TÉCNICO LISBOA

RRHE: Remote Replication of Human Emotions

Joaquim António Véstia Guerra

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisor: Prof. Artur Miguel do Amaral Arsénio

Examination Committee

Chairperson:	Prof. Miguel Nuno Dias Alves Pupo Correia
Supervisor:	Prof. Artur Miguel do Amaral Arsénio
Member of the Committee:	Prof. Francisco António Chaves Saraiva de Melo

October 2015

To my parents, Durval and Cidália.

To my brother, João.

To my girlfriend, Débora.

And to all my friends.

Acknowledgments

A thanks to all the teachers who have supported me throughout this career in Instituto Superior Técnico, they taught me all the knowledge that I have achieved on all the different subjects. I want to thank to professor David Carreira, by the fantastic way of transmitting knowledge which made me look to this course in a different way. A special thanks to professor David Matos by his geniality, thank you for the passion transmitted to us in this course that sometimes becomes very tiring.

A very special thanks to my advisor, Prof. Artur Arsénio, thank you for always keeping me in the right direction and for all the patience shown on the revision of this work.

I also want to thank to YDreams and once again to Prof. Artur Arsénio for inviting me to work with that amazing team. A thanks to David Gonçalves for the support provided in the graphic part.

I would also like to thank to my family for supporting me throughout these years. Thank you for all the moral support and motivation that you gave me in bad and in good moments. I want to thank particularly to my father for always being in accordance with my decisions, thank you for the trust deposited on me, to my mother for always being available to help in everything she could, thank you too for the meals prepared with all the affection, and to my brother for helping me to get rid of stressful moments and for the prompt availability and motivation to help me whenever possible.

All this would not have been possible without my colleagues and friends who were present in good and bad times: Alexandre Quitério, Daniel Magarreiro, David Afonso, Emídio Silva, João Oliveira, Miguel Neves, Pedro Pinto, Rodrigo Bruno and many other. A special thanks to my friend Francisco Silva for helping me mainly in the technical aspects.

All this people had a very important role and helped me getting to where I am today, however, there is a person to who I want to thank specially, my girlfriend Débora. Thank you for your understanding and patience in those moments when I was not able to give you the attention I wanted, thank you for the strength you have given me during the most difficult times and thank you for all the good times we spent together.

Resumo

A deteção de emoções humanas por software é um assunto que vem sendo debatido há muito tempo. Várias propostas foram feitas pela literatura mas ainda persistem falhas que não permitem a exploração deste tipo de soluções a nível comercial. Os utilizadores ainda não confiam neste tipo de sistemas devido à alta percentagem de erros na classificação, optando por interação física ou video-conferência para visualmente (e possivelmente utilizando pistas vocais) transmitir as suas emoções.

Uma possibilidade para a melhoria da exatidão dos sistemas atuais poderá ser o uso de fontes de conteúdo emocional multimodais. Isto requereria a integração de várias técnicas de extração de emoções de diferentes tipos. Além disso, as atuais interfaces emocionais devolvem normalmente resultados grosseiros. Na verdade, os algoritmos emocionais apenas emitem palavras correspondentes à emoção detectada. Acreditamos também que uma interface de utilizador para a detecção de emoções mais inteligente pode aumentar drasticamente o número de casos de uso para esta tecnologia, aumentado muito significativamente a usabilidade destes sistemas.

Esta tese endereça os problemas mencionados anteriormente, propondo uma abordagem multimodal para detecção de emoções. O algoritmo de detecção é executado em servidores remotos (e.g. na cloud) e a informação é depois apresentada ao utilizador através de agentes emocionais, como simples emoticons, avatares mais complexos ou interfaces robóticas que podem replicar remotamente as expressões emocionais.

Uma vez que a replicação emocional é feita remotamente, a aplicação cliente e o atuador não precisam de estar no mesmo espaço físico ou mesma sub-rede que o sistema de detecção. Tudo está ligado a um servidor que pode ser alojado na Internet.

O Sistema multimodal proposto combina duas das modalidades de extração de emoções mais usadas, nomeadamente expressões faciais e propriedades vocais. Com um algoritmo deste tipo conseguimos reduzir significativamente a indução em erro causada pela ironia, i.e expressões faciais que contradizem o tom de voz expressado em simultâneo.

Este trabalho foi avaliado comparativamente a dois cenários base, consistindo na avaliação individual dos algoritmos de detecção de emoções facial e vocal. Os resultados mostram que a implementação de um algoritmo multimodal permite um aumento dos acertos de classificações, que por sua vez torna as classificações por software mais próximas daquelas feitas manualmente por utilizadores.

Palavras-chave: Deteção de Emoções, Espressões Faciais, Reconhecimento de Voz Emocional, Classificadores, Máquinas de Vetores de Suporte, Sistema de Codificação de Atividade Facial.

Abstract

Software-based human emotion detection is an issue that has been debated for a long time. Several solutions have been proposed in the literature, but there are still flaws that impair the effective commercial exploitation of such solutions. Users still do not trust this kind of systems due to the high percentage of classification errors, opting by physical interaction or video-conference communication for visually (and possibly using as well audio clues) transmitting their emotions.

One possibility for improving current systems accuracy could be exploiting multimodal sources of emotional content. This will require the integration of multiple techniques of emotion extraction from different sensing modalities. Furthermore, current emotional interfaces are usually bulky. Indeed, emotional algorithms output words corresponding to the detected emotion. We believe that smart user interfaces for emotional detection systems can drastically augment the number of use-cases for this technology, increasing very significantly such systems usability.

This thesis addresses the aforementioned problems. It proposes a multimodal emotion detection approach. The detection algorithm runs on remote backend servers (e.g. on the cloud), and the information is then presented to the user through emotional agents, such as simple emoticons or more complex avatars, or robotic interfaces that may remotely mimic the emotional expressions.

Since the emotion replication is done remotely, the client application and the actuator do not need to be in the same physical space or in the same sub-network as the detection system. Everything is connected to a server that can be hosted in the Internet.

The proposed multimodal system merges two of the most used modalities for emotion extraction, namely facial expressions and voice properties. With such an algorithm we were able to significantly reduce the error induction from irony, i.e facial expressions that contradict the simultaneously expressed vocal tone.

This work was comparatively evaluated with respect to two baseline scenarios, consisting of individually evaluating each of the facial and voice emotion detection algorithms. The results show that this implementation of a multimodal algorithm allows an increase of classification hits, which in turn makes software-based classifications much closer to user-made manual classifications.

Keywords: Emotion Detection, Face Expressions, Voice Emotional Recognition, Classifier, Support Vector Machines, Facial Action Coding System.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The concept of Human-Computer Interaction (HCI) emerged with the necessity of functionality and usability of systems [55]. The level of functionality is measured with the quantity and efficiency of services in the system [52]. The meaning of usability is the level by which a system can be used efficiently and how it adequates to accomplish some goals for specific users. One of the techniques used to improve accuracy in HCI is precisely the recognition of human emotions, which can be used in a large number of systems. However, according to the Nass and Brave [36] studies on HCI emotions, these kind of stimulus can bring problems since it tends to examine photographs and voices with deliberately performed emotions as opposed to emotions experienced naturally.

Interpersonal communications is dominated by non-verbal expressions [3]. This means that the interaction between human and machines could be richer if machines could perceive and respond to human non-verbal communication, such as emotions. This document places a special focus on emotions recognized from face expressions and speech. However there are more non-verbal gestures very important to detect human emotions, such as posture and hand signals [10].

But which emotions should a system recognize? To address this question, lets introduce briefly the concept of basic emotions, based on Ortony and Turner study [38]. These basic emotions (e.g. fear) are the building blocks for more complex emotions (e.g. jealousy). Plutchik, around 2001, has demonstrated an important property of basic emotions; he argues that these emotions are innate and universal across all cultures [47], which adds universality to this kind of systems (that uses emotional parameters). Defining the set of basic emotions, Ekman and Friesen [14] around 1975 limited the list to the following six: Happiness, Surprise, Fear, Disgust, Anger, Sadness.

These basic emotions are the ones recognized by the majority of recent Emotions Recognition Software's, such as the framework for the facial classifications developed by Ekman and Friesen [13] (shown in Figure 1.1). Hence, this thesis aims to recognize the six previous emotions and the neutral state as a seventh emotion, as depicted in Figure 1.2.

For the past 50 years, social scientists community worked hard on facial expressions analysis. They believed that facial expressions are a portal to one's internal mental state [15], [20] and, when an emotion occurs, a series of biological events follow it producing changes in a person (e.g. facial muscles

Surprise: brows raised, eyelids opened and more of the white of the eye is visible, jaw drops open without tension or stretching of the mouth
Fear: brows raised and drawn together, forehead wrinkles drawn to the center, mouth is open, lips are slightly tense or stretched and drawn back
Disgust: upper lip is raised, lower lip is raised and pushed up to upper lip or it is lowered, nose is wrinkled, cheeks are raised, lines below the lower lid, brows are lowered
Anger: brows lowered and drawn together; vertical lines appear between brows; lower lid is tensed and may or may not be raised; upper lid is tense and may or may not be lowered due to brows' action; eyes have a hard stare and may have a bulging appearance; lips are either pressed firmly together with corners straight or down or open, tensed in a squarish shape; nostrils may be dilated (could occur in sadness too) unambiguous only if registered in all three facial areas
Joy: corners of lips are drawn back and up, mouth may or may not be parted with teeth exposed or not, a wrinkle runs down from the nose to the outer edge beyond lip corners, cheeks are raised, lower eyelid shows wrinkles below it and may be raised but not tense, crow's-feet wrinkles go outward from the outer corners of the eyes.
Sadness: inner corners of eyebrows are drawn up, skin below the eyebrow is triangulated with inner corner up, upper lid inner corner is raised, corners of the lips are drawn or lip is trembling

Figure 1.1: Characteristics of six emotions discernible through facial expressions as extracted from Ekman and Friesen [13].

Figure 1.2: Basic emotions to be recognized under the scope of this thesis.

movements). Other authors presented the idea that facial expressions can be used as a strategic tool to accomplish elicit behaviours or social goals in an interaction [18].

In the field of emotion recognition from speech the analysis of Prosody has been the main focus of research. Features like pitch and energy with their meanings, medians, standard deviations, minimum and maximum values [8] are normally combined with some higher level features, such as speaking rate, phone or word duration. A sad agent, for example, typically displays slower, with little high-frequency energy and lower pitched speech. In the other hand, an agent experiencing anger will speak faster and louder, with strong high-frequency energy and more explicit enunciation [46].

This project was done during an internship at YDreams Robotics[1] which has the goal of provide a contribution to the project "Smart Lamp": a robot that adapts its posture according to the human emotion detected.

1.1 Motivation

Software for recognizing human emotions has been in use for a long time, but in fact, people are still feeling limited by computers. Indeed, humans communicate through means that are not typically perceived by machines, in particular spatial relations. Even though we do not realize it consciously, we interact with each other through very simple spatial cues that everyone understands. For instance, something as simple as walking in the direction of someone indicates a wish to speak with that person.

Because the software is not usually capable of perceiving this implicit communication, it forces users to inform the system through explicit interactions, like the press of a button. This results in people feeling like the computer is in their way, instead of supporting their tasks as it was intended, because they are forced to repeat something they have already communicated. Hence, users tend to ignore the system, using older methods and not effectively adopting the solutions developed for them. As such, there is the need for software that deals with this implicit protocol.

1.2 Objectives

One of this thesis main goals is to evaluate the advantages and disadvantages of a system that uses both facial expressions and speech when compared with systems that have just one of these modalities. We have developed a solution that allows the users to remotely interact with a robot which will express human emotions recognized on the controller.

Since this work has been developed inside a company, YDreams Robotics, it means that the obtained result must be a generic and reusable project. For this purpose, one of this thesis goals is to separate the requirements for "Smart Lamp" project from the requirements of a generic application where this system can be included. With this division the final project is a generic emotions recognition tool before its inclusion on "Smart Lamp" project. This way, the software produced could later be reused for other projects, and even further developed outside the scope of this internship.

