

Chapter 2

On the Use of Multi-attribute Decision Making for Combining Audio-Lingual and Visual-Facial Modalities in Emotion Recognition

Maria Virvou, George A. Tsihrintzis, Efthimios Alepis,
Ioanna-Ourania Stathopoulou and Katerina Kabassi

Abstract In this chapter, we present and discuss a novel approach that we have developed for the integration of audio-lingual and visual-facial modalities in a bi-modal user interface for affect recognition. Even though researchers acknowledge that two modalities can provide information that is complementary to each other with respect to affect recognition, satisfactory progress has not yet been achieved towards the integration of the two modalities. In our research reported herein, we approach the combination of the two modalities from the perspective of a human observer by employing a multi-criteria decision making theory for dynamic affect recognition of computer users. Our approach includes the specification of the strengths and weaknesses of each modality with respect to affect recognition concerning the 6 basic emotion states, namely *happiness*, *sadness*, *surprise*, *anger* and *disgust*, as well as the emotionless state which we refer to as *neutral*. We present two empirical studies that we have conducted involving

M. Virvou (✉) · G.A. Tsihrintzis · E. Alepis · I.-O. Stathopoulou
Software Engineering Lab, Department of Informatics, University of Piraeus, 185 34 Piraeus,
Greece
e-mail: mvirvou@unipi.gr

G.A. Tsihrintzis
e-mail: geoatsi@unipi.gr

E. Alepis
e-mail: talepis@unipi.gr

I.-O. Stathopoulou
e-mail: iostath@unipi.gr

K. Kabassi
Department of Restoration and Conservation of Cultural Heritage, Technological Educational
Institute of the Ionian Islands, 291 00 Zakynthos, Greece
e-mail: kkabassi@teion.gr

human users and human observers concerning the recognition of emotions from audio-lingual and visual-facial modalities. Based on the results of the empirical studies, we assign weights to criteria for the application of a multi-criteria decision making theory. Additionally, the results of the empirical studies provide information that may be used by other researchers in the field of affect recognition and is currently unavailable in the relevant literature.

Keywords Empirical studies • Affective computing • Facial expression analysis • Multi-modal interfaces • Audio-lingual affect recognition • Visual-facial affect recognition • Multi-criteria decision making

2.1 Introduction

The recognition of emotions can lead to affective user interfaces that take into account users' feelings and can adapt their behaviour accordingly. Neurologists and psychologists have made progress in demonstrating that emotion is at least as and perhaps even more important than reason in the process of decision making and action deciding [1]. Moreover, the way people feel may play an important role in their cognitive processes as well [2]. This important motivation has led to a wealth of recent research efforts toward the recognition of emotions of users while they interact with software applications. Picard points out that one of the major challenges in affective computing is to try to improve the accuracy of recognising people's emotions [3]. Improving the accuracy of emotion recognition may imply the combination of many modalities in user interfaces. Indeed, human emotions are usually expressed in several ways. Human faces, people's voices, or people's actions may all show emotions. As we articulate speech, for example, we usually move the head and exhibit various facial emotions [4]. Moreover, humans convey emotional information both intentionally and unintentionally via speech patterns; these vocal patterns are perceived and understood by listeners during conversation [5]. Hence the typical keyboard-mouse input device may now seem too limited for the goal of emotion recognition.

Ideally, evidence from many modalities of interaction should be combined by a computer system so that it can generate as valid hypotheses as possible about users' emotions. It is hoped that the multimodal approach may provide not only better performance, but also more robustness [6]. Similar views about the benefits of the combination of modalities have been supported by many researchers in the field of human-computer interaction [7–11]. However, progress in emotion recognition based on multiple modalities has been rather slow. Although several approaches have been proposed to recognize human emotions based on facial expressions or speech unimodally, relatively limited work has been done to fuse these two and other modalities to improve the accuracy and robustness of the emotion recognition systems [12].

In view of the above, it is the aim of this chapter to combine modalities in order to improve the accuracy and overall performance of emotion recognition. In this chapter, we focus on the combination of the audio-lingual modality and the visual-facial modality. Our approach for the combination of the two modalities is by employing Multi-Attribute Decision Making, henceforth referred to as MADM.

As a first goal of an attempt to combine the audio-lingual modality and the visual-facial modality, one has to determine the extent to which these two different modalities can provide emotion recognition from the perspective of a human observer. Moreover, one has to specify the strengths and weaknesses of each modality. In this way, one can determine the weights of the criteria that correspond to the respective modalities from the perspective of a human observer. Hence, empirical studies need to be conducted first concerning emotion recognition based on two modalities: the audio-lingual and the visual-facial.

Such empirical studies constitute an important milestone and yield important results. Not only do they provide the basis towards the combination of modalities into the affective user modelling component of a bi-modal system, but they also give evidence for other researchers to use since, currently, there are not many results from such empirical studies in the literature. Indeed, after an extensive search of the literature, one finds that there is a shortage of empirical evidence concerning the strengths and weaknesses of these modalities. In this chapter, we present each of the empirical studies and show and discuss results from their comparison. Moreover, we present and discuss a novel integration approach that we have developed, which is based on the results of the empirical studies as they have been incorporated in a multi-criteria decision making theory. Specifically, in this chapter, emphasis is placed on six basic emotions, namely *happiness*, *sadness*, *surprise*, *anger* and *disgust*, as well as the emotionless state, which we refer to as *neutral*.

The main body of the chapter is organised as follows. In Sect. 2.2, we present and discuss related work. We also show how important it is for the efficient application of MADM into the problem of affect recognition to have conducted empirical studies using human subjects. In Sect. 2.3, we present and discuss the aims of our empirical studies and their settings. In Sect. 2.4, we describe the empirical study concerning the audio-lingual modality. In Sect. 2.5, we describe the empirical study concerning the visual-facial modality. In Sect. 2.6, we discuss the results produced by the empirical studies. In Sect. 2.7, we describe the architecture of our bi-modal affect recognizer and we show how the results of the empirical studies have been combined using MADM in order to integrate information for affect recognition from the audio-lingual and visual-facial modalities. In Sect. 2.8, we present the conclusions drawn from the empirical studies and point to related future work.

2.2 Related Work

The issue of combining two modalities raises the problem of how these modalities may be combined. In fact, the mathematical tools and theories that have been used for affect recognition can lead to a classification of affect recognizers. Such classification

has been made in [13] where affect recognizers have been classified into two groups on the basis of the mathematical tools that these recognizers have used: (1) The first group using traditional classification methods in pattern recognition, including rule-based systems, discriminate analysis, fuzzy rules, case-based and instance-based learning, linear and nonlinear regression, neural networks, Bayesian learning and other learning techniques. (2) The second group of approaches using Hidden Markov Models, Bayesian networks etc. Indeed, a recent piece of research that uses the above approaches for the integration of audio-visual evidence is reported in [14]. Specifically, for person-dependent recognition, Zeng and his colleagues [11] apply the voting method to combine the frame-based classification results from both audio and visual channels. For person-independent tests, they apply multi-stream hidden Markov models (HMM) to combine the information from multiple component streams.

In contrast with all of the above tools, in this chapter, we present a solution that we have developed to the problem of *bi-modal* affect recognition based on the use of the Multiple Attribute Decision Making (MADM) methodology. MADM involves making preference decisions (such as evaluation, prioritisation, selection) over the available alternatives that are characterised by multiple, usually conflicting attributes [14]. In our case, the information from the two modalities is combined using a multi-criteria decision theory [14, 15]. More specifically, each modality is considered as a criterion that a human observer would use in order to recognise another human's emotion. Thus, we have focused on how humans perceive other humans' emotions and utilise this information to create criteria for the system to use. Our aim is to render bi-modal affect recognition as human-like as possible.

The motivation underlying this particular approach stems from our belief that affect recognition in computerized systems should be performed in ways that are similar to those of human-human interaction, so that affective human-computer interaction may become more natural. This is a belief that is also expressed by other researchers in the field. For example, Nasoz and Lisetti [16] express the belief that human-human interactions should form the basis for the design of human-computer interfaces since it is important to facilitate a natural and believable interaction between the user and the computer. The incorporation of other human reasoning theories, such as Human Plausible Reasoning, has been done successfully in past user interfaces [17].

MADM can provide a formal mathematical tool for the computer to judge which particular emotion the human computer-user is experiencing based on evidence from the visual-facial and audio-lingual modalities. In many cases, an emotion may be better recognized by a particular modality. For example, if a user is seen by a camera smiling without saying anything then this user is likely to have experienced *happiness*, based on the visual-facial modality while there is no evidence of emotion from the audio-lingual modality. On the other hand, if a user swears at the computer then this user is likely to have experienced *anger*, based on the audio-lingual modality irrespective of whether there is any additional evidence on the user's emotions from the visual-facial modality. A smaller and different part of our approach that combines two other modalities using MADM, namely the keyboard and the audio-lingual, is described in [18].

