “affectiveness” of the whole interaction. Therefore, the system is enriched with an agent capable to express emotions and, as a result, enforces the user’s temper to interact with more noticeable evidence in his/her behaviour.

Figure 3 illustrates a form where an instructor may change speech attributes. Within this context the in-
Fig. 4. Programming the behaviour of animated agents depending on particular students’ actions.

The instructor may create and store for future use many kinds of voice tones such as happy tone, angry tone, whisper and many others depending on the need of a specific affective agent-user interaction. In some cases a user’s actions may be rewarded with a positive message by the agent accompanied by a smile and a happy tone in the agent’s voice, while in other cases a more austere behaviour may be desirable for educational needs. Figure 4 illustrates how an instructor may set possible actions for the agent in specific interactive situations while a user takes a test.

6. Application of the decision making method

For the evaluation of each alternative emotion the system uses SAW for a particular category of users. This particular category comprises of the young (under the age of 19) and novice users (in computer skills). The likelihood for a specific emotion (happiness, sadness, anger, surprise, neutral and disgust) to have occurred by a specific action is calculated using the formula below:

\[
em_{1e_1} + em_{1e_2}
\]

\[
\frac{2}
\]

where

\[
em_{1e_1} = w_{1e_1k_1}k_1 + w_{1e_1k_2}k_2 + w_{1e_1k_3}k_3
\]

\[
+ w_{1e_1k_4}k_4 + w_{1e_1k_5}k_5 + w_{1e_1k_6}k_6
\]

\[
em_{1e_2} = w_{1e_1m_1}m_1 + w_{1e_1m_2}m_2 + w_{1e_1m_3}m_3
\]

\[
+ w_{1e_1m_4}m_4 + w_{1e_1m_5}m_5
\]

\[
+ w_{1e_1m_6}m_6 + w_{1e_1m_7}m_7
\]

\[
(3)
\]

\[
(4)
\]

**em**\(_{1e_1}\) is the probability that an emotion has occurred based on the keyboard actions and **em**\(_{1e_2}\) is the probability that refers to an emotional state using the users’ input from the microphone. These probabilities result from the application of the decision making model of SAW and are presented in Eqs (3) and (4) respectively. \(em_{1e_1}\) and \(em_{1e_2}\) take their values in \([0,1]\).

In Eq. (3) the \(k\)’s from \(k_1\) to \(k_6\) refer to the six basic input actions that correspond to the keyboard. In Eq. (4) the \(m\)’s from \(m_1\) to \(m_7\) refer to the seven basic input actions that correspond to the microphone. These variables are Boolean. In each moment the system takes data from the bi-modal interface and translates them in terms of keyboard and microphone actions. If an action has occurred the corresponding criterion takes the value 1, otherwise its value is set to 0. The \(w\)’s represent the weights. These weights correspond to a specific emotion and to a specific input action and are acquired by the stereotype database. More specifically, the weights are acquired by the stereotypes about the emotions.

In order to identify the emotion of the user interacting with the system, the mean of the values that have occurred using Eqs (3) and (4) for that emotion is estimated. The system compares the values from all the different emotions and determines whether an emotion is taking effect during the interaction. As an example we give the two formulae with their weights for the two modes of interaction that correspond to the emotion of happiness when a user (under the age of 19) gives the correct answer in a test of our educational application.
In case of \( em_{1e1} \) considering the keyboard we have:

\[
em_{1e1} = 0.4k_1 + 0.4k_2 + 0.1k_3 + 0.05k_4 + 0.05k_5 + 0k_6
\]

In this formula, which corresponds to the emotion of happiness, we can observe that the higher weight values correspond to the normal and quickly way of typing. Slow typing, often use of the backspace key and use of unrelated keys are actions with lower values of stereotypic weights. Absence of typing is unlikely to take place. Concerning the second mode (microphone) we have:

\[
em_{1e2} = 0.06m_1 + 0.18m_2 + 0.15m_3 + 0.02m_4 + 0.14m_5 + 0.3m_6 + 0.15m_7
\]

In the second formula, which also corresponds to the emotion of happiness, we can see that the highest weight corresponds to \( m6 \) which refers to the ‘speaking of a word from a specific list of words showing an emotion’ action. The empirical study gave us strong evidence for a specific list of words. In the case of words that express happiness, these words are more likely to occur in a situation where a novice young user gives a correct answer to the system. Quite high are also the weights for variables \( m2 \) and \( m3 \) that correspond to the use of exclamations by the user and to the raising of the user’s voice volume. In our example the user may do something orally or by using the keyboard or by a combination of the two modes. The absence or presence of an action in both modes will give the Boolean values to the variables \( k1 \ldots k6 \) and \( m1 \ldots m7 \).