[1] http://www.ydreamsrobotics.com/ last accessed on 15/09/2015

1.3 Requirements

We need to separate SmartLamp from general requirements that make this solution able to be integrated with other projects.

Starting with SmartLamp, due to its nature, we need to guarantee that both RRHE-Actuator and RRHE-Client can run on mobile devices. Since SmartLamp will support both Android and iOS, we need to guarantee that our apps are also available on these operating systems. Since SmartLamp has personality - according to the detected emotions - we must also ensure that RRHE-Actuator is able to reproduce the correct emotions with the intensity corresponding to their confidence degrees.

As general requirements for this solution we must consider: performance, by reducing to a minimum the amount of operations to be performed by client and actuator; portability, so that RRHE-Client and RRHE-Actuator can be integrated in as many projects and platforms as possible; responsiveness, so that the emotions are quickly reproduced in the actuator after being captured by the client.

1.4 Document Structure

In this document we present the challenges to be addressed, our approach to solve them, and the experimental evaluation of the final solution. First we discuss the related work in chapter 2, where we explore the recurring problems of recognizing human emotions and how using a multimodal technique might provide a solution to those problems. In chapter 3 we explain how we have structured our solution, the **Remote Replication of Human Emotions (RRHE)**, to fulfil all the requirements of this project. The implementation decisions and details are explained in chapter 4. The solution experimental evaluation is described in chapter 5, which explains how we have tested RRHE. Lastly, chapter 6 concludes this document, presenting a final discussion on our work, discussing its main contributions, and directions for further improvements in future work.

Chapter 2

Related Work

Building a system to remotely recognize emotions during human-machine interactions requires knowledge in three main areas - facial expressions analysis, speech features extraction and ubiquitous computing (ubicomp). Firstly, the main techniques used to detect and classify facial expressions will be explored. Afterwards, we will focus on previous work for recognizing emotions from speech analysis, to extract the main features given by Prosody. The most relevant previous work that addressed the combination of multiple modalities for emotion recognition is reviewed afterwards. Since this is a system that needs to run remotely over the Internet, this thesis will explore some methods used on actual ubiquitous systems. Finally, it will be summarized all the questions and concerns elicited during this section, which will serve as the basis for this work.

2.1 Facial Expressions

According to Mehrabian [34], 55% of the effect conveyed by a human communication message is reflected by facial expressions. It is extremely important to have an effective representation of the human face to successfully recognize facial expression. Nowadays, there are two common methods used to obtain facial features:

- Using geometric features;
- Using appearance features [24].

The locations and shape of facial components that represent the face geometry are represented by geometric features. Valstar et al. [56] demonstrated that geometric feature-based methods have an identical or superior performance than appearance-based approaches in Action Unit recognition. In appearance-based methods, the idea is to apply image filters to specific face portions as well as to the whole face, to extract appearance changes over the time.

There are different methodologies studied in the literature for developing classifiers for emotion recognition [49], [11]. In a static approach, the classifiers evaluate each frame in videos to one of the facial expression category. Bayesian network classifiers and Naïve Bayes classifiers were often used on these

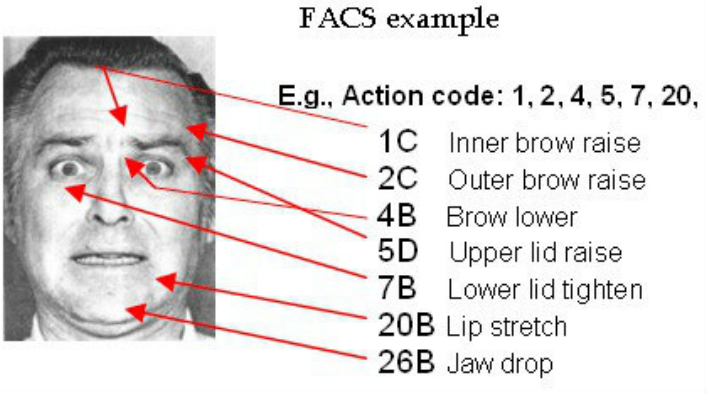

Figure 2.1: Example of Facial Action Coding System (FACS) output

approaches. In dynamic approaches, the classifiers watch for temporal patterns to recognize facial expressions [4]. Classifiers based on Hidden Markov Models (HMM) and multi-level HMM [27] are often used in dynamic approaches.

Facial Action Coding System (FACS) was developed by Ekman and Friesen [15] to represent movements on the face as facial expressions codes. They described a set of action units (AUs). An action unit has a direct link to a muscle movement (e.g. blinking) and they proposed 44 AUs to mask all the possible movement combinations. FACS does not contain any system to classify facial expressions, it needs to be done in an independent system which is preconfigured, manually, with a set of rules. Figure 2.1 shows an example of FACS analysis.

Another approach to recognize emotions from facial expressions was proposed by Mase [33], using optical flow (OF). Black and Yacoob [6] used another classification technique, employing local parametrized samples of image motion to retrieve non rigid motion. Once estimated, these parameters were used as entry to a rule-based classifier to identify the six essential emotions. Yacoob and Davis [61] developed another optical flow technique applying identical rules to execute the classification of the six basic emotions. Rosenblum, Yacoob and Davis [48] also developed optical flow for face fractions and later implemented a function to classify expressions.

Ohya and Otsuka [39] developed yet another optical flow approach. However, they additionally introduced 2D Fourier transform coefficients that were used as feature collections for hidden Markov model (HMM) to classify facial expressions present on each frame. Finally, tracked motions were employed to command the facial expression of an animated Kabuki system [40]. For each one of the six basic expressions it was obtained a detection.

Martinez [32] brought in an indexing approach based on the recognition of frontal face images beneath distinct facial expressions, occlusions and illuminations conditions. A Bayesian approach was implemented to get the right combination between learned features model and local observations. Furthermore, since new conditions could be different from the previous ones, an Hidden Markov Model was applied to increase recognition rates.

Oliver et al. [37] used lower face tracking as a strategy to select mouth features, using the obtained values as information to an Hidden Markov Model based system. Figure 2.2 shows an example of

Figure 2.2: Real time computer graphics animation on Oliver's system

how this mouth tracking works. The mentioned techniques are akin because they initially extract a few features from each frame, which will then be used as input to a classification system. In addition, the outcome of these techniques is one of the emotion categories previously picked.

2.2 Prosody

The computer speech community, has traditionally focused on "what was said" and "who said it", instead of "how it was said". Languages cannot be considered equally. The large variety of languages, and the correspondent number and variability of features in each one, makes it difficult to predict how to connect these features to obtain better results on the recognition rate. [45].

Detecting emotions in speech is a challenging task. At this moment, researchers are still exploring what features are more relevant to the recognition of emotions in speech. There are also some doubts about which are the best algorithms to classify emotions, and which emotions to class together.

The recent studies for emotion recognition in speech have been using diverse classification algorithms like HMM (Hidden Markov Models), GMM (Gaussian Mixture Model), MLB (Maximum-Likelihood Bayes), KR (Kernel Regression), k-Nearest Neighbour and NN (Neural Network) [17].

Prior research works on both speech and psychology features presented evidence supporting the processing of emotional information from a combination of tonal, prosodic, speaking rate, spectral information and stress distribution [23]. Fundamental frequency and intensity, in particular, are two important parameters extracted from Prosody that need to be properly normalized due to significant variations across speakers.

Affective applications are being developed and gradually appearing in the market. However, the development of efective solutions depends strongly on resources like affective stimuli databases, either for recognition of emotions or for synthesis. The information is normally recorded by the affective databases, by means of sounds, psychophysiological values, speech, etc. and actually there is a great amount of effort on increasing and improving its applications [19]. Some other important resources include libraries of machine learning algorithms such as classification via artificial neural networks (ANN); Hidden Markov Models (HMM); genetic algorithms; etc.

By analysing speech patterns the user's emotions are identified by emotional speech. Parameters extracted from voice and Prosody features such as intensity, fundamental frequency and speaking rate are deeper correlated with the emotion expressed in speech. Fundamental frequency (F_0), normally known as pitch (since it represents the perceived fundamental frequency of a sound) is one of the most important attributes for determining emotions in speech [35].

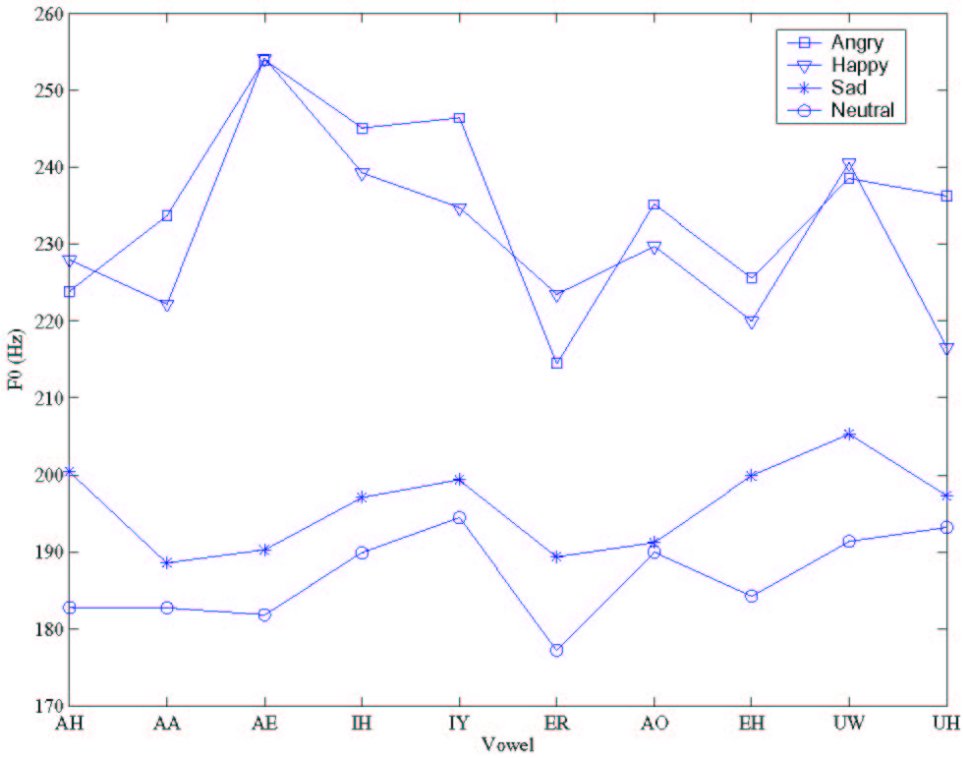

Figure 2.3: Average fundamental frequencies for vowels [63]

One possible way to extract and analyse features from human speech is statistical analysis. Using this method, the features connected with the pitch, Formants of speech and Mel Frequency Cepstral Coefficients, can be chosen as inputs to the classification algorithms. Bäzinger et al. said that statistics related to pitch carry important information about emotional status [9]. Nevertheless, pitch was also considered to be the most gender dependent feature [1].

According to Kostoulas et al. [22], the emotional state of an individual is much related to energy and pitch. From these features of the speech signal, it may be easier to understand happiness or anger, but not so easy to detect, for instance, sadness. In Figure 2.3 we can observe that anger/happy and sad/neutral show similar F_0 values on average. We can note that for neutral speech the mean vowel F_0 values are less when compared with other kinds of emotions.

Besides pitch, there are some other important features that are linked to speaking: rate, formants, energy and spectral features, such as MFCCs. The spectrum peaks of the sound spectrum |P(f)|' of the voice can be defined as formants; this term is a polysemic word and it also refers to an acoustic resonance of the human vocal tract. It is usually calculated as an amplitude peak in the frequency spectrum of the sound. It is useful to distinguish between genders and to predict ages.

Wang & Guan [59] used MFCCs, formant frequency and prosodic features to represent the characteristics of the emotional speech.