Decision making theories seem very promising although they have not yet been used for this purpose for combining modalities of interaction from the point of view of a human observer. Decision processes with multiple attributes deal with human judgement that takes into account several criteria and pinpoints to the best possible result among conflicting hypotheses. 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 who s/he interacts with, may be regarded as criteria that may be used by the human observer to select the emotion that the fellow human is most likely to have.

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 incorporating 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 [19].

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 [20], improving intelligent user interfaces [21], modelling user preferences in recommender systems [22], selecting the best route in mobile guides [23], or individualising e-commerce web pages [24, 25]. However, MADM methods have not been used for affect recognition by providing an integrating mechanism of different modalities.

2.2.1 *Multi-attribute Decision Making*

In more detail, 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 g_1, g_2, \dots, g_n ($n \geq 3$) 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 [26]. According to Triantaphyllou and Mann [27], 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 life-cycle and is performed by the developer or is based on an empirical study which may involve experts in the domain.

There are several MADM theories. Among them, we have used in our research the Simple Additive Weighting (SAW) [14, 15] method, which is probably the best known and most widely used decision making method. SAW 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 $U(X_j)$ can be calculated in the SAW method as a linear combination of the values of the n attributes:

$$U(X_j) = \sum_{i=1}^n w_i x_{ij}, \quad (2.1)$$

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

An empirical study, at the early stages of the development of a MADM system, aims at determining the weights of importance of the criteria that human experts take into account while observing users interacting with a computer. In the next section, we describe the empirical studies that we conducted concerning affect recognition by human observers.

2.3 Aims and Settings of the Empirical Studies

Affect recognition aims at recognizing human emotions. However, there are too many emotions for all to be distinct and equally basic. The thesis that there are basic emotions is not implausible [28]. In our empirical studies, we focus on the recognition of six basic emotions with respect to audio-lingual and visual-facial modalities. Specifically, the six basic emotional states are *happiness*, *sadness*, *surprise*, *anger* and *disgust* as well as the emotionless state, which we refer to as *neutral*. These six emotional states in our study were selected because they were fundamental in the majority of previous emotion recognition studies. Indeed, most of these emotional states have been selected by several researchers and theorists over the past years. For example, the emotions of anger happiness and surprise are considered in many research works [29–45].

For the purposes of our research we conducted two empirical studies concerning the audio-lingual and visual-facial modalities of interaction. The audio-lingual and the visual-facial empirical studies were conducted simultaneously in two different phases and their results were compared and combined in the end. There were two kinds of roles for the participants in both empirical studies: (1) The human participants who were used to express emotions in the modalities in question. The expressions of emotion in different situations were recorded in computer protocols. (2) The human participants who served as observers and recognisers of the emotions that were recorded in the computer protocols. The human participants of both roles were selected from the same groups for both experiments. In fact, there were 300 participants of various ages and educational background that

were used for the elicitation of emotions. Moreover, there were 50 participants of similar background that were used as observers. The experiment took place in Greece, therefore the participants of both roles were all Greek, so that their emotion expressions and recognition respectively, were compatible with the Greek culture and average temper of the people living in Greece. In our view, empirical studies of this kind should be culture-dependent for more reliable results.

The aims of each of the empirical studies were the following:

1. To elicit and record emotion expressions of human participants in the respective modalities.
2. To create databases of known expressions of emotions in the respective modalities. Specifically, in the audio-lingual modality, the aim was to create a database of typical words/phrases connected to each of the basic emotions. On the other hand, in the visual-facial modality, the aim was to create a database of facial expressions for each of the basic emotions.
3. To specify degrees of recognisability of each of the basic emotions by human listeners or visual observers for the respective modalities. This means that we aimed at estimating the extent to which a particular emotion (e.g. happiness) is recognizable by each of the modalities investigated. Such findings are useful towards combining the audio-lingual and the visual-facial modalities in the sense that we specify the strengths and weaknesses of each modality with respect to each of the basic emotions. The comparison of the results is important for the application of the multi-criteria theory.
4. To specify important parts (criteria) of recognisability in the respective modalities. For example, in the audio-lingual modality, the goal was to find out how recognition could be achieved based on the actual words that a user had said as well as the volume of the user's voice. In the visual-facial modality, the goal was to classify emotion recognisability from specific portions of the face (e.g. mouth, eyes, eye brows etc.).

The settings of the two empirical studies were the same. However, one important difference of the experiments was connected to the questions asked to human observers concerning recognisability of emotions via the two different modalities. The elicitation of emotions of subjects and the creation of databases of emotions were based on the human participants of the empirical studies who expressed emotions. On the other hand, the analysis of recognisability of emotions was based on human participants that were used as observers of the recorded expression of emotions.

2.3.1 Elicitation of Emotions and Creation of Databases

As a first step of the empirical studies we had to ensure that we could elicit emotions of human participants so that these could be recorded in databases for further analysis and use. The elicitation of users' emotions constitutes an important part of the settings of empirical studies concerning affective computing. For

example, Nasoz and Lisetti [16] conducted an experiment, where they elicited six emotions (sadness, anger, surprise, fear, frustration, and amusement) and measured physiological signals of subjects by showing them movie clips from different movies (The Champ for sadness, Schindler's List for anger, The Shining for fear, Capricorn One for surprise, and Drop Dead Fred for Amusement) and by asking them hard math questions (for frustration). They conducted a panel study to choose an appropriate movie clip to elicit each emotion. In contrast, our empirical studies aimed at recording emotions for different modalities; we did not aim at recording physiological signals of users but rather aimed at recording and analyzing visual-facial and audio-lingual expressions of emotions. For the purposes of our studies, we used an educational application to elicit emotions of subjects. Indeed, the educational procedure can provoke emotions. For example, Nasoz and Lisetti [16] have used hard mathematical problems for the elicitation of frustration of subjects.

In our empirical study we created and used an educational application for all six emotions that we were aiming at recording and analyzing. The educational application incorporated a monitoring module that was running unnoticeably in the background. Moreover, users were also video taped while they interacted with the application.

More specifically, the experimental system consisted of the main educational application with the presentation of theory and tests, a programmable human-like animated agent, a user monitoring component, and a database. While using the educational application from a desktop computer, participants were being taught a particular medical course. We have selected medicine as a subject area as it is an area of high interest for many people nowadays and we also considered it more appropriate for the elicitation of many different emotions.

The information was given in text form while at the same time the animated agent read it out loudly using a speech engine. Participants could choose a specific part of the human body and all the available information was retrieved from the system database. An example of using the main application is illustrated in Fig. 2.1. An animated agent was present in these modalities to make the interaction more human-like. The educational application incorporated the capability to manipulate the agent behaviour with regards to movements and gestures of the agent on screen, as well as speech attributes such as speed, volume, and pitch. As a result, the system was enriched with an agent, who was capable to express emotions and, thereby the user was further encouraged to interact with more noticeable emotional evidence in his/her behaviour. Participants were also expected to take tests concerning part of the selected medical theory they had chosen. Both in the reading theory interaction, as well as in the interaction with the medical examinations, participants were expected to express their feelings freely, as a consequence of their interaction with the agent and with the educational environment in general.

Indeed, for the elicitation of the six basic emotions the educational application provided the following situations:

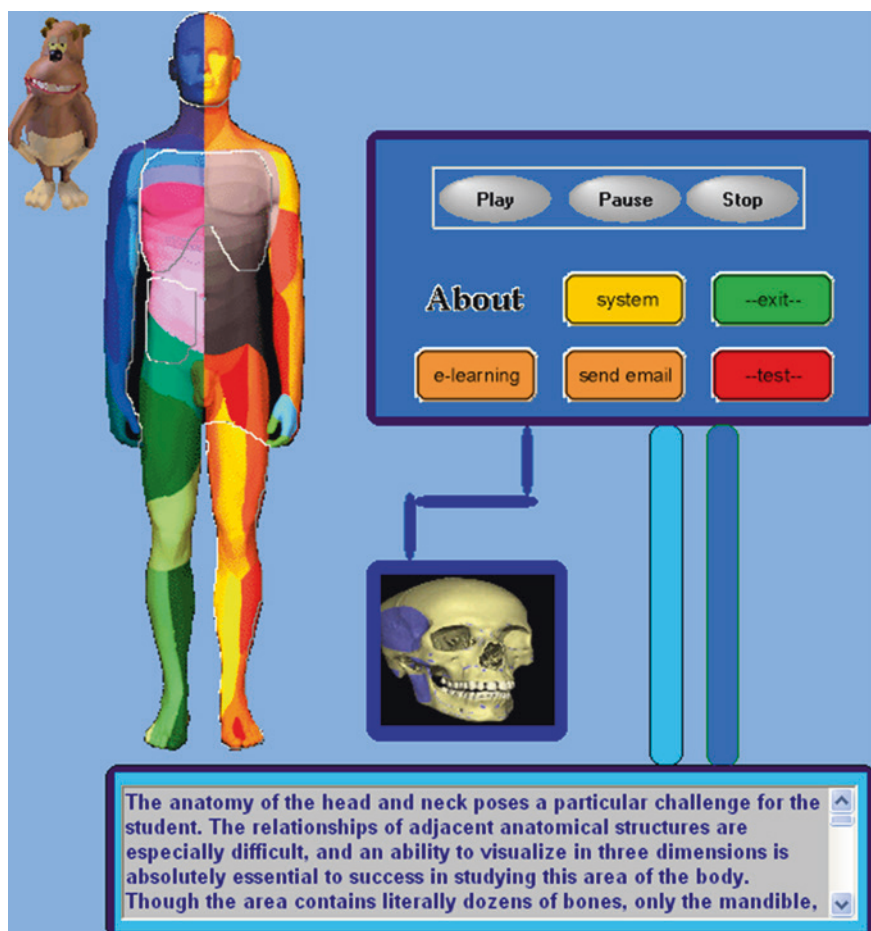


Fig. 2.1 A screen-shot of theory presentation in the medical educational application

1. For happiness: the agent would say jokes and would make funny faces or the user would receive excellent grades in his/her performance on tests.
2. For anger: the agent would be rude and unfair to users for their performance at tests.
3. For disgust: the medical application would show videos of surgical operations showing blood and human organs.
4. For sadness: the medical application would show pictures of patients while they were receiving treatments for serious diseases.
5. For surprise: the agent would pop up into the screen completely unexpectedly.
6. For neutral: the user would use the medical application under normal circumstances.

To ensure that users would experience the above emotions, all of the participants did not have a medical background and, therefore, they were not exposed frequently to this kind of material.

2.3.2 Creation of Databases of Known Expressions of Emotions

As stated above, we aimed at creating databases of known expressions of emotions for the visual-facial and audio-lingual modalities, respectively. This part of the empirical studies could have been skipped if there already existed databases of this kind that we could use for further analysis. Indeed, in the case of visual-facial modality, several face databases are available today, which have been developed by various researchers and could be accessed through the World Wide Web. These include: (1) The AR Face Database [46], which contains over 4,000 color images of 126 persons' faces in front view, forming various facial expressions under various illumination conditions and possible occlusion (e.g., by sun glasses and/or scarf). The main disadvantage of this database is its limitation to containing only four facial expressions, namely "neutral", "smiling", "anger", and "scream". (2) The Japanese Female Facial Expression (JAFPE) Database [47], which contains 213 images of the neutral and 6 additional basic facial expressions as formed by 10 Japanese female models. (3) The Yale Face Database [48], which contains 165 gray-scale GIF-formatted images of 15 individuals. These correspond to 11 images per subject of different facial expression or configuration, namely, center-light, with glasses, happy, left-light, without glasses, normal, right-light, sad, sleepy, surprised, and wink. (4) The Cohn-Kanade AU-Coded Facial Expression Database [49], which includes approximately 2000 image sequences from over 200 subjects and is based on the Facial Action Coding System (FACS), first proposed by Paul Ekman [30]. (5) The MMI Facial Expression Database [50], which includes more than 1500 samples of both static images and image sequences of faces in front and side view, displaying various expressions of emotion and single and multiple facial muscle activation.

Although many of the aforementioned face databases were considered for further analysis by human observers, we found that either the number of different facial expressions or the scope of the facial expression formation process did not entirely match our goals and studies. Furthermore, the aforementioned databases were constructed by photographing persons coming from diverse cultures. In our view, this would create a problem for the recognisability of emotions expressed because people in different cultures may have different ways of expressing emotions. Thus, we made the decision to create our own facial expression database [51] by photographing Greek people expressing the six emotional states mentioned previously.

In the case of audio-lingual modality there was no existing database that we could have used. Therefore, we certainly had to create our own database for the purposes of our research.

2.3.3 Analysis of Recognisability of Emotions by Human Observers

Recognizability of emotions by human observers consisted of specifying degrees of recognisability of each of the basic emotions by human listeners and visual observers for the respective audio-lingual and visual-facial modalities.

The empirical studies involved a total number of 300 male and female users of various educational backgrounds, ages, and levels of familiarity with computers. These users were asked to use the medical educational application and their actions were video recorded. Then, after they had completed their interaction with the application, participants were asked to watch the video clips concerning exclusively their personal interaction and to determine in which situations they experienced changes in their emotional state. Then, they associated each change in their emotional state with one of the six basic emotion states in our study and the data was recorded and time-stamped. In this way, we had recorded what the actual emotions of the users were.

2.4 Empirical Study for Audio-Lingual Emotion Recognition

The first empirical study that we have conducted concerns audio-lingual emotion recognition. In this empirical study, the audio-lingual modality of interaction is based on using a microphone as input device. The empirical study aimed at identifying common user reactions that express user emotions while they interact with computers. As a next step, we associated the reactions with particular emotions.

2.4.1 The Experimental Educational Application for Elicitation of Emotions

The participants were asked to use the medical educational application, which incorporated a user monitoring module. Figure 2.2 illustrates a snapshot of the monitoring application module while it records the user microphone input and the exact time of each event. In this study, we took into consideration only the data from the audio-lingual modality of interaction.

2.4.2 Audio-Lingual Modality Analysis

The basic function of this experiment was to capture the data inserted by the user orally. The data was recorded into a database of audio-video clips. The monitoring component recorded the actions of users with the microphone.

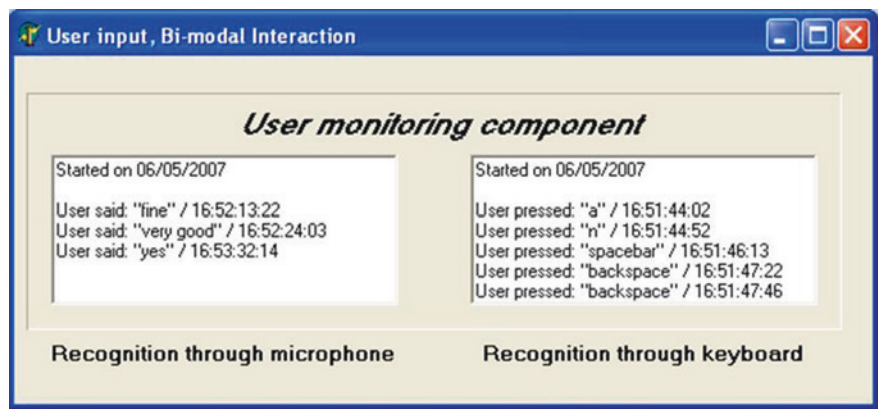


Fig. 2.2 Snapshot of operation of the user monitoring component

In the next step, the collected transcripts were given to 300 users as human expert-observers, who were asked to perform emotion recognition with regards to the six emotional states, namely happiness, sadness, surprise, anger, disgust and neutral. Successful recognition was considered, when an observer could recognize the emotion that the user had confirmed that s/he had experienced at a particular state of the video. All human expert-observers were asked to analyze the data corresponding to the audio-lingual input only. Therefore, they were asked to listen to the video tapes without seeing them. They were also given what the user had said in printed form from the computer audio recorder. The human expert-observers were asked to justify the recognition of an emotion by indicating the weights of the criteria that they had used in terms of specific words and exclamations, as well as the volume of voice.

This part of the study has supplied us with a significant number of words, phrases and exclamations, all associated with emotional states, which are used in the creation of a database of words for emotion recognition. Thus, a database has been constructed from the users' oral input and words, phrases and exclamations. At the same time changes in voice volume and voice pitch were also recorded in relation to the oral input.

Table 2.1 illustrates the results of the empirical study in terms of the oral input via microphone and the six basic emotions (neutral, happiness, sadness, surprise, anger, disgust) recognized by human expert-observers. For each emotion, we note the percentages of the users' oral reaction or the absence of audio input. Furthermore, Table 2.1 illustrates the changes in the users' voice while uttering a word or phrase or saying an exclamation in each emotional situation. For example, a user in surprise may have uttered an exclamation (58 %) rather than having spoken a word (24 %) or having remained silent (18 %). This action is recognised at a degree of 66 % accompanied by an increase in the user's voice volume. A summary of the results is illustrated in Table 2.1.