A possible situation where a user would use both the keyboard and the microphone could be the following: The specific user knows the correct answer and types in a speed higher than the normal speed of writing. The system confirms that the answer is correct and the user says a word like ‘bravo’ that is included in the specific list of the system for the emotion of happiness. The user also speaks in a higher voice volume. In that case the variables \( k1, m3 \) and \( m6 \) take the value 1 and all the others are zeroed. The above formulae then give us \( em_{1e1} = 0.4 \times 1 = 0.4 \) and \( em_{1e2} = 0.15 \times 1 + 0.3 \times 1 = 0.45 \).

In the same way the system then calculates the corresponding values for all the other emotions using other formulae. For each basic action in the educational application and for each emotion the corresponding formula have different weights deriving from the stereotypical analysis of the empirical study. In our example in the final comparison of the values for the six basic emotions the system will accept the emotion of happiness as the most probable to occur.

7. Evaluation

Fifty medical students were involved in the evaluation of the multi-criteria emotion recognition system. All users were asked to work with Edu-Affe-Mikey and the whole interaction was video recorded. Then the videos collected were presented to the users that participated the experiment in order to perform emotion recognition for themselves with regard to the six emotional states, namely happiness, sadness, surprise, anger, disgust and the neutral emotional state. Whenever a participant recognized an emotional state, the emotion was marked and stored as data in the system’s database. Finally, after the completion of the experimental study, the data were compared with the systems’ corresponding hypothesis in each case an emotion was detected.

The evaluation revealed that the emotion that is most difficult to be recognized by the two modes of system was ‘surprise’. The system’s success in identifying the particular emotion was 45%. However, the system’s success in identifying the particular emotion would be decreased to 8% if the system performed emotion recognition only through the keyboard. The emotions that were most easily successfully recognized were sadness and anger. More specifically, Edu-Affe-Mikey’s success in identifying these emotions was 70%. The system’s success in recognizing sadness and anger would be decreased to 34% and 42%, respectively, if the system relied on emotion recognition only through the keyboard.

The system’s success in recognizing different emotions ranged from 45% to 70%, which is quite satisfactory. Indeed, recent empirical studies (e.g. [20]) have revealed that even humans are not 100% successful in recognising other people’s emotions. Furthermore, the evaluation revealed that the success of the system was significantly increased using two modes instead of one and combining their results through SAW.

8. Discussion and conclusions

In this paper we have described an affective educational application that recognizes students’ emotions based on their words and actions that are identified by
the microphone and the keyboard, respectively. The system uses an innovative approach that combines evidence from the two modes of interaction based on user stereotypes and a multi-criteria decision making theory.

The main advantage of the proposed approach is that the whole process is based on experimental studies that involve real users. More specifically, potential users of the software have participated in an empirical study during requirements specification and analysis for the design of the user stereotypes as well as during the final evaluation of the system. Additionally, human experts have been involved in requirements analysis and design for the selection the criteria that are used in their reasoning process and estimating their weights of importance.

Another experimental study was conducted for evaluating the system’s capability in performing emotion recognition. More specifically, 50 users interaction with the educational application was video recorded and the videos collected were presented to the same users, who were asked to comment on their emotion. The emotions the users identified were compared to the emotions identified by the system. This comparison revealed that the system could adequately identify the users’ emotion. However, its hypotheses were more accurate when there was a combination of the evidence from two different modes using the multi-criteria decision making theory.

This kind of work is novel in the area of affect recognition through multiple modalities. First, there is a shortage of empirical studies that involve humans as users and as observers. Most empirical studies give results from the point of view of a computer system that performs affect recognition [1,9,23]. However, empirical studies that show how humans perceive and reason about other people’s emotions is very important for understanding the area of affect recognition. A similar empirical study in this respect is the one conducted by De Silva and colleagues [5]. Still, in that empirical study, De Silva and colleagues had focused on different aspects from the ones of our empirical study. For example, De Silva et al. focus on acoustic and visual issues that result from the signal processing of the voice and the image of users, whereas in our study we focused on linguistic aspects of the audio mode (what the users say) and combine this with keyboard actions. Moreover, on top of the results of the empirical study, we have proposed a quite novel and original approach for combining modalities sing multi-criteria decision making so that we may reproduce the reasoning of a human observer.

In future work we plan to improve our system by the incorporation of stereotypes concerning users of several ages, educational backgrounds and computer knowledge levels. Moreover, there is ongoing research work in progress that exploits a third mode of interaction, visual this time [19], to add information to the system’s database and complement the inferences of the user modelling component about users’ emotions. The third mode is going to be integrated to our system by adding cameras and also providing the appropriate software, as for future work.

Acknowledgements

Support for this work was provided by the General Secretariat of Research and Technology, Greece, under the auspices of the PENED-2003 program.

References