MFCCs are an universal way to make a spectral representation of speech. They are used in many areas, such as speech and speaker recognition. Kim et al. [42] referred that statistical assumptions with MFCCs also brings emotional information. MFCCs are generated with a Fast Fourier Transform followed by a non-linear warp of the frequency axis. Afterwards it is calculated the power spectrum, to

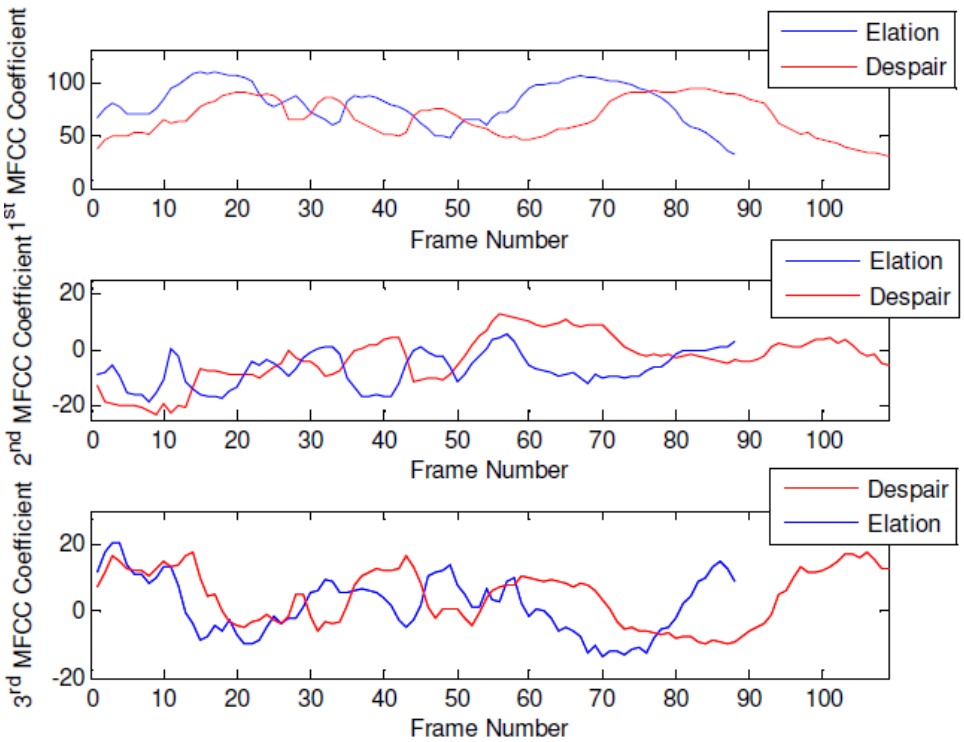

Figure 2.4: Variation in MFCCs for 2 emotional states [51]

have frequencies logarithmically spaced. In the end, MFCCs result from the appreciation with the cosine basis functions of the first N coefficients of this strained power spectrum.

Figure 2.4 shows up the variation in three MFCCs, calculated with the first 13 components, for a female speaker saying the sentence ”Seventy one” (with desperation and euphoria emotional states).

2.2.1 EmoVoice

Vogt et al. [58] developed the EmoVoice framework for online recognition of emotions from human voice. Their project is divided into two major modules: creation and analysis of an emotional speech corpus; and real-time tracking of emotional states.

The initial module includes several tools for audio segmentation, classification of an emotional speech corpus, feature extraction and feature selection. EmoVoice offers a graphical user interface to easily record speech files and create a classifier properly trained according to the speech files. Hence, the returned classifier will be adapted to the context where it will be used for. This classifier can later be used for the second module, the real-time emotion recognition. Here, the results of classification are collected constantly during speaking.

During the first phase, audio segmentation, they decided to use Voice Activity Detection (VAD) to segment the entire sound in fragments of voice activity without pauses longer than 200ms. This approach is really fast and the results come close to segmentation into phrases, without any linguistic knowledge.

On the second step, feature extraction, the goal is to find the set of properties of the acoustic signal

that best characterise emotions. Since an optimal feature set is not yet defined, there was the necessity to choose just a part of the entire set. The properties observed and the views over these values are as follows (concepts extracted from Vogt et al. [58]):

- Logarithmised Pitch: *"the series of local minima and local maxima, the difference between that, the distance between local extremes, the slope, the first and second derivation and, obviously, the basic series."*
- Signal Energy: *"the basic series, the series of the local maxima and local minima, the difference between them, the distance between local extremes, the slope, the first and second derivation as well as the series of their local extremes."*
- MFCCs: *"The basic, local maxima and minima for basic, first and second derivation for each of 12 coefficients alone."*
- Frequency Spectrum: *"the series of the center of gravity, the distance between the 10 and 90% frequency, the slope between the strongest and the weakest frequency, the linear regression."*
- Harmonics-to-Noise Ratio (HNR): *"only the basic series."*
- Duration Features: *"segment length (in seconds); pause as the proportion of unvoiced frames in a segment (obtained from pitch calculation); pause as the number of voiceless frames in a segment (obtained from voice activity detection); zero-crossings rate."*

Finally, the third step runs a classification over the aforementioned extracted features. Actually there are two classification methods supported by EmoVoice: Naïve Bayes (NB), which is very fast even for high-dimensional feature vectors; Support Vector Machine (SVM) which returns more accurate values but shows poor performance.

Figure 2.5 illustrates the overall EmoVoice architecture that was just reviewed. EmoVoice SDK (Software Development Kit) will be adapted for this thesis work.

2.3 Multimodal Implementations

Emotion recognition from multimodal techniques is still an open challenge. Pantic and Rothkrantz [43] presented a survey where the focus was on audiovisual affect recognition. Since then, an increasing number of studies were made on this matter. As evidenced by the state-of-the-art for Prosody and facial expressions implementations (single-modal techniques), most of the existing studies focus on the recognition of the six basic emotions.

Pal et al. [41] presented a system for detecting hunger, pain, sadness, anger and fear, extracted from child facial expressions and screams. Petridis and Pantic [44] investigated the separation of speech from laughter episodes taking into account facial expressions and Prosody features.

Actually, there are three main strategies for data fusion used on audiovisual affect recognition studies:

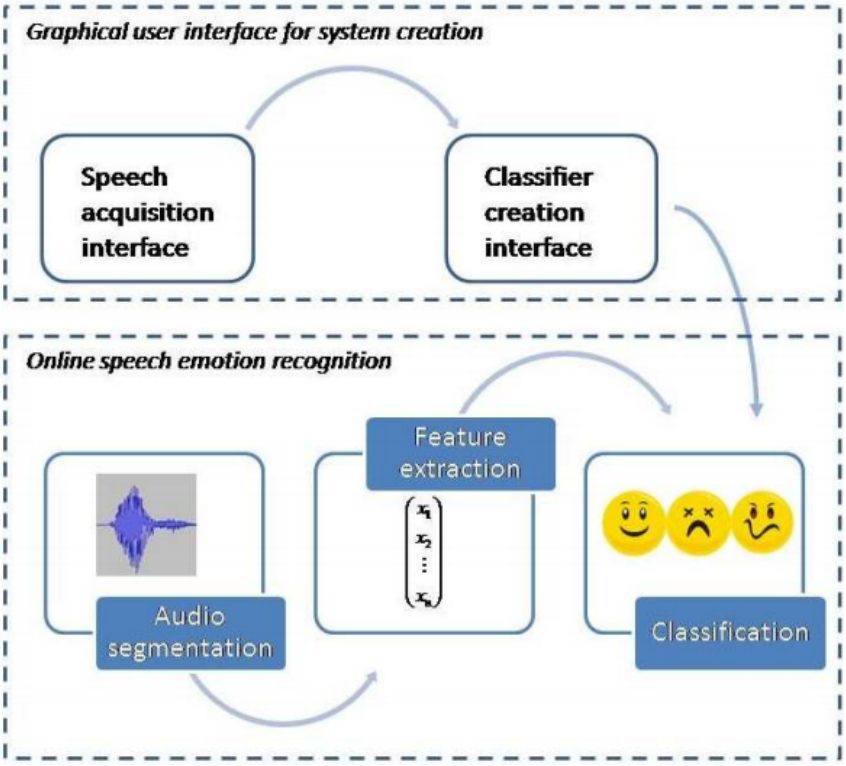

Figure 2.5: EmoVoice architecture overview.

- Feature-Level Fusion (partially extracted from [66]): *"Prosodic features and facial features are concatenated to construct joint feature vectors, which are then used to build an affect recognizer. However, the different time scales and metric levels of features coming from different modalities, as well as increasing feature-vector dimensions influence the performance."*

- Decision-Level Fusion (partially extracted from [66]): *"The input coming from each modality is modelled independently, and these single-modal recognition results are combined in the end. Audio and visual expressions are displayed in a complementary redundant manner what turns incorrect the assumption of conditional independence between audio and visual data streams in decision-level fusion, which results in loss of information of mutual correlation between the two modalities."*

- Model-Level Fusion (partially extracted from [66]): *"To address the problem above, a number of model-level fusion methods have been proposed. It aims at making use of the correlation between audio and visual data streams and relaxing the requirement of synchronization of these streams."*

Zeng et al. [64] introduced a method to fuse multi-streams using HMMs. The goal is to form, according to the maximum common information, an ideal link between several streams extracted from audio and visual channels. Afterwards, Zeng et al. [65] further evolved this technique, presenting a middle-level training approach. Under this layer, several learning schemes can be used to combine multiple component HMMs. Song et al. [53] introduced a solution where upper face, lower face and prosodic behaviours are modelled into individual HMMs to model the correlation features of these elements. Fragopanagos and Taylor [17] proposed an artificial neural networks (NN) based approach. This proposal also incorporates a feedback loop, named ANNA, to assimilate the data extracted from facial expressions, lexical

content and prosody analysis. Sebe et al. [50] utilized a Bayesian network (BN) to combine features from facial expressions and prosody analysis.

Figure 2.6 shows up an overview of the currently existing systems for audiovisual emotions recognition using prosody and visual features as decision factors.

References	Feature	Fusion	Classifier	Performance							
				exp	per	cue	class	sub	samp	acc (%)	other
Busso et al. 04 [16]	102 markers, prosody	F, D	SVM	P	D		4	1	256 sentences	89	
Caridakis et al. 06 [17]	facial points, prosody	M	RNN	S	I		4	SAL 4	1000 tunes	79	
Fragopanagos and Taylor 05 [53]	17 FAPs, prosody	M	ANNA	S	I	L	4	SAL 4	500 epochs	44-71	various labels/ features
Go et al. 03 [56]	Eigenfaces, MFCC	D	LDA	P	I		6	20	360 utterances	95-98	
Hoch et al. 05 [66]	Gabor feature, prosody	D	SVM	P	D		3	7	840 sequences	90.7	car setting
Karpouzis et al. 07 [74]	19 FPs, prosody	M	RNN	S	I	B	4	SAL 4	1000 tunes	82	
Pal et al. 06 [97],	Vertical gray level, F0-F3	D	Rules, k-means	S	D		5	1	?	75.2	
Petridis & Pantic 08 [108]	facial points, cepstral features	F, D	Adaboost + NN	S	I	B	2	8	96 laughter/ speech episodes	86.9	AMI data[6]
Schuller et al. 07 [118]	AAM, prosody, articulatory, voice quality, lexical	F	SVM	S	I	B	3	21	10.5 hours	recall: 41.7-63.9 (RR)	balance training
Sebe et al. 06 [122]	12 motion units, prosody	M	BN	P	D		11	SD 38	1254 sentences	90	
Song et al. 04 [124]	54 FAPs, prosody	M	THMM	P	?		7	?	?	84.7	
Wang & Guan 05 [138]	Gabor wavelets, prosody, MFCC, formants.	D	FLDA	P	I		6	8	500 sentences	82.14	6 lan-guages
Zeng et al. 06 [150]	12 motion units, prosody	M	MFHMM	P	I		11	CH 20	660 sentences	83	
Zeng et al. 07 [151]	Texture with LLP, prosody	D	Adaboost + MHMM	S	D		2	AAI 2	137 uttarances	89	
Zeng et al. 04 [153]	motion units, prosody	D	SNoW	P	D		11	CH 38	1254 sentences	89-90	
Zeng et al. 05 [154]	motion units, prosody	M	MFHMM	P	I		11	CH 20	660 sentences	80.61	
Zeng et al. 07 [155]	motion units, prosody, formants	D	HMM	P	D, I		11	CH 20	660 sentences	I: 72.42 D: 96.3	
Zeng et al. 05 [156]	motion units, prosody	F	Fisher-Boosting	P	D		4	CH 20	660 sentences	84-87	

Figure 2.6: Fusion: Feature/Decision/Model-level, exp: Spontaneous/Posed expression, per: person-Dependent/Independent, class: the number of classes, sub: the number of subjects, samp: sample size (the number of utterances), cue: other cues (Lexical/Body), acc: accuracy, RR: mean with weighted recall values, FAP: facial animation parameter, and ?: missing entry. AAI, CH, SAL, and SD are existent databases. This table was extracted from [66]

Multimodal techniques have already been implemented on several HCI systems. Some examples are (examples extracted from [66]):

- Lisetti and Nasoz [26] proposed a system that would identify user's emotions by fusing physiological signals and facial expressions. The goal is mirror the user's emotion, like fear and anger, by adjusting an animated interface agent.