It should be noted that bored (neutral emotion) users and sad users were orally less expressive than users in other emotional states. However, in cases where bored

Table 2.1 Empirical study results

	Say an exclamation	Utter a certain word or phrase	Keep silent
Emotions	Change in volume (%)	Change in volume (%)	Change in volume (%)
Neutral	6	22	72
	45	37	
Happiness	31	45	24
	40	55	
Sadness	8	28	64
	52	44	
Surprise	58	24	18
	66	60	
Anger	39	41	20
	62	70	
Disgust	50	39	11
	64	58	

Human recognition of basic emotional states through microphone

Table 2.2 Percentages of successful emotion recognition by human experts

Emotional state	Percentage of recognition by human experts (audio data) (%)
Neutral	18
Happiness	46
Sadness	48
Surprise	62
Anger	79
Disgust	57

users actually said something, this could trigger a decrease in their voice volume. Users experiencing the emotions of surprise and disgust usually express them by saying an exclamation (58 and 50 %, respectively) while happy users and users in anger would likely say a word or phrase contained in our ‘emotional database’ of words and phrases (45 and 41 %, respectively). Particularly, with regards to the emotions of surprise and anger, users would increase the volume of their voice while saying something. Table 2.2 illustrates the percentages of successful emotion recognition by human experts concerning the participants’ emotional states and the audio modality of interaction. These percentages result after comparison of the recognized emotional states by the human experts and the actual emotional states as recorded by the participants themselves.

In the next step of our study, the participants were asked to specify, which input action from 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. Considering the users’ basic input actions through the microphone we have 7 cases: (a) user speaks using strong language, (b) user uses exclamations, (c) user speaks with a high

voice volume (higher than the average recorded level), (d) user speaks with a low voice volume (lower 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.

These input actions are considered as criteria for evaluating all different emotions and selecting the one that seems to be prevailing. More specifically, in the resulting audio-lingual emotion recognition system, each emotion is evaluated using the criteria (input actions) from the microphone. These criteria are weighted according to the analysis of the empirical studies so that for each emotional state, these seven input action criteria obtain specific values. For the evaluation of each alternative emotion, the system uses SAW for each particular category of users. The overall functionality of this approach (uni-modal recognition of emotions through audio-lingual data), described in [52], requires more comprehensive writing and is beyond the scope of the present chapter. In this chapter, we focus specifically on the combination of pre-given bi-modal information concerning emotion recognition, through a multi-attribute decision making theory, in order to improve the emotion recognition capability of a system.

2.5 Empirical Study for Visual-Facial Emotion Recognition

The basic aim of this experiment was to identify and quantify the most common facial expressions of a user during a typical human-computer interaction session. Firstly, we observed humans during human-computer interaction sessions and concluded that the facial expressions corresponding to the “neutral”, “happiness”, “sadness”, “surprise”, “anger”, and “disgust” emotional states arose very commonly and, thus, form the corresponding classes for our recognition/classification tasks.

2.5.1 Visual-Facial Empirical Study on Subjects

To acquire image data, we built a two-camera system. Specifically, two identical cameras of 800-by-600 pixel resolution were placed with their optical axes on the same horizontal plane and forming a 30° angle. This allowed us to video record and photograph faces in front and side view, simultaneously while users were interacting with the medical educational application.

Videos of the resulting facial expressions were then showed to the participant to verify his/her facial expression. If the subject agreed that the expressed emotions were genuine, with regards to the facial expression, they were saved and labeled; as photographs. The final dataset consists of the images of 300 different individuals, each forming the six expressions: “neutral”, “happiness”, “sadness”, “surprise”, “anger”, and “disgust”, as described in [51].

2.5.2 Visual-Facial Empirical Study by Human Observers

In order to understand how humans classify someone else’s facial expression and set a target error rate for automated systems, we developed a questionnaire that was then answered by 300 participants acting as observers. Our aim was to identify differences between the “neutral” expression of a modality and its deformation into other expressions. We also aimed at quantifying these differences into measurements of the face (such as dimension ratio, distance ratio, texture, or orientation) and, thus, extracting corresponding features that convert pixel data into a higher-level representation of shape, motion, color, texture, and spatial configuration of the face and its components. With these goals in mind, the questionnaire was divided into two separate parts:

1. In the first part of the questionnaire, the participants were asked to map the facial expressions that appeared in 14 images into emotions. Each participant could choose from the 6 common emotions that we mentioned earlier, namely “anger”, “happiness”, “neutral”, “surprise”, “sadness”, and “disgust”, or specify any other emotion that he/she thought appropriate. Next, the participant had to specify the degree (0–100 %) of his/her confidence in the identified emotion. Finally, he/she had to indicate which features (such as the eyes, the nose, the mouth, the cheeks, etc.) had helped him/her make that decision. A typical print-screen of the first part of the questionnaire is illustrated in Fig. 2.3.
2. In the second part of the questionnaire, the participant had to classify the emotion from portions of the face. Specifically, we showed the participant the “neutral” facial image and an image of some facial expression of a subject. The

4a. What emotion does the image represent

Other...

4b. In what percent: %

4c. Which Facial Features helped you understand the emotion?
(you can choose more than one)

<input type="checkbox"/> Mouth	<input type="checkbox"/> Forehead texture
<input type="checkbox"/> Eyes	<input type="checkbox"/> Texture between the brows
<input type="checkbox"/> Shape of the face	<input type="checkbox"/> Texture of the cheeks
Other....	

Fig. 2.3 The first part of the questionnaire

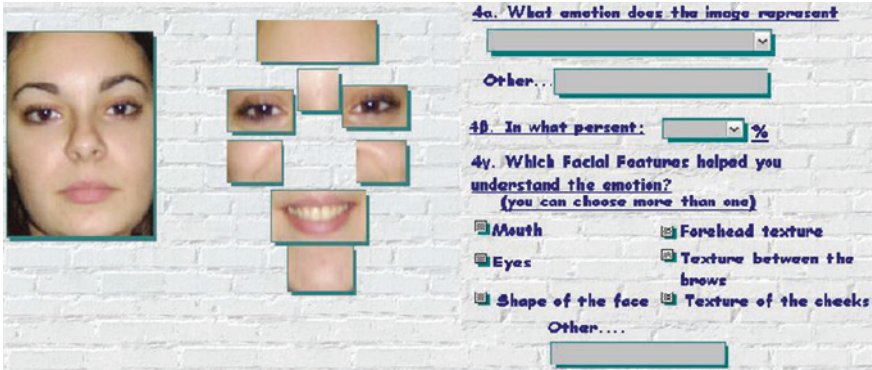


Fig. 2.4 The second part of the questionnaire

latter image was cut into the corresponding facial portions, namely, the eyes, the mouth, the forehead, the cheeks, the chin and the brows. A typical print-screen of this part of the questionnaire is shown in Fig. 2.4. Again, each participant could choose from the 6 emotions mentioned earlier or specify any other emotion that he/she thought appropriate. Next, the participant had to specify the degree (0–100 %) of his/her confidence in the identified emotion. Finally, he/she had to indicate which features (such as the eyes, the nose, the mouth, the cheeks, etc.) had helped him/her make that decision.

When it comes to recognizing an emotion from someone else’s facial expression, the majority of the participants consider it as a difficult task. Generally, “positive” emotions, such as “surprise” or “happiness”, achieved the highest correct recognition rate of 90 and 70 %, respectively. The questionnaire indicates the emotion of “disgust” as the most difficult to recognize, as it was correctly recognized at a rate of only 37 %. Other correct recognition rates by human experts, corresponding to all emotions in our questionnaire, are shown in Fig. 2.5 and Table 2.3.

From our study of the images and the questionnaire, we identified significant deviations from the “neutral” to other emotions, which can be quantified into a classifying feature vector. Typical such deviations are shown in Table 2.3.

The observations of facial changes that arise during formation of various facial expressions, as indicated in Table 2.3, led us to the identification of the most important facial features that can represent these changes in mathematical terms and allow us to form *feature vectors*. The aim of the feature extraction process is to convert pixel data into a higher-level representation of shape, motion, color, texture, and spatial configuration of the face and its components. Specifically, we locate and extract the corner points of specific regions of the face, such as the eyes, the mouth and the eyebrows, and compute variations in size or orientation from the “neutral” expression to another one. Also, we extract specific regions of the face, such as the forehead or the region between the eyebrows, so as to compute variations in texture. Namely, the extracted features are: (1) mouth ratio, (2) left eye ratio, (3) right eye ratio, (4) ratio of the face dimensions, (5) texture

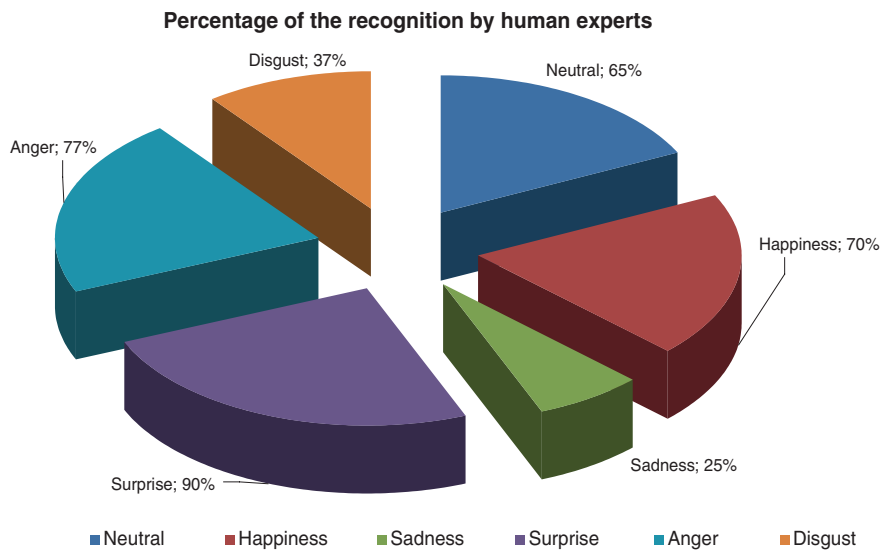








Fig. 2.5 Percentage of the recognition by human experts

Table 2.3 Emotions and facial expression

Results from empirical studies			
Emotion	Facial Image	Deviation from the “neutral expression”	Recognition percentage by human observers (%)
Neutral		<ul style="list-style-type: none">• Lack of facial skin movement• All the variations depart from this expression	65
Happiness		<ul style="list-style-type: none">• Bigger-broader mouth• Slightly narrower eyes• Changes in the texture of the cheeks• Occasionally changes in the orientation of brows	70
Sadness		<ul style="list-style-type: none">• Changes in the direction of the mouth• Wrinkles formed on the chin (different texture)• Occasionally, wrinkles formed in the forehead and different direction of the brows	25

(continued)

Table 2.3 (continued)

Results from empirical studies			
Emotion	Facial Image	Deviation from the “neutral expression”	Recognition percentage by human observers (%)
Surprise		<ul style="list-style-type: none">• Longer head• Bigger-wider eyes• Mouth opened• Wrinkles in the forehead (changes in the texture)• Changes in the orientation of brows (the brows are raised)	90
Anger		<ul style="list-style-type: none">• Wrinkles between the brows(different textures)• Smaller eyes• Wrinkles in the chin• The mouth is tight• Occasional wrinkles over the brows. in the forehead	77
Disgust		<ul style="list-style-type: none">• The distance between the nostril and the eyes is shortened• Wrinkles between the brows and on the nose• Wrinkles formed on the chin and the cheeks	37

measurement of the forehead, (6) texture measurement of the chin, (7) texture measurement of the region between the brows, (8) texture measurement of the left cheek, (9) texture measurement of the right cheek and, (10) brow orientation. The resulting feature vector is fed into a classifier (e.g. artificial neural network, svm-based, etc.) that will attempt to recognize the person’s facial expression. The feature extraction process and the recognition results of a neural network-based system are analyzed and presented at various stages of system development in [53–55]. However, a detailed analysis is beyond the scope of the present chapter that focuses on combining information from two modalities.

2.6 Discussion and Comparison of the Results from the Empirical Studies

On the basis of the results of the empirical studies that were described in the previous sections, we can compare the emotion recognisability achieved by the audio-lingual and the visual-facial modalities from the perspective of a human observer.

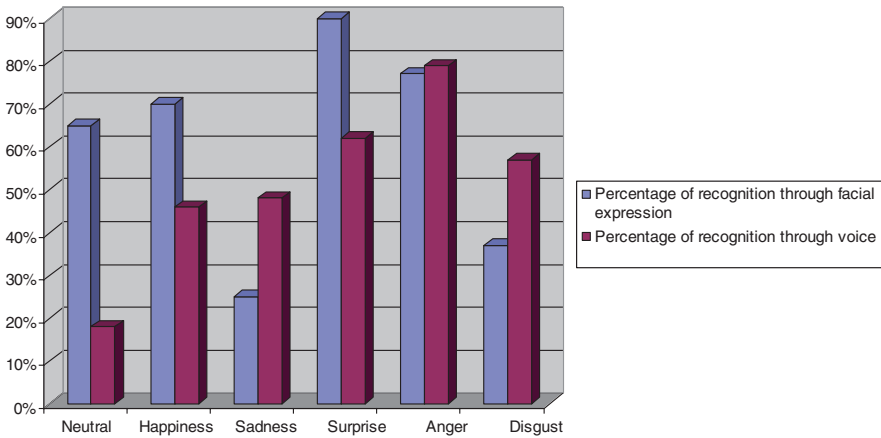


Fig. 2.6 Recognition of emotions through visual-facial and audio-lingual modalities

Figure 2.6 illustrates the percentages of successful emotional recognition through the audio-lingual and visual-facial modalities. The analysis of Fig. 2.6 leads to important conclusions.

In an overall comparison of the two modalities, the visual-facial modality appears to have stronger affect recognition potential than the audio-lingual modality. There are cases where both the audio-lingual and the visual-facial emotion analysis recognize an emotion significantly well. Such cases are for the emotions of anger and surprise where successful recognition is over 50 %. However, surprise can be recognized more easily by the visual-facial modality than the audio-lingual one, as the visual-facial analysis has been successful for 90 % of the cases whereas the audio-lingual modality has been successful for 62 % of the cases. In the case of the emotion of sadness, there is not satisfactory recognition by either modality since in both modalities affect recognition is under 50 %. However, regarding the emotion of sadness, the audio-lingual modality appears to have stronger affect recognition potential, since it achieves a recognition rate of 48 % rather than the rate of 25 % of the visual-facial modality.

When comparing the two modalities, the visual-facial modality is significantly better than the audio-lingual in the recognition of the following three emotional states:

1. Neutral (visual-facial: 65 % versus audio-lingual: 18 %).
2. Happiness (visual-facial: 70 % versus audio-lingual: 46 %).
3. Surprise (visual-facial: 90 % versus audio-lingual: 62 %).

The audio-lingual modality is significantly better than the visual-facial one in the recognition of the following two emotional states:

1. Disgust (audio-lingual: 57 % versus visual-facial: 37 %).
2. Sadness (audio-lingual: 48 % versus visual-facial: 25 %).

The emotional state that is easily recognizable by either modality is:

1. Anger (audio-lingual: 79 % versus visual-facial: 77 %).

The most relevant research work performed by other researchers in the past is that of De Silva et al. [8] who performed an empirical study and reported results on human subjects’ ability to recognize emotions. The aims of the empirical study performed by De Silva et al. were similar to the aims of our empirical study, i.e. discovering the best modality for recognizing certain emotions. For this reason, they compared human recognition results in three tests: video only, audio only, and combined audio and video. In particular, De Silva et al. [8] showed video clips of facial expressions and corresponding synchronised emotional speech clips to 18 human subjects and asked them to recognize the emotions of the humans, who were appeared in the video and speech clips. De Silva et al. concluded that some emotions, such as sadness and fear, are better identified via audio clips. On the other hand, they concluded that other emotions, such as anger and happiness, are better identified via video clips. However, De Dilva et al. focused on the audio signals of voice rather than lingual keywords that conveyed affective information. In fact, in order for them to ensure that the lingual information would not interfere with the results of their empirical study, they used human subjects, who were not familiar with the languages used in the video clips, namely Spanish and Sinhala. In contrast with the De Silva et al. approach [8], in our research we have included the lingual aspect of users’ spoken words on top of the pitch and volume of their voice and compare the audio-lingual results with the visual-facial results to identify the modality that conveys the largest chunk of information for human observers. Table 2.4 compares our findings with the findings in Da Silva et al. [8].

Table 2.5 illustrates percentages of successful emotion recognition through audio and facial data, based on the work of Busso et al. [12]. In their work, the

Table 2.4 Comparison of our results with De Silva and Huang [9]

Emotions	Audio modality		Video modality	
	Our results (%)	De Silva et al. [8] (%)	Our results (%)	De Silva et al. [8] (%)
Happiness	46	43	70	84
Sadness	48	36	25	16
Anger	79	43	77	66
Disgust	57	–	37	–
Surprise	62	44	90	56
Neutrral	18	–	65	–

Table 2.5 Computerized recognition results from Busso et al. [12]

Emotions	Audio (%)	Facial (%)
Anger	68	79
Sadness	64	81
Happiness	70	100
Neutral	81	81

emotional states of humans are recognized through a sophisticated emotion recognition system and not by other humans. Moreover, only four (4) emotional states were recognized, namely anger, sadness, happiness, and neutral. For these reasons, this specific study could not give us significant empirical information for the purposes of our study.