- Duric et al. [12] proposed a system that implement a model of embodied cognition which is a particularized mapping between the kinds of interface adaptations and the user's affective states.

- Maat and Pantic [31] presented a proactive HCI tool that is able to learn and analyse the context-

dependent behavioural patterns of the users. According to the data captured from multisensory, this tool is able to adapt the interaction accordingly.

- Kapoor et al. [21] proposed an automated learning companion that fuse data from cameras, wireless skin detector, sensing chair and task state to detect frustration and anticipate when the user needs help.

- In the Beckman Institute, University of Illinois, Urbana-Champaign[1] (UIUC), has been developed a multimodal computer-aided learning system, in this system the computer avatar gives a convenient tutoring strategy. This information is based on user's facial expression, keywords, task state and eye movement.

2.4 Irony Detection

Human expressions are often employed to express irony. For instance, bad news (like "you are fired") may turn someone's face with a sad emotional expression, while the person, with a happy voice, states a positive sentiment such as "but these are good news". Indeed, irony is an important instrument in human communication, both verbally as well as written. Indeed, irony is quite often used in literature, website, blogs, theatrical performances, etc.

To the best of our knowledge, no previous work addressed irony detection from multimodal sensing modalities. However, there are some works addressing irony detection in written texts. But even such work is very recent, as demonstrated by Filatova [16] first corpus including annotated ironies in texts.

Buschmeier et al. [7] analyzed the impact of several features, as well as combinations of them, used for irony detection in written product reviews. They evaluated different classifiers, reaching an F1-measure of up to 74% using logistic regression.

Another work [54] used a sentiment phrase dictionary combined method to address multiple semantic recognition problems, such as text irony. Machine learning methods were also employed for satire detection in Web Documents [2].

2.5 Ubiquitous Computing

Ubiquitous computing, now also known as pervasive computing, was described by Mark Weiser as: *"The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it"* [60].

Pervasive computing is the result of the link of upgraded technologies from both distributed systems and mobile computing. The area of distributed systems emerged when personal computers and local networks converged. This field brings lots of knowledge that covers many fundamental areas to pervasive computing:

- Fault tolerance;

[1] http://beckman.illinois.edu/ , last accessed on 27/08/2015

- Remote communication;
- Remote information access;
- Security;
- High availability.

In the early nineties, the emergence of full-function laptops and wireless LANs brought about the problem of building distributed systems with mobile clients. Mobile computing addresses the following areas:

- Mobile information access;
- Mobile networking;
- System-level energy saving techniques;
- Location sensitivity;
- Support for adaptive applications.

The study plan of ubiquitous computing subsumes that of mobile computing, including four extra concepts:

- Masking uneven conditioning: The integration of pervasive computing devices into available smart spaces depends on a number of factors that are not related to technology, such as internal policies and business models.
- Effective use of smart spaces: A space can be a limited area, for instance a room or kitchen, or it can be an open area with well defined barriers like a school campus. By embedding computing devices in space infrastructure, a smart space brings the concept of an intelligent space which allows to develop unexpected connections.
- Localized scalability: The increasing number of connections between devices (user's devices and infrastructure devices) increases smart spaces complexity. This brings critical problems in power consumption, required bandwidth and, of course, it provides more distractions to the mobile user.
- Invisibility: Weiser's ideal is to eliminate all of the pervasive computing technology from the user consciousness.

In the following chapter we explain our approach to provide an answer to these questions.

Chapter 3

Architecture

RRHE's architecture is based on a Client-Server-Actuator model as shown in Figure 3.1. The internal architecture of RRHE-Client module is composed by 3 major logical components: communication module, video segmentation, and live evaluator. The communication module is responsible for the communication with the server, sending the captured data – images and audio files. The video segmentation component incorporates the algorithm which splits a video into several segments containing still images and an audio file. This data is subsequently analyzed by the server. The live evaluator module in the RRHE-Client allows the overall system to run in real-time, using the microphone and video camera available at the hosting device.

The RRHE-Server module is the brain of the system. Similar to the client module, the server also has a communication module. The communication module is listening for requests, either from the client – with data for classification – or from the actuator – with requests for users' status updates. The components responsible for the capture of emotions are a layer above the communication module. For the facial expressions we have a component capable of detecting and extracting faces in images. After running this process, the Face Emotion Recognizer component evaluates the extracted face and outputs an emotional category. In the vocal expression side we have the Voice Emotion Recognizer that incorporates algorithms capable of extracting properties from the audio signal for further analysis in the classification phase. This thesis also proposes another module performing the multimodal integration of facial and audio emotional content for better classifying emotions and for enabling the detection of ironies.

Finally, the RRHE-Actuator module replicates the emotions detected by the server and remotely transmitted through the communication channel. The actuator starts a cycle of requests to the server where it requests the most recent emotional state of a user. RRHE-Actuator can replicate the detected emotions in several ways, from the display of an emoticon to the status update in a social network or mimic interpretation by a robotic agent. In the context of this thesis, only the representation by emoticons and the integration with Facebook social network were considered.

The next sections present in detail each module of the RRHE architecture.

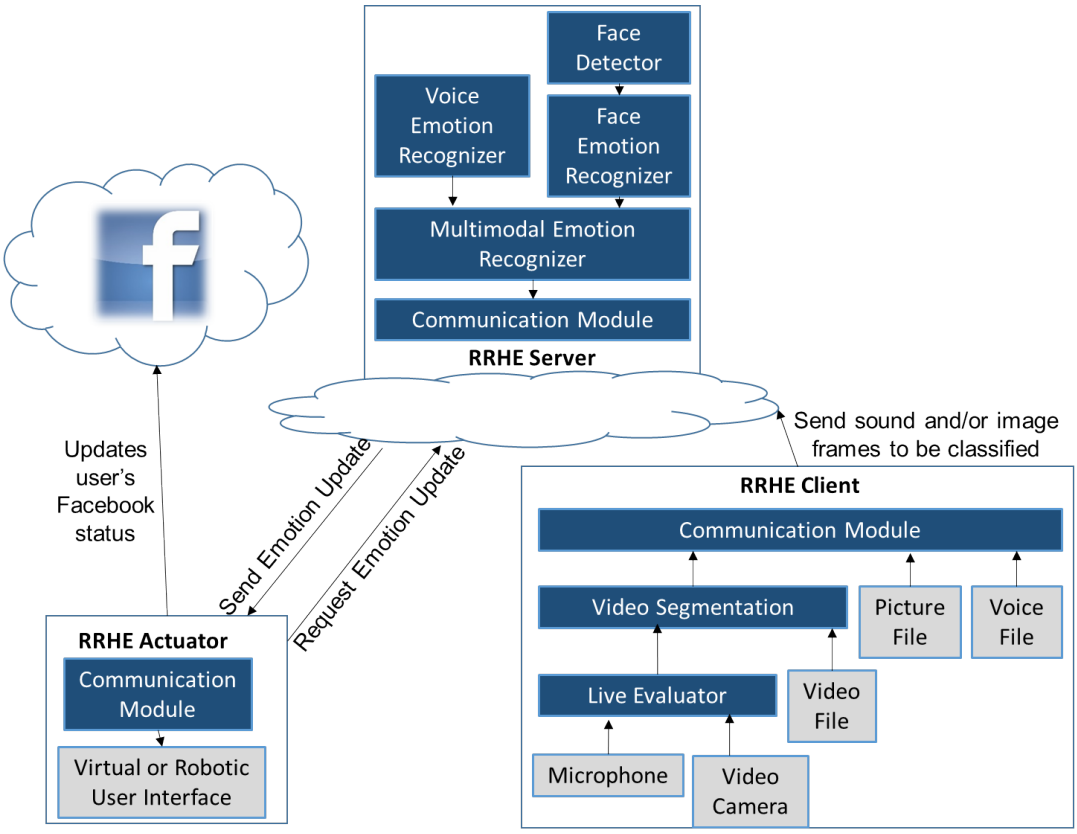

Figure 3.1: RRHE Architecture.

3.1 Client

The client module (RRHE-Client) is an application that can be installed in any PC or mobile device. The purpose of the RRHE-Client is the collection of sound and image data (from the microphone and video camera devices, respectively), and their transmission to the server for proper classification. However, some processing needs to be carried out at the client. Hence, the RRHE-Client has three fundamental modules: Core Functionality, Video Segmentation, and Live Evaluation.

3.1.1 Core Functionality

A main interface,as shown in Figure 3.2, enables users to access the core functionality. The latter consists of a set of commands used for testing the system. The available commands and respective functions are as follows:

1. Send image - allows to select an image file to be sent to the server for evaluation and classification as an emotion detected in the face transmitted - if there is one. This is a very important function since it allows to separately test the recognition of facial emotions.

2. Send sound - allows to select a sound file to be sent to the server, for evaluation and classification as an emotion detected on the voice transmitted - if there is one. It is equally important function since it allows to separately test the recognition of voice emotions.

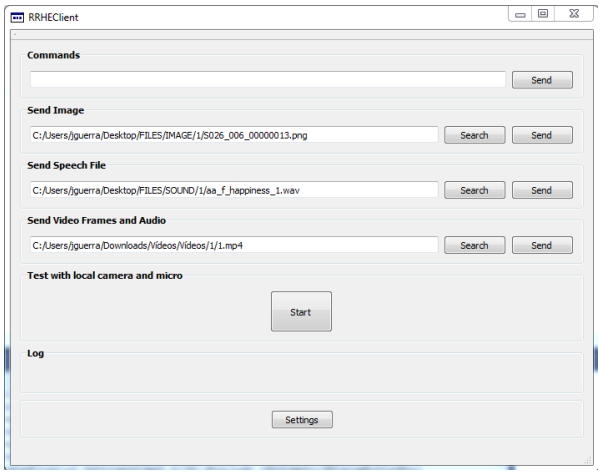

Figure 3.2: RRHE-Client main interface.

Figure 3.3: RRHE-Client Log.

3. Send video - allows to select a video file for testing the system as a whole, combining facial and voice emotions. Image and sound segment pairs are extracted from the provided video file as previously detailed in section 3.1.2.

4. Begin Live Evaluation - enters the Live Evaluation module (section 3.1.3).

5. Log - shows a list of error messages (a log example is presented in Figure 3.3) that help the user understand the system's behavior and fix any problem.