2.7 Combining the Results from the Empirical Studies Through MADM

The empirical data from the two modalities are combined using a multi-criteria decision theory, where each modality is a criterion. The percentages of emotion recognition for each emotion and for each modality are analyzed and then used as weights in order to improve the accuracy of the bi-modal system and determine the prevailing emotion (Fig. 2.7).

According to SAW, emotion recognition values are calculated as a linear combination of the inputs provided by the two distinct modalities and the main aim is to determine the prevailing emotion in order to improve the accuracy of the system. More specifically, using SAW, emotion recognition values are calculated by applying the corresponding criteria weights to the following Eq. (2.2)

$$EM_i = \sum_{m=1}^2 W_{mi} * V_{mi}, \quad i = 1, \dots, 6, \quad (2.2)$$

where EM_i represents the value of successful recognition of the i -th of the six basic emotions through the visual-facial and audio-lingual modality. Moreover, W_{mi} is the weight of the criterion V_{mi} which corresponds to the successful recognition of the i -th emotion through the visual-facial ($m = 1$) or the audio-lingual ($m = 2$) modality. The values of the criteria V_{mi} are calculated in run time by each modality separately and then are incorporated to the resulting system that combines them. Then the system decides for the prevailing emotion (REM in Eq. 2.3) as:

$$REM = \max(EM_i), \quad i = 1, \dots, 6, \quad (2.3)$$

Taking into account the vectors \bar{V}_{1i} and \bar{V}_{2i} , the utility function for the estimation of each emotion recognition is done based on the formulae presented in Table 2.7. Table 2.7 results from Table 2.6, in which the ability of each modality in recognizing an emotional state is illustrated. The values in Table 2.7 emerge from the values in Table 2.8 after normalization to 1 and can be incorporated into the MADM model.

The weights W_{mi} , on the other hand, are static and have resulted from the analysis of the results of the empirical study. In particular, the weights W_{mi} show the system efficiency in recognizing the i -th emotion through the m -th modality. For example, the neutral emotion ($i = 1$) is mostly recognized through the visual-facial modality ($m = 1$) and, therefore, the weight of recognition of the neutral emotion through the visual-facial modality (W_{11}) is higher than the weight of recognition of the neutral

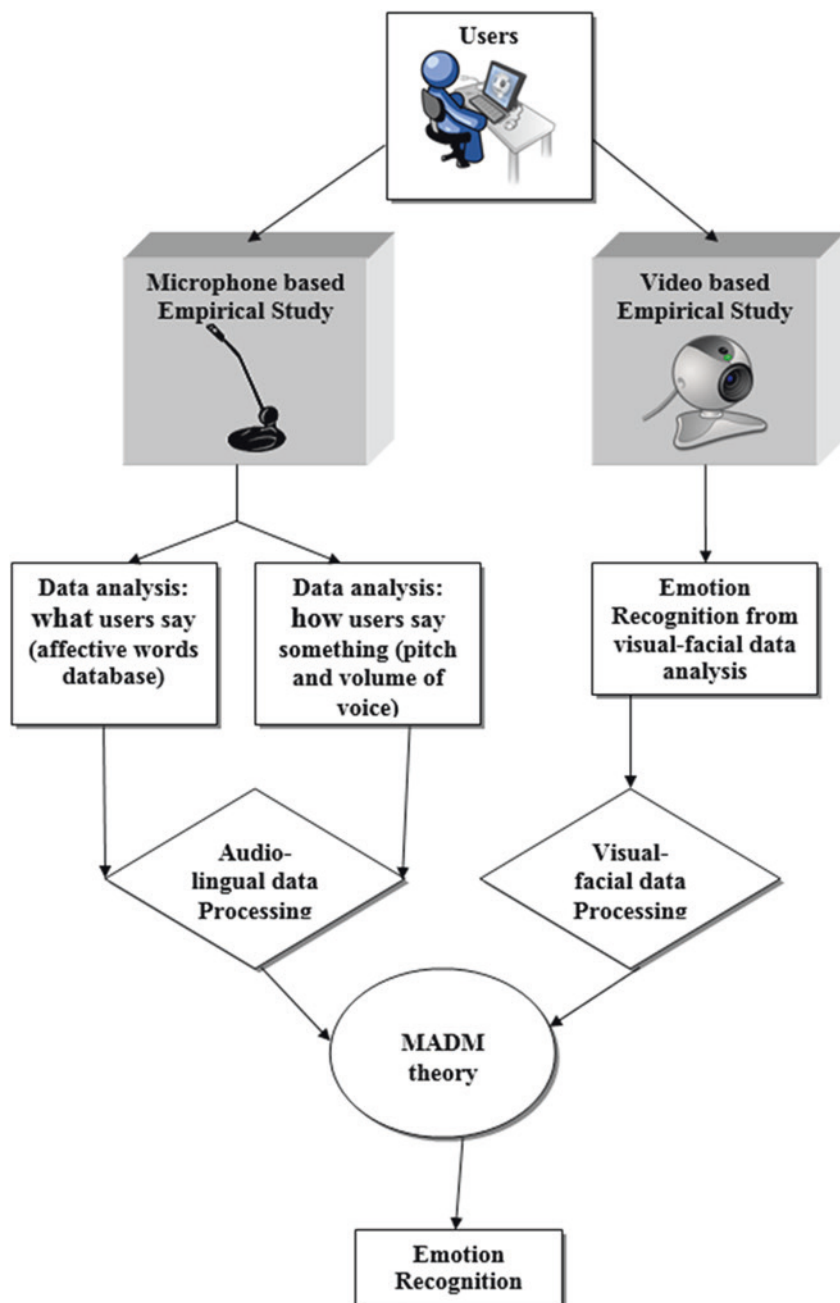


Fig. 2.7 Combining audio-lingual and visual-facial data through MADM

Table 2.6 Recognition of emotions through audio-lingual and visual-facial data

Emotions	Video modality (%)	Audio modality (%)
Neutral	65	18
Happiness	70	46
Sadness	25	48
Surprise	90	62
Anger	77	79
Disgust	37	57

Table 2.7 The formulae for emotion recognition and selection by the system

Emotion	$EM_i = W_{1i} * V_{1i} + W_{2i} * V_{2i}$
Neutral	$EM_1 = 0.78 * V_{11} + 0.22 * V_{21}$
Happiness	$EM_2 = 0.60 * V_{12} + 0.40 * V_{22}$
Sadness	$EM_3 = 0.34 * V_{13} + 0.66 * V_{23}$
Surprise	$EM_4 = 0.59 * V_{14} + 0.41 * V_{24}$
Anger	$EM_5 = 0.49 * V_{15} + 0.51 * V_{25}$
Disgust	$EM_6 = 0.39 * V_{16} + 0.61 * V_{26}$

emotion through the audio-lingual modality (W_{2i}). The utility function for the first emotion (neutral emotion) is: $EM_1 = W_{11} * V_{11} + W_{21} * V_{21}$ and after replacing the corresponding weights from Table 2.7 becomes: $EM_1 = 0.78 * V_{11} + 0.22 * V_{21}$. The values for V_{11} and V_{21} are calculated in run time by the system and represent the results of emotion recognition through the visual-facial and the audio-lingual modality, respectively. More specifically, V_{11} and V_{21} are elements of vectors \bar{V}_{1i} and \bar{V}_{2i} correspondingly and they take their values in $[0,1]$. \bar{V}_{1i} is the vector that refers to modality $m = 1$ (visual-facial modality) and \bar{V}_{2i} is the vector that refers to modality $m = 2$ (audio-lingual modality). In the proposed system, each modality sends data to the emotion recognition decision system through vectors $\bar{V}_{1i} = (V_{11}, V_{12}, V_{13}, V_{14}, V_{15}, V_{16})$ and $\bar{V}_{2i} = (V_{21}, V_{22}, V_{23}, V_{24}, V_{25}, V_{26})$. Specifically, the data for the six emotions are represented as vector elements V_{11} – V_{16} for the visual-facial modality and V_{21} – V_{26} for the audio-lingual modality.

In cases where one modality or both modalities contain significant information about a specific recognized emotion, the multi-criteria approach confirms the recognition. However, there are cases where evidence from modalities leads to two possible recognized emotional states. These are cases where a modality fails to recognize an emotion correctly as an emotional state is confused with another. As an example, we may state that for the visual-facial emotion recognition the emotional state of anger projects as a facial expression that may be confused with the facial expression projected by the emotional state of sadness. Hence, visual-facial emotion recognition systems frequently fail to correctly identify these two emotions. The incorporation of the multi-criteria model, with weights that derive from the analysis of the empirical studies, gives the bi-modal system the additional capability to use evidence from more than one modalities of interaction that is complementary to a high extent. Stated in other words, when a modality cannot distinguish between two or more emotional states, the incorporation of evidence from a second modality might provide a solution.