6. Settings - enters the Settings module to configure the system as later described in section 3.1.4.

3.1.2 Video Segmentation

It is necessary to fragment a video, so that one can recognize emotional expressions on its content. Each segment, with a configurable duration, consists of:

1. A sound file: audio signal and the duration of the segment;

2. An image: that can be captured in the beginning, middle, or end of the segment.

Knowing the total size of the video, the duration of the segments and the configurable interval between them we can immediately split the video in several segments – using the open-source software FFmpeg. As soon as we have several segments for classification, and once again with FFmpeg, we can extract the sound from this segment. Finally, to extract a frame from each of the segments, we used functional object VideoCapture from OpenCV. With this object we have access to the total number of frames in a segment and, with the configured position, we just need to capture the frame in the right place. Figure 3.4 illustrates this process.

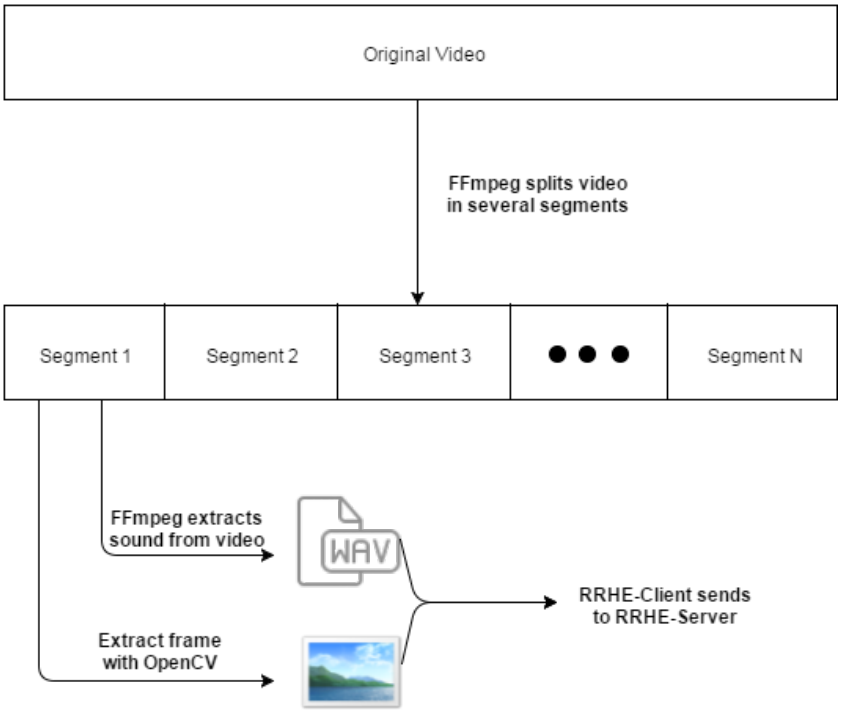

Figure 3.4: Video Segmentation illustration

3.1.3 Live Evaluation

The Live Evaluation module continuously collects still images and sound segments, which are captured from the selected camera and microphone, and sends the data to the server for evaluation. The last captured image is always shown in the Live Evaluation interface, as shown in Figure 3.5. The RRHE solution runs in real-time whenever the Live Evaluation module is used. In addition, whenever facial recognition is on, images and respective sound segments are discarded if a face is not detected.

3.1.4 Settings

The Settings module (see Figure 3.6) contains the following list of settings:

1. User ID - Since RRHE supports multiple users, the user must specify its user identifier.

2. Remote Server Address - The address - IP address or name and IP port - of the RRHE-Server to connect to in the standard format (e.g.: 192.168.0.1:50000, server:50000). If no port is specified, the default port - 50000 - is assumed.

3. Image Frame Position - The frame of a video segment to be used as the image. The options for this setting are:

   (a) Beginning - The first frame of the video segment

   (b) Middle - The middle frame of the video segment

   (c) End - The last frame of the video segment

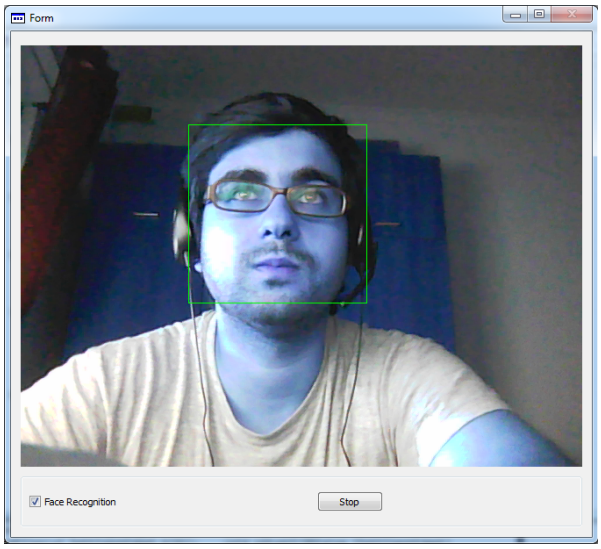

Figure 3.5: Live Evaluation interface.

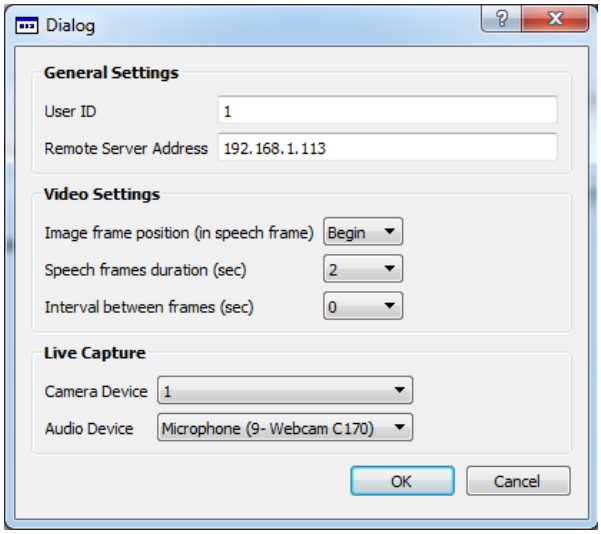

Figure 3.6: Settings interface.

4. Speech Frame Duration - The duration of a video segment. The available options are 1, 2 and 3 seconds

5. Interval Between Frames - The duration of the interval between analysed segments, in other words, the periods of time that are not evaluated. The available options are 0, 1 and 2 seconds

6. Camera Device - The camera used by RRHE-Client to capture video. The available options are all the available cameras.

7. Audio Device - The microphone used by RRHE-Client to capture sound. The available options are all the available microphones.

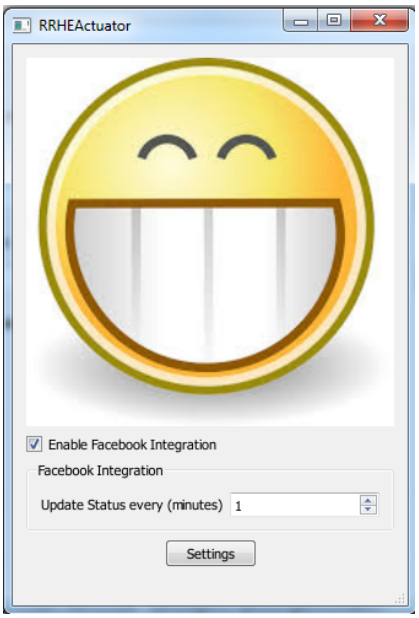

Figure 3.7: RRHE-Actuator user interface showing an emotion

Figure 3.8: RRHE-Actuator user interface showing an emotion

3.2 Actuator

The actuator module (RRHE-Actuator) is an application that can be installed on any PC or mobile device. The purpose of RRHE-Actuator is to represent emotions detected in the data sent by RRHE-Client. RRHE-Actuator establishes a request-on-demand connection with RRHE-Server and periodically asks for the update of the emotional status of the user. After receiving the notification from RRHE-Server (section 3.3), RRHE-Actuator updates the emotional state. The representation of the emotional status is made through two different aspects:

1. Emoticon (Figure 3.7) - An emoticon representing the detected emotion is displayed in the user interface.

2. Facebook integration (Figure 3.8) - To demonstrate a possible way for RRHE to be integrated with external systems, RRHE-Actuator can also update the Facebook status of the user using a 'Feeling' emoticon according to the detected emotion. The Facebook integration can be enabled or disabled in the RRHE-Actuator user interface and the login is requested in the first status update.

3. SmartLamp (Figure 3.9) - SmartLamp is a desktop lamp with robotic behaviors and personality.

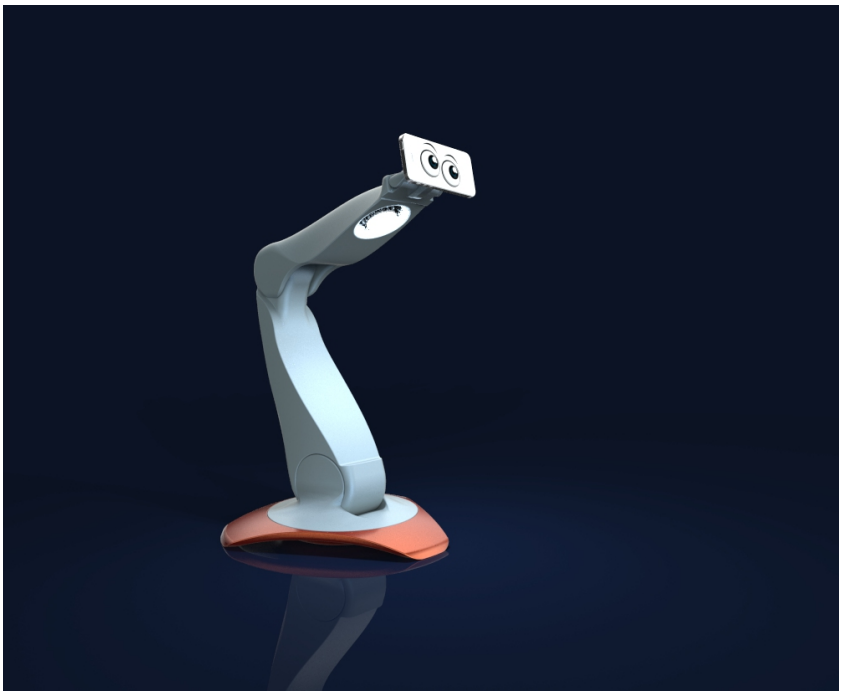

Figure 3.9: SmartLamp Prototype.

This product is being developed by YDreams Robotics and has as main features: face tracking during video calls; video surveillance with motion detection; play games; express emotions. SmartLamp incorporates a smartphone/tablet and it is compatible with Android and iOS. RRHE was developed aiming its integration into the SmartLamp, by including the RRHE-Client and RRHE-Actuator as part of SmartLamp's applications. We will have multiple SmartLamps communicating with one RRHE-Server, sending data to be classified or requesting emotions updates. The representation of each emotion can be modelled and/or mimicked in robotic movements as well as using its screen. Unfortunately, the integration of RRHE in SmartLamp was not possible because the SmartLamp prototype is not yet ready for such integration.

3.3 Server

The server module (RRHE-Server) is a console application (as shown in Figure 3.10) dedicated to the treatment of the information captured and sent by RRHE-Client. RRHE-Server is the core module of RRHE since it is responsible for processing audio and image data to recognize the respective emotion. RRHE-Server consists of:

1. TCP server - a typical TCP server listening to requests.
2. Server Manager - maintains the execution context of the server (e.g., the emotional state of each active user).
3. Worker Threads - launched (one per core) at the start of the application. They work together with the Server Manager in a Single Producer-Multiple Consumer type of environment, processing the

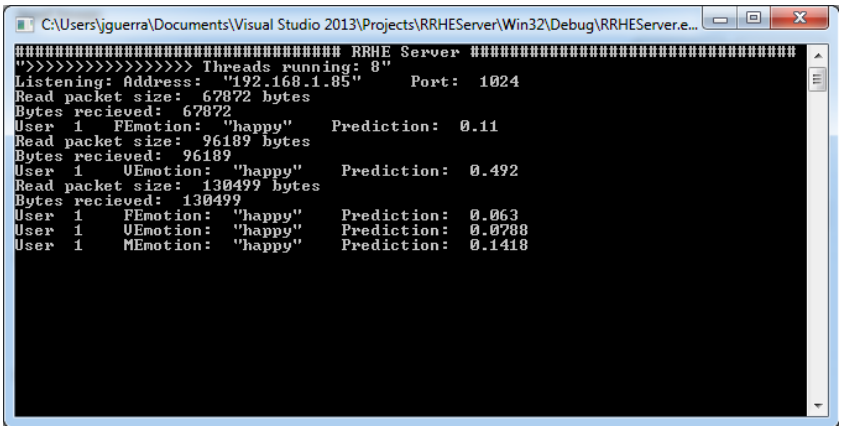

Figure 3.10: RRHE-Server console application showing initialization

RRHE-Client and RRHE-Actuator requests.