Table 2.8 A possible response from the audio-lingual emotion recognition subsystem

<i>Audio-lingual emotion recognition system</i>						
Emotions	Neutral (%)	Happiness (%)	Sadness (%)	Surprise (%)	Anger (%)	Disgust (%)
Percentage of recognition	6	6	3	37	44	4

As a characteristic example, we consider the situation where a user is angry about something while s/he uses an educational application. Evidence from the user's voice and face is provided by a video camera that records users' interaction with the computer. Each modality uses the captured video data and analyzes them in terms of only optical and only acoustical information. Subsequently, both modalities make emotion recognition assumptions. In this particular situation, where a user is experiencing the emotional state of anger, we have the following evidence: the user raises his/her voice volume. At the same time linguistic information, such as the presence of specific exclamations, reveals an unusual emotional state of the user. Finally, characteristic changes appear on the user's face.

The audio-lingual information (higher voice volume and presence of specific exclamations) is compatible to a high degree with emotional states of anger and surprise. However other emotional states may produce similar audio-lingual information. After the analysis of the six possible emotional states for the seven basic input action characteristics, described in Sect. 2.4, a possible recognized emotion response from the audio-lingual recognizer is the following (Table 2.8):

This response may accordingly result in a vector $\bar{V}_{2i} = (0.06, 0.06, 0.03, 0.37, 0.44, 0.04)$. The resulting response from the audio-lingual recognizer emerges also from the analysis of stereotypical data, both from user characteristics, as well as from user input actions that indicate possible emotional states. A thorough examination of both the stereotypical analysis for audio-lingual emotion recognition and the underlying mechanism for the combination of stereotypical emotional data is presented in previous works of the authors [52]. Table 2.8 indicates that there are two possible recognized emotional states, namely anger and surprise, with the emotional state of anger being dominant.

Continuing our example, snapshots from the video camera provide the other modality for emotion recognition, i.e. facial evidence of the user. Specifically, the visual-facial modality returns a six dimensional vector containing one ace to indicate a recognized expression and five zeroes. For example, when a 'neutral' facial expression is recognized, an output value of [1.00; 0.00; 0.00; 0.00; 0.00; 0.00] should be returned ideally. Similarly, when the emotional state of happiness is recognized, the ideal output should be [0.00; 1.00; 0.00; 0.00; 0.00; 0.00] and similarly for the other expressions. Practically, the output vector components will have values in the [0, 1] interval each and all of them adding up to one. Thus, the output vector components can be regarded as a set of six degrees of membership of the face image in each of the six emotional states, namely "neutral", "happiness", "surprise", "anger", "disgust", and "sadness" [55]. For example, a typical output vector from the visual-facial modality could be: [0.10; 0.01; 0.44; 0.01; 0.40;

0.04]. This implies that the visual-facial recognizer considers the input expression as “neutral”, “happiness”, “sadness”, “surprise”, “anger”, and “disgust” with a confidence of 10, 1, 44, 1, 40 and 4 %, respectively.

In this particular situation the emotional states of anger and sadness are recognized with similar degrees of confidence. Furthermore, the visual-facial emotion recognition system would wrongly recognize sadness, rather than anger, as the most probable emotional state.

Thus, the two modalities return the output vectors $\bar{V}_{1i} = (0.10, 0.01, 0.44, 0.01, 0.40, 0.04)$ (visual-facial modality) and $\bar{V}_{2i} = (0.06, 0.06, 0.03, 0.37, 0.44, 0.04)$ (audio-lingual modality). Their elements represent the degree of confidence of correct emotion recognition for each of the six emotions. In this case, we can calculate the values EM_i for each one of the six emotions:

$$EM_1 = 0.78 * V_{11} + 0.22 * V_{21} \Rightarrow EM_1 = 0.78 * 0.10 + 0.22 * 0.06 \Rightarrow EM_1 = 0.091,$$

Correspondingly $EM_2 = (0.60 * 0.01 + 0.40 * 0.06) = 0.03$, $EM_3 = (0.34 * 0.44 + 0.66 * 0.03) = 0.169$, $EM_4 = (0.59 * 0.01 + 0.41 * 0.37) = 0.157$, $EM_5 = (0.49 * 0.40 + 0.51 * 0.44) = 0.462$, $EM_6 = (0.39 * 0.04 + 0.61 * 0.04) = 0.04$. Applying formula (2.3) $REM = \max(EM_i) = 0.462 = EM_5$ reveals that the utility function is maximized for the first emotion. Therefore, the combination of information from the audio-lingual and visual-facial modalities recognizes the correct emotional state (anger) for the user interacting with the system. Table 2.9 illustrates the conclusive normalized results of emotion recognition, after the application of MADM, taking into consideration weights that resulted from the analysis of our empirical studies.

In view of the above, every time the user interacts with the system and an emotion is expected to occur, the values of the utility functions for all the six emotions are estimated (EM_1 – EM_6) and the one that maximizes the utility function is selected as the prevailing emotion. At this point we should emphasize the fact that, in some cases, a modality may perform quite successfully in recognizing a certain emotional state based on uni-modal data, while information from another modality may seem unusable. However, the objective purpose of an emotion recognition system is to provide considerable and accurate information for a set of emotional states of users. In our case (six discrete emotional states), the two modalities complement each other to a large extent in their ability to recognize emotions. Clearly, properly combined complementary information from several modalities statistically improves the overall emotion recognition accuracy of a system.

Table 2.9 Resulting emotion recognition from the combined bi-modal data

	Video (%)	Audio (%)	Bi-modal result	Bi-modal normalized (%)
Neutral	10	6	9.12	10.03
Happiness	1	6	3	3.30
Sadness	44	3	16.94	18.64
Surprise	1	37	15.76	17.36
Anger	40	44	42.04	46.27
Disgust	4	4	4	4.40

2.8 Discussion and Conclusions

In this chapter, we have described and discussed a novel approach towards combining two modalities in bi-modal affect recognition using Multi-Attribute Decision Making (MADM). The bi-modal interaction consists of the visual-facial and the audio-lingual modalities. The main focus of this chapter has been on two empirical studies that we conducted concerning the two modalities and linking the results to the integration of the modalities through MADM.

The empirical studies constitute an important milestone for our approach described in this chapter. However, the settings and results of our empirical studies provide an important research investigation and results for other researchers to use or to compare with their own results. The field of affect recognition is not currently well understood and there are not many results of empirical work yet in the literature. Specifically, our empirical studies provide results as to how affect recognition is achieved via the visual-facial and audio-lingual modalities for the perspective of a human observer rather than physiological signals [56]. The two modalities are to a high extent complementary to each other and, thus, can be used in a bi-modal affective computer system designed to perform affect recognition taking into account the strengths and weaknesses of each modality.

From the empirical studies, we found that certain emotion states such as neutral and surprise are more clearly recognized from the visual-facial modality rather than the audio-lingual one. Other emotion states, such as anger and disgust are more clearly recognized from the audio-lingual modality. These results are only partially in accordance with a previous empirical study [8] which, however, did not take into account the lingual aspect in the audio modality. This is not surprising, since certain words convey emotions and this constitutes evidence that can be added to such information as the pitch and volume of the voice. Moreover, we have constructed a basic affective vocabulary for the Greek language that can be used in the audio-lingual modality and a database of facial expressions of emotions.

In our research so far, we have performed significant research work leading to the construction of a bi-modal affect recognizer. As a plan for the near future we will conduct extensive experiments for the evaluation of our affect recognizer.