3.3.1 Facial Expressions Recognizer

RRHE-Server receives an image file for which it must extract an emotion. This module has been developed purely in C++ with Qt and OpenCV frameworks. The first problem to address is to crop only the significant part - human face - of the entire image. This is achieved using the Haar Cascade methods given by OpenCV. This framework is also useful for rotating the recognized faces to proper positions. Afterwards, we have developed our Gabor Bank implementation, which will filter the image, returning a features vector. We can use our facial classifier, which uses the OpenCV SVM as the learning algorithm, to process such input vector. In the end, this module returns all the confidence values found for each supported emotion.

3.3.2 Vocal Expressions Recognizer

RRHE-Server receives an audio file - WAV format - to extract an emotion from the voice in the audio signal. We first need to extract some audio properties used during the classification process. Since it is easy for Matlab to handle sound files, we decided to use this language for such work. In addition, we also decided to implement our vocal classifier using the Matlab SVM as the underlying learning algorithm. This module will return all the confidence values found for each of the supported emotions.

3.3.3 Emotions Fusion

Once both facial and vocal emotions have been estimated, the fusion algorithm (later described in Section 4.3 of the next chapter) is applied to estimate a final multimodal emotion. This module has been developed purely in C++ with Qt-Framework. Figure 3.10 shows a log of its output. The last three lines in the figure inform: the ID of the user being processed, which is the number one; the facial emotion recognized and its confidence value, which is happy with 0.063; the vocal emotion recognized and its confidence value, which is happy with 0.0788; the final emotion assigned to this video and its confidence

value, which is happy with 0.1418.

3.4 Summary

In this chapter we started by presenting the RRHE architecture as a whole. Afterwards we explained the architecture and functionality of RRHE-Client, including components, technologies and the way it integrates with the remaining system. Afterwards, we presented the RRHE-Actuator, its capabilities and how they can be used in the system. In the end we presented the RRHE-Server and how it was built, its role, the technologies involved in each of its components and its function in the system.

In the following chapter we present the implementation details, explaining the design decisions that led to our final solution.

Chapter 4

Implementation

This chapter will focus on the implementation of RRHE three main modules for emotion recognition. namely: voice emotion recognition; facial emotion recognition; and the proposed emotion fusion technique. Hence, the following sections describe the techniques and algorithms used to implement each module.

4.1 Voice Emotions Extraction

This module uses the Support Vector Machine (SVM) algorithm from Matlab. SVM uses a binary classification based on statistical learning with data represented in a vectorial space. It finds an hyperplane of maximum margin through internal kernel functions to get the final classification. SMVs have the capacity to generalize new information accurately using trained models, which are created during the learning phase.

As in other problems, this classification is multi-class. In the scope of this thesis, there are 7 different classes that can be returned. For problems like this, there are various algorithms that can be applied, such as one-against-all (OAA), multi-class ranking and pairwise SVM.

We opted the one-against-all (OAA) algorithm which means that for a given input, all emotion classes are going to classify this parameter and return its degree of confidence. In the end, the chosen class is the one that presents the highest degree of confidence, after comparing all classes.

Using the highest degree of confidence alone as the decision factor, the number of missclassifications is increased if the two confidence values are relatively close. To mitigate this problem, we decided to implement the algorithm of hybrid kernel and thresholding fusion proposed by Yang et al. [62]. The implemented algorithm is shown in figure 4.1. During the training phase, for each utterance, 60 attributes 4.1.1 and the respective labels are used to feed the models X_i, where $i = 1, ..., 7$ corresponding to the number of emotions (emotional classes) used in this project. The best kernel function between quadratic, linear, polynomial, radial basis function (RBF) and multilayer perceptron (MLP) is calculated for each of the trained models, resulting in an hybrid kernel for the generated classifiers. The average (μ_i) and standard deviation (σ_i) of the confidence values returned during the training phase are calculated for

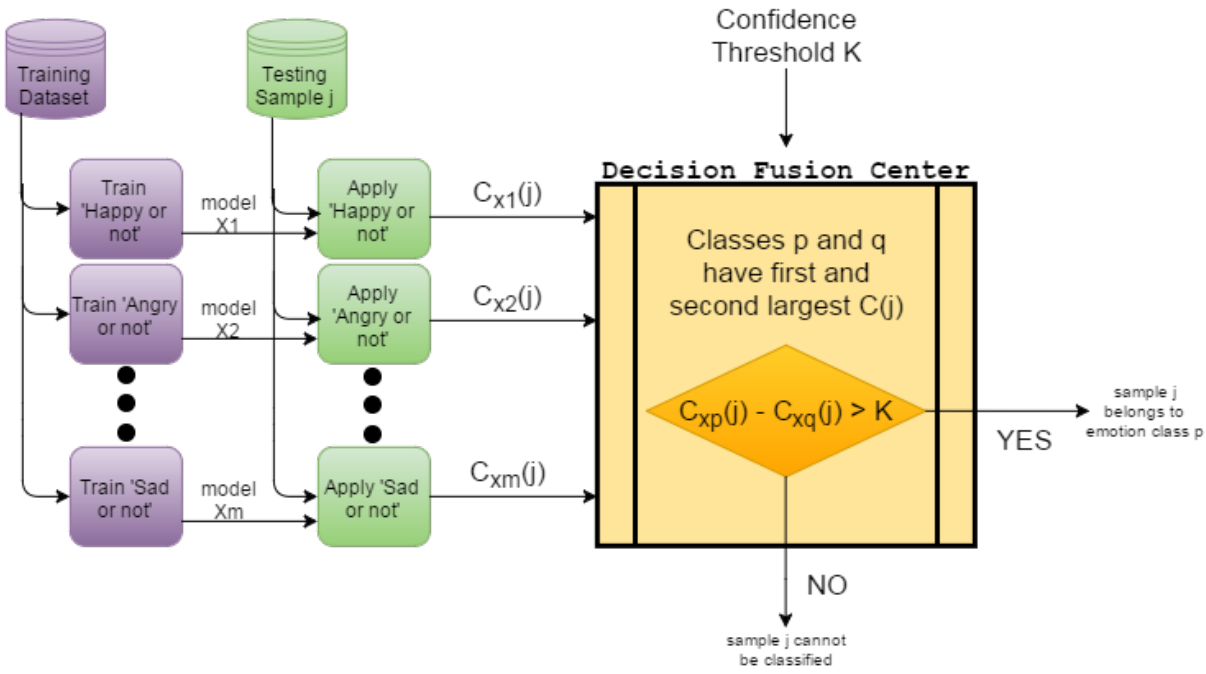

Figure 4.1: Hybrid kernel and thresholding fusion algorithm, adapted from [62].

each classifier. On the end, results the construction of the models trained for each class of classification (emotion).

4.1.1 Utterance Attributes

As just mentioned, 60 attributes were used to analyse each utterance, as follows. The algorithm uses 12 features:

- Pitch (1 feature): Defined as the relative lowness or highness with which a tone is perceived by the human hearing. Its value depends on the number of vibrations per second produced by the vocal chords. The pitch values are extracted and represented by cepstrum - the Inverse Fourier Transform (IFT) of the logarithm of the signal frequency spectrum - in the frequency domain.

- Energy (1 feature): The energy represents the speech intensity. It is calculated for each 60ms segment, by adding the amplitude of the squared values of each 1ms sample in the segment.

- Pitch difference and Energy difference (2 features): Is the difference between the pitch values and energy values of two contiguous segments. The higher the fluctuation of these values, the most evident is the presence of emotions.

- Formant: Calculated from the format (frequency and bandwidth) of the vocal channel. In the context of this project the frequency and bandwidth were used for the first four formants of each segment (2 x 4 = 8 features). Each formant was determined by the Linear Predictive Coding (LPC) method.

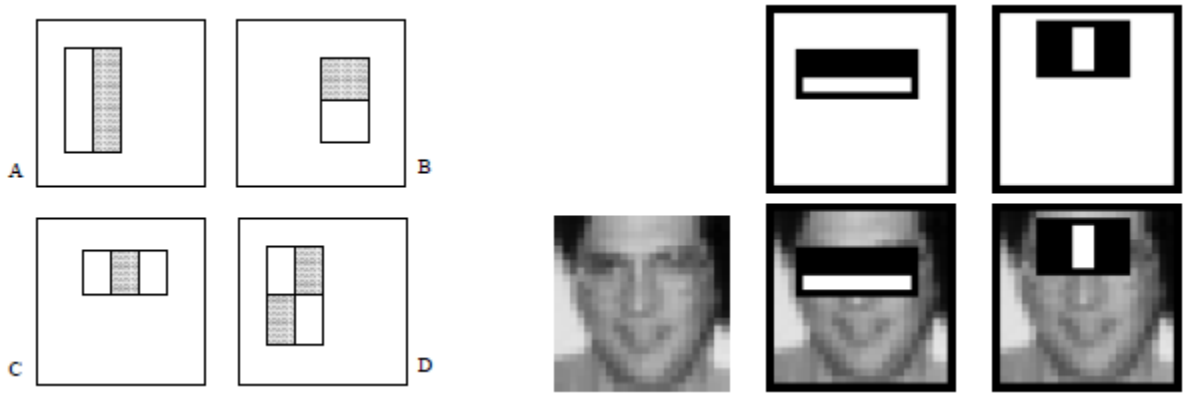

Figure 4.2: Haar features example.

The average, maximum, minimum, range and standard deviation in each 60ms segment are calculated for each of these 12 features. This way, for each segment we have 12 x 5 = 60 attributes that will be used in the classifier, either for training or for classification.

The concept of speaker-dependent emotion classification was not implemented, meaning that parameters specific to the speaker (e.g., sex) are not considered in the evaluation. This could be a future update to the algorithm which we believe could improve its results.

4.2 Face Emotions Extraction

The approach followed to implement this module is based on two previous research works [5], [28], where several algorithms are combined to achieve the emotional states recognition. The methodology proposed in this thesis is the following:

- Face detection and extraction;
- Facial features extraction through Gabor Filters Bank;
- Training SVM classifiers with the labelled data (e.g., emotion labels).

4.2.1 Face Detection

The face area represents the region of interest (ROI) in the context of this module. Therefore, we first focused on the detection of faces in an image. This kind of task has been widely discussed in the literature and several algorithm implementations exist. We adopted a Haar feature based cascade classifier [57], which is also part of the OpenCV framework.

This algorithm includes:

- Haar Features: these features are calculated in small windows of the image. In each window a binary mask is virtually applied and the value of the feature is the difference between the sum of the pixel values above the part of the mask with value 1 and the sum of the other part. An example of this kind of feature is provided in figure 4.2.

- Cascade Classification: this classification approach is based on several classification steps. At each step a different feature is considered and if a feature value does not match the trained model, the process is aborted and further stages are not evaluated.

A simple face rotation correction algorithm was implemented, which is based on the position of the eyes detected via specifically trained cascade classifiers. Once the eyes position is obtained, a simple trigonometry calculation is performed to get the value of the angle between the eye-line and the x-axis. This value is then used for the definition of the rotation matrix which is applied to the whole image.

4.2.2 Gabor Filters

Gabor filters are, roughly speaking, linear filters obtained by modulating a complex sinusoid with a Gaussian. These filters are typically used in image processing for tasks like edge detection, texture classification and face recognition. They are particularly effective in case of a time-frequency analysis, which is an analysis technique that aims to simultaneously study a signal in both time and frequency domain. This is due to multi resolution and multi-orientation properties. Multi-resolution is a method for orthonormal base creation by slicing the signal space into subspaces at different scales. One reason behind the great success of this kind of filters is the discovery that simple cells in the human visual cortex can be modelled with this particular filter. In this thesis we have adopted the following Gabor function formula:

$$g(x, y, \lambda, \theta, \psi, \sigma, \gamma) = exp(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2})exp(i(2\pi\frac{x'}{\lambda} + \psi)) \tag{4.1}$$

where $x' = xcos\theta + ysin\theta$ and $y' = -xsin\theta + ycos\theta$ represent the rotated component of the complex sinusoid, λ is the wavelength of the sinusoid, θ is the spatial orientation of the filter, ψ is the phase offset, σ is the standard deviation of the Gaussian support and γ is the aspect ratio factor (e.g. 1.0 for a circular shape).

Derived parameters can also be considered. For instance, the spatial frequency bandwidth of the filter is defined as:

$$b = log_2(\frac{\frac{\sigma}{\lambda}\pi + \sqrt{\frac{ln(2)}{2}}}{\frac{\sigma}{\lambda}\pi - \sqrt{\frac{ln(2)}{2}}}), \frac{\sigma}{\lambda} = \frac{1}{\pi}\sqrt{\frac{ln(2)}{2}}\frac{2^b + 1}{2^b - 1} \tag{4.2}$$

These additional relationship between λ, θ and b is useful for generating Gabor filter banks.

4.2.3 Gabor Filter Bank

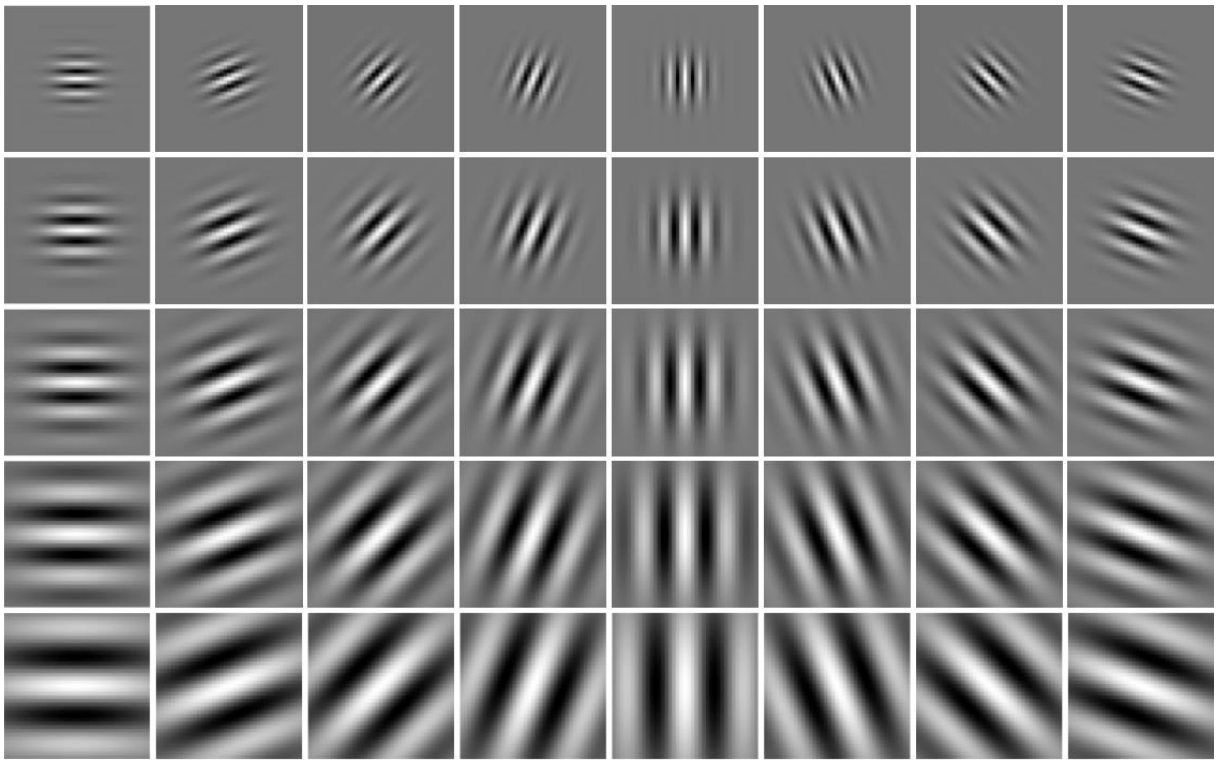

Figure 4.3: Example of a gabor filter bank.

Gabor Filter banks (see example in figure 4.3) are one of the main methods for selection of Gabor filters, typically adopted in texture segmentation problems. Families of filters are typically obtained by generating Gabor kernels with spatial frequencies λ, sinusoid orientation θ and bandwidth in ad hoc intervals, while scaling parameters are sometimes selected intuitively and assumed to be constant.

4.2.4 SVM Classifiers

We used the OpenCV implementation's of linear SVM, which realizes a C-Support linear SVM, this is, a linear SVM with Soft Margin, as follows:

$$min\frac{1}{2}\omega^T\omega + C * \sum_{i=1}^{l}\xi_i \tag{4.3}$$

$$y_i * (\omega^T\phi(x_i) + b) \geq 1 - \xi_i, \tag{4.4}$$

$$\xi_i \geq 0, i = 1, ..., l \tag{4.5}$$

where $\omega \cdot x - b = 0$ is the hyperplane: ω is the normal vector to the hyperplane and $x_i \in \mathbb{R}^n, i = 1, ...l$ are the training vector; $y \in \mathbb{R}^l, y_i \in \{1, -1\}$ is a class indicator, ξ_i is a non-negative slack variable measuring the degree of missclassification on x_i. The parameter C gives a weight to these missclassification variables. In other words, C is a trade-off between margin maximization and error minimization.

4.3 Emotion Fusion Technique

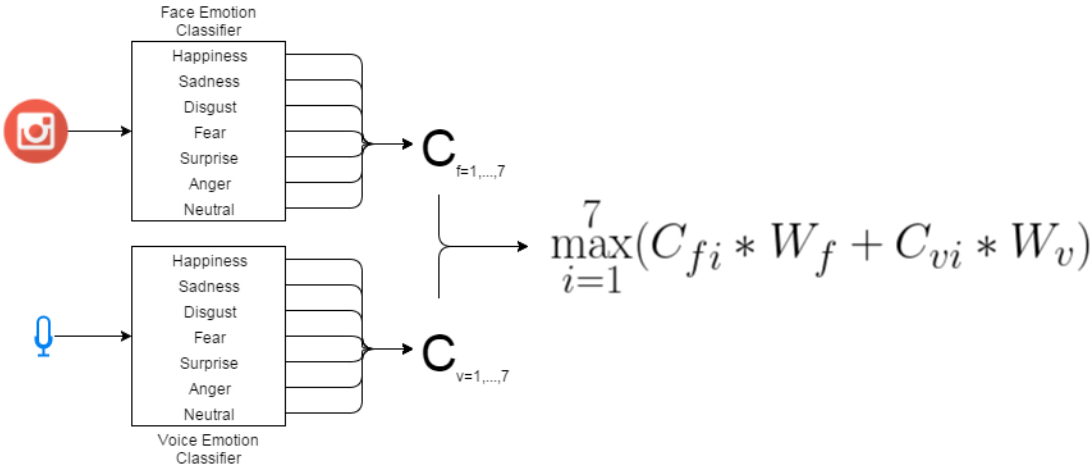

Figure 4.4: RRHE fusion algorithm.

Figure 4.4 illustrates our idea to combine both facial and vocal emotions. C_{fi} is the confidence degree for facial emotion i, W_f is the weight for face classifier, C_{vi} is the confidence degree for vocal emotion i and W_v is the weight for voice classifier, with $i = 1, ..., 7$ representing the 7 emotions supported by RRHE.

The idea behind this algorithm was as simple as to weigh up each one of the classifiers, and apply this weight to their emotional confidences degrees. Afterwards, the weighted emotional classes are summed together, and the final emotion will be the one with the maximum value.

We considered that face expressions are the most relevant element when evaluating an emotional scenario. Humans tend to reflect what is on their mind by actively issuing facial expressions. After several experiences we decided to weigh up the facial classifier with 60% and voice classifier with 40%.

Chapter 5

Evaluation

To validate this thesis – which presents a new approach for detecting and replicating human emotions – multiple evaluations were conducted that allow the comparison between results obtained with our system and other annotations (human and computerized).

Regarding system evaluation the extracted data aims to demonstrate the accuracy and performance of our solution. To train the SVM applied for face emotion recognition, it was used a set of images from the Cohn-Kanade Expression Database [29]. Likewise, to train the SVM applied for voice emotion recognition, it was used a set of recordings from LDC database [25].

The first tests had the objective to individually determine performance of the voice and face classifiers. Then, in order to have a comparison between RRHE and human annotations, we decided to make a questionnaire where 30 participants classify 30 videos regarding face emotion, voice emotion and overall emotion. To finalize, we compared the results obtained with the new Kinect V2 with our results.

During the tests the server was hosted in a machine with 16GB RAM, an Intel i7 Quad-Core 3.60GHz processor and an SSD disk, connected via Ethernet to a optical fiber network with a downstream of 99.77 Mbps and an upstream of 20.08 Mbps. The client PC was hosted in a machine with 6GB of RAM, Intel i5 Quad-Core 2.40GHZ processor and a 5400RPM hard disk connected via WiFi to a optical fiber network with 48.19 Mbps downstream and 20.20 Mbps upstream.

5.1 Experiences with Individual Classifiers

The goal of this experiment was to individually test each emotion classifier. For each classifier:

1. 100 images and sounds for each emotion were extracted as validation data from the DB used during the classifier's training phase, with the goal of testing if the algorithms recognize well the data used for their training (no generalization);

2. 50 images and sounds for each emotion were extracted from the validation set of the same DB used to train the classifier thus already requiring some generalization capability;

3. 100 images and sounds for each emotion were extracted from a different DB than the one used

to train the classifier, which corresponds to a more demanding test concerning the generalization capability.

For each classifier, it is presented the confusion matrix to characterize the classification error rate. Performance results for several metrics are also presented for each test.

5.1.1 Face emotion classifier

The test 1 of the face emotion classifier employed 100 images for each emotion used for the classifier's training. The following confusion matrix and performance results were obtained from the experimental results.

- Confusion Matrix

	Happiness	**Sadness**	**Fear**	**Anger**	**Disgust**	**Surprise**	**Neutral**
Happiness	96	0	0	0	0	0	4
Sadness	0	94	1	1	2	0	2
Fear	0	2	95	1	2	0	0
Anger	0	2	4	94	0	0	0
Disgust	0	2	3	0	95	0	0
Surprise	0	0	0	0	3	93	4
Neutral	5	3	0	0	0	3	89

Table 5.1: Confusion Matrix for the first test - lines represent labelled images and columns show RRHE classifications.

Table 5.1 shows that RRHE was right in 656 out of the 700 tests. The emotional class where RRHE achieved the best results was Happiness while the worst results were obtained in the Neutral class. The following data demonstrates the performance of this test both concerning classification accuracy and processing delays.

- Performance Analysis

	% Correct Classifications	**# False Positives**	**# False Negatives**	**Processing Delays**			
				Average	**Standard Deviation**	**Max**	**Min**
Happiness	96%	5	4	1.48s	0.20s	1.97s	1.24s
Sadness	94%	9	6	1.46s	0.29s	2.01s	1.01s
Fear	95%	8	5	1.90s	0.09s	1.97s	1.54s
Anger	94%	2	6	1.57s	0.16s	1.97s	1.37s
Disgust	95%	7	5	1.48s	0.12s	1.90s	1.37s
Surprise	93%	3	7	1.99s	0.09s	2.05s	1.52s
Neutral	89%	10	11	2.44s	0.13s	2.60s	2.13s
System	93.7%	44	44	1.76s	0.15s	2.60s	1.01s

Table 5.2: Performance Analysis for the first test.