References

1. Leon, E., Clarke, G., Callaghan, V., Sepulveda, F.: A user-independent real-time emotion recognition system for software agents in domestic environments. *Eng. Appl. Artif. Intell.* **20**, 337–345 (2007)
2. Goleman, D.: *Emotional Intelligence*. Bantam Books, New York (1995)
3. Picard, R.W.: Affective computing: challenges. *Int. J. Hum. Comput. Stud.* **59**, 55–64 (2003)
4. Stathopoulou, I.-O., Tsihrintzis, G.A.: Visual affect recognition. *Front. Artificial Intell. Appl.* **214**, 1–267 (2010)
5. Morrison, D., Wang, R., De Silva, L.C.: Ensemble methods for spoken emotion recognition in call-centres. *Speech Commun.* **49**, 98–112 (2007)
6. Pantic, M., Rothkrantz, L.J.M.: Toward an affect-sensitive multimodal human-computer interaction. *Proc. IEEE* **91**, 1370–1390 (2003)

7. Chen, L.S., Huang, T.S., Miyasato, T., Nakatsu, R.: Multimodal human emotion/expressions recognition. In: Proceedings of the 3rd International Conference on Face & Gesture Recognition: IEEE Computer Society (1998)
8. De Silva, L., Miyasato, T., Nakatsu, R.: Facial emotion recognition using multimodal information. In: IEEE International Conference on Information, Communications and Signal Processing (ICICS'97), pp. 397–401 (1997)
9. Huang, T.S., Chen, L.S., Tao, H.: Bimodal emotion recognition by man and machine. In: ATR Workshop on Virtual Communication Environments Kyoto, Japan (1998)
10. Oviatt, S.: User-modeling and evaluation of multimodal interfaces. In: Proceedings of the IEEE, pp. 1457–1468 (2003)
11. Zeng, Z., Tu, J., Liu, M., Huang, T., Pianfetti, B., Roth, D., Levinson, S.: Audio-visual affect recognition. *IEEE Trans. Multimed.* **9**, 424–428 (2007)
12. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of the 6th International Conference On Multimodal Interfaces, State College, PA, USA, ACM (2004)
13. Liao, W., Zhang, W., Zhu, Z., Ji, Q., Gray, W.D.: Toward a decision-theoretic framework for affect recognition and user assistance. *Int. J. Hum.-Comput. Stud.* **64**, 847–873 (2006)
14. Alepis, E., Stathopoulou, I.-O., Virvou, M., Tsihrintzis, G., Kabassi, K.: Audio-lingual and visual-facial emotion recognition: Towards a bi-modal interaction system. In: Proceedings of International Conference on Tools with Artificial Intelligence, ICTAI, 2, art. no. 5670096, pp. 274–281 (2010)
15. Fishburn, P.C.: Additive utilities with incomplete product set: applications to priorities and assignments. *Oper. Res.* **15**, 537–542 (1967)
16. Nasoz, F., Lisetti, C.L.: MAUI avatars: mirroring the user's sensed emotions via expressive multi-ethnic facial avatars. *J. Vis. Lang. Comput.* **17**, 430–444 (2006)
17. Virvou, M., Kabassi, K.: Adapting the human plausible reasoning theory to a graphical user interface. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **34**(4), 546–562 (2004)
18. Alepis, E., Virvou, M.: Object oriented design for multiple modalities in affective interaction. *Intell. Syst. Reference Libr.* **64**, 87–99 (2014)
19. Kabassi, K., Virvou, M.: A knowledge-based software life-cycle framework for the incorporation of multicriteria analysis in intelligent user interfaces. *IEEE Trans. Knowl. Data Eng.* **18**, 1265–1277 (2006)
20. Naumann, F.: Data fusion and data quality. In: Proceedings of the New Techniques and Technologies for Statistics (1998)
21. Kabassi, K., Virvou, M.: A knowledge-based software life-cycle framework for the incorporation of multicriteria analysis in intelligent user interfaces. *IEEE Trans. Knowl. Data Eng.* **18**(9), 1265–1277 (2006), art. no. 1661516
22. Schütz, W., Schäfer, R.: Bayesian networks for estimating the user's interests in the context of a configuration task. In: UM2001 Workshop on Machine Learning for User Modeling, pp. 23–36 (2001)
23. Bohnenberger, T., Jacobs, O., Jameson, A., Aslan, I.: Decision-theoretic planning meets user requirements: enhancements and studies of an intelligent shopping guide. In: Pervasive Computing, pp. 279–296 (2005)
24. Chin, D.N., Porage, A.: Acquiring user preferences for product customization. In: Proceedings of the 8th International Conference on User Modeling 2001, Springer, Berlin (2001)
25. Kudenko, D., Bauer, M., Dengler, D.: Group decision making through mediated discussions. In: User Modeling, 2003, pp. 147–147
26. Vincke, P.: *Multicriteria Decision-Aid*. Wiley, New York (1992)
27. Triantaphyllou, E., Mann, S.H.: An examination of the effectiveness of multi-dimensional decision-making methods: a decision-making paradox. *Decis. Support Syst.* **5**, 303–312 (1989)
28. Ortony, A., Clore, G., Collins, A.: *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge (1998)
29. Ekman, P., Friesen, W.V.: *Unmasking the Face. A Guide to Recognizing Emotions from Facial Clues*. Prentice-Hall, Englewood Cliffs (1975)

30. Ekman, P., Friesen, W.V.: *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto (1978)
31. Ekman, P.: *Emotion in the Human Face*. Cambridge University Press, New York (1982)
32. Ekman, P., Levenson, R.W., Friesen, W.V.: Autonomic nervous system activity distinguishes between emotions. *Science* **221**(4616), 1208–1210 (1983)
33. Ekman, P., Davidson, R.J.: *The Nature of Emotion. Fundamental Questions*. Oxford University Press Inc., Oxford (1994)
34. Ekman, P.: Basic emotions. In: Dalglish, T., Power, M.J. (eds.) *Handbook of Cognition and Emotion*. Wiley, Sussex (1999)
35. Tomkins, S.S.: Affect as amplification: some modifications in theory. In: Plutchik, R., Kellerman, H. (eds.) *Emotion: Theory, Research and Experience*, vol. 1: *Theories of Emotion*. Academic Press, New York (1980)
36. Tomkins, S.S.: Script theory: differential magnification of affects. In: Howe, J.H.E., Dienstbier, R.A. (eds.) *Nebraska symposium on motivation*, vol. 26. Lincoln University of Nebraska Press, Lincoln (1999)
37. Plutchik, R.: *The Emotions: Facts, Theories, and a New Model*. Random House, New York (1962)
38. Plutchik, R.: A general psychoevolutionary theory of emotion. In: Plutchik, R., Kellerman, H. (eds.) *Emotion: Theory, Research and Experience*, vol. 1: *Theories of Emotion*. Academic Press, New York, pp. 3–33 (1980)
39. Oatley, K., Johnson-Laird, P.N.: Towards a cognitive theory of emotions. *Cogn. Emot.* **1**, 29–50 (1987)
40. McDougall, W.: *An Introduction to Social Psychology*. Luce and Co., Boston (1926)
41. Izard, C.E.: *The Face of Emotion*. Appleton-Century-Crofts, New York (1972)
42. Izard, C.E.: *Human Emotions*. Plenum, New York (1977)
43. Frijda, N.: *The Emotions*. Cambridge University Press, New York (1987)
44. Arnold, M.B.: *Emotion and Personality*. Columbia University Press, New York (1960)
45. Weiner, B.: An attributional theory of achievement motivation and emotion. *Psychol. Rev.* **92**, 548–573 (1985)
46. Martinez A.M.: *The AR face database*, CVC Technical Report (1998)
47. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: *Proceedings of the 3rd International Conference on Face and Gesture Recognition*. IEEE Computer Society (1998)
48. The Yale Database. <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>
49. Kanade, T., Tian, Y., Cohn, J.F.: Comprehensive database for facial expression analysis. In: *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*: IEEE Computer Society (2000)
50. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: *IEEE International Conference Multimedia and Expo (ICME'05)*, Amsterdam, The Netherlands (2005)
51. Stathopoulou, I.-O., Tsihrintzis, G.A.: Facial expression classification: specifying requirements for an automated system. In: *Knowledge-Based Intelligent Information and Engineering Systems*, pp. 1128–1135 (2006)
52. Alepis, E., Virvou, M., Kabassi, K.: Development process of an affective bi-modal Intelligent Tutoring System. *Intell. Decis. Technol.* **1**, 1–10 (2007)
53. Stathopoulou, I.-O., Tsihrintzis, G.: Emotion recognition from body movements and gestures. In: *Smart Innovation, Systems and Technologies*, 11 SIST, pp. 295–303 (2011)
54. Lampropoulos, A.S., Stathopoulou, I.-O., Tsihrintzis, G.A.: Comparative performance evaluation of classifiers for facial expression recognition. *Stud. Comput. Intell.* **226**, 253–263 (2009)
55. Stathopoulou, I.-O., Tsihrintzis, G.A.: NEU-FACES: a neural network-based face image analysis system. In: *Proceedings of the 8th International Conference on Adaptive and Natural Computing Algorithms*, Part II Warsaw. Springer, Poland (2007)
56. Stathopoulou, I.-O., Alepis, E., Tsihrintzis, G., Virvou, M.: On assisting a visual-facial affect recognition system with keyboard-stroke pattern information. *Knowl.-Based Syst.* **23**(4), 350–356 (2010)

Intelligent Interactive Multimedia Systems and Services
in Practice

Tsihrintzis, G.A.; Virvou, M.; Jain, L.C.; Howlett, R.J.;
Watanabe, T. (Eds.)

2015, IX, 122 p. 49 illus., 1 illus. in color., Hardcover

ISBN: 978-3-319-17743-4