The test 2 of the face emotion classifier employed 50 images for each emotion used from the validation set of the same DB used for the classifier's training. The following confusion matrix and performance

results were obtained from the experimental results.

- Confusion Matrix

	Happiness	**Sadness**	**Fear**	**Anger**	**Disgust**	**Surprise**	**Neutral**
Happiness	43	0	0	0	0	4	3
Sadness	0	43	1	0	4	0	2
Fear	0	0	39	0	5	0	6
Anger	0	2	2	37	8	0	1
Disgust	0	3	7	4	35	0	1
Surprise	5	0	0	1	0	41	3
Neutral	3	3	0	0	4	0	40

Table 5.3: Confusion Matrix for the second test - lines represent labelled images and columns show RRHE classifications.

Table 5.3 shows that RRHE loses quality when the images do not belong to the set of images used to train the classifier. Given the nature of the algorithm this was the expected behaviour. RRHE was still right in 278 of the 350 tests, showing the worst results in the Disgust emotional class - with only 70% of correct guesses.

- Performance Analysis

	% Correct Classifications	**# False Positives**	**# False Negatives**	**Processing Delays**			
				Average	**Standard Deviation**	**Max**	**Min**
Happiness	86%	8	7	1.44s	0.18s	1.96s	1.25s
Sadness	86%	8	7	1.69s	0.14s	1.85s	1.32s
Fear	78%	10	11	1.79s	0.27s	2.10s	1.11s
Anger	74%	5	13	1.88s	0.22s	2.11s	1.29s
Disgust	70%	21	15	1.78s	0.16s	1.89s	1.09s
Surprise	82%	4	9	1.99s	0.11s	2.31s	1.87s
Neutral	80%	16	10	2.21s	0.14s	2.40s	1.91s
System	79.43%	72	72	1.83s	0.17s	2.40s	1.09s

Table 5.4: Performance Analysis for the second test.

The test 3 of the face emotion classifier employed 100 images for each emotion used from JAFFE Database [30]. The following confusion matrix and performance results were obtained from the experimental results.

- Confusion Matrix

	Happiness	Sadness	Fear	Anger	Disgust	Surprise	Neutral
Happiness	52	0	0	19	0	2	27
Sadness	0	49	5	0	14	0	32
Fear	0	16	49	5	26	0	4
Anger	0	15	21	42	17	0	5
Disgust	0	17	14	10	41	12	6
Surprise	16	0	0	12	17	50	5
Neutral	6	9	5	9	3	7	61

Table 5.5: Confusion Matrix for the third test - lines represent labelled images and columns show RRHE classifications.

Table 5.5 shows RRHE results with images from the JAFFE database. RRHE correctly identified only 344 of the 700 emotions. The poor performance in this test is explained by the nature of the images. While the training set consisted of only Caucasian people, the evaluation set consisted of only Asian people - Japanese to be more precise. The considerable physical differences between the two races (e.g. the slanted eyes) have a big impact on learning algorithms as is the case of SVM. With two disjunct mixtures of the two databases for training and evaluation respectively RRHE would probably get better results.

- Performance Analysis

	% Correct Classifications	# False Positives	# False Negatives	Processing Delays			
				Average	Standard Deviation	Max	Min
Happiness	52%	22	48	1.57s	0.12s	1.69s	1.22s
Sadness	49%	57	51	1.84s	0.01s	1.86s	1.81s
Fear	49%	45	51	1.79s	0.15s	2.09s	1.58s
Anger	42%	55	58	2.03s	0.04s	2.12s	1.97s
Disgust	41%	77	59	1.55s	0.22s	1.97s	1.21s
Surprise	50%	21	50	1.73s	0.07s	1.99s	1.66s
Neutral	61%	79	39	2.26s	0.16s	2.49s	1.95s
System	49.14%	356	356	1.82s	0.11s	2.49s	1.21s

Table 5.6: Performance Analysis for the third test.

5.1.2 Voice emotion classifier

The test 1 of the voice emotion classifier employed 100 sounds for each emotion used for the classifier's training. The following confusion matrix and performance results were obtained from the experimental results.

- Confusion Matrix

	Happiness	Sadness	Fear	Anger	Disgust	Surprise	Neutral
Happiness	100	0	0	0	0	0	0
Sadness	0	97	0	0	1	0	2
Fear	0	0	98	0	0	0	2
Anger	0	0	0	97	0	0	3
Disgust	0	0	0	0	99	0	1
Surprise	0	0	0	0	0	100	0
Neutral	3	2	0	0	0	0	95

Table 5.7: Confusion Matrix for the first test - lines represent labelled sounds and columns show RRHE classifications.

As shown in table 5.7 our voice classifier was right in 98% of the tests. Therefore, we can say that our algorithm works well when evaluation data and training data are extracted from the same dataset.

- Performance Analysis

	% Correct Classifications	# False Positives	# False Negatives	Processing Delays			
				Average	Standard Deviation	Max	Min
Happiness	100%	3	0	2.96s	0.44s	3.56s	2.02s
Sadness	97%	2	3	3.22s	0.15s	3.47s	2.97s
Fear	98%	0	2	2.99s	0.05s	3.01s	2.58s
Anger	97%	0	3	3.44s	0.27s	3.88s	2.96s
Disgust	99%	1	1	3.57s	0.27s	4.01s	3.11s
Surprise	100%	0	0	3.07s	0.18s	3.54s	2.85s
Neutral	95%	8	5	2.98s	0.09s	3.22s	2.87s
System	98%	14	14	3.18s	0.21s	4.01s	2.02s

Table 5.8: Performance Analysis for the first test.

The test 2 of the voice emotion classifier employed 50 sounds for each emotion used from the validation set of the same DB used for the classifier's training. The following confusion matrix and performance results were obtained from the experimental results.

- Confusion Matrix

	Happiness	Sadness	Fear	Anger	Disgust	Surprise	Neutral
Happiness	48	0	0	0	0	0	2
Sadness	0	49	0	0	0	0	1
Fear	0	0	45	3	0	0	2
Anger	0	2	0	47	0	0	1
Disgust	0	2	0	0	43	0	5
Surprise	1	0	0	0	0	47	2
Neutral	3	2	0	0	0	0	45

Table 5.9: Confusion Matrix for the second test - lines represent labelled sounds and columns show RRHE classifications.

Table 5.9 shows that our voice classifier performs well with data extracted from the training database. This proves that vocal properties used for training and evaluation were well defined since we got good results in a set which the classifier does not know.

- Performance Analysis

	% Correct Classifications	# False Positives	# False Negatives	Processing Delays			
				Average	Standard Deviation	Max	Min
Happiness	96%	4	2	2.98s	0.36s	3.45s	2.22s
Sadness	98%	6	1	3.02s	0.26s	3.45s	2.58s
Fear	90%	0	5	2.69s	0.21s	3.00s	2.28s
Anger	94%	3	3	3.00s	0.10s	3.12s	2.76s
Disgust	86%	0	7	3.88s	0.23s	4.10s	3.21s
Surprise	94%	0	3	2.95s	0.14s	3.23s	2.76s
Neutral	90%	13	5	2.93s	0.10s	3.11s	2.78s
System	92.57%	26	26	3.06s	0.20s	4.10s	2.22s

Table 5.10: Performance Analysis for the fifth test.

The test 3 of the voice emotion classifier employed 100 sounds for each emotion used from UGA Database. The following confusion matrix and performance results were obtained from the experimental results.

- Confusion Matrix

	Happiness	Sadness	Fear	Anger	Disgust	Surprise	Neutral
Happiness	21	0	0	10	0	14	55
Sadness	0	17	0	0	9	0	74
Fear	0	0	8	4	0	5	83
Anger	1	0	0	23	0	1	75
Disgust	2	10	0	0	9	0	79
Surprise	0	0	0	0	0	15	85
Neutral	0	0	0	17	0	3	80

Table 5.11: Confusion Matrix for the sixth test - lines represent labelled sounds and columns show RRHE classifications.

The worst results obtained by RRHE are shown in table 5.11. We can see that the voice classifier behaves badly with data unrelated with the training data. One of the crucial reasons for that is the fact that we have a database with sound files that contain a lot of background noise. A background noise removal method could prove very efficient in this case. The 80% correct classifications and 451 false positives in the neutral emotional state reveal that in case of ”doubt” the chosen state is neutral most of the times.

- Performance Analysis

	% Correct Classifications	# False Positives	# False Negatives	Processing Delays			
				Average	Standard Deviation	Max	Min
Happiness	21%	3	79	2.76s	0.45s	3.67s	2.12s
Sadness	17%	10	83	3.23s	0.47s	3.97s	2.37s
Fear	8%	0	92	3.50s	0.49s	4.03s	2.14s
Anger	23%	31	77	2.95s	0.29s	3.56s	2.55s
Disgust	9%	9	91	3.99s	0.09s	4.11s	3.79s
Surprise	15%	23	85	3.04s	0.28s	3.42s	2.43s
Neutral	80%	451	20	2.97s	0.20s	3.34s	2.65s
System	24.71%	527	527	3.21s	0.32s	4.11s	2.12s

Table 5.12: Performance Analysis for the sixth test.

5.1.3 Summary

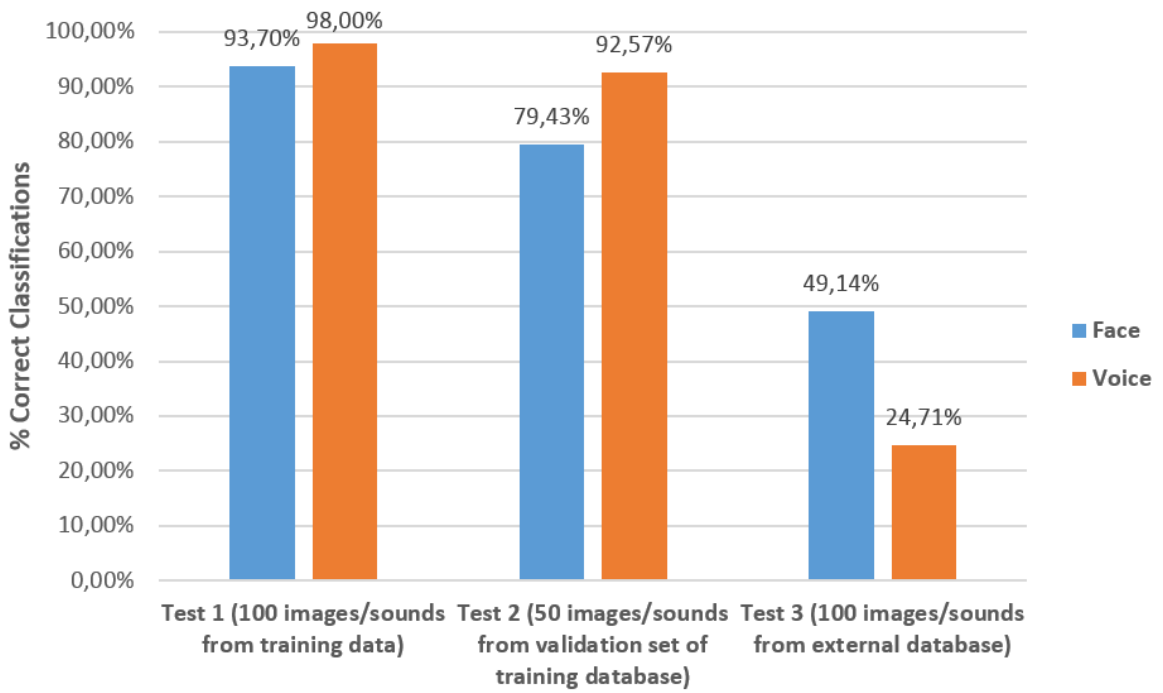

Figure 5.1: Percentage of correct classifications.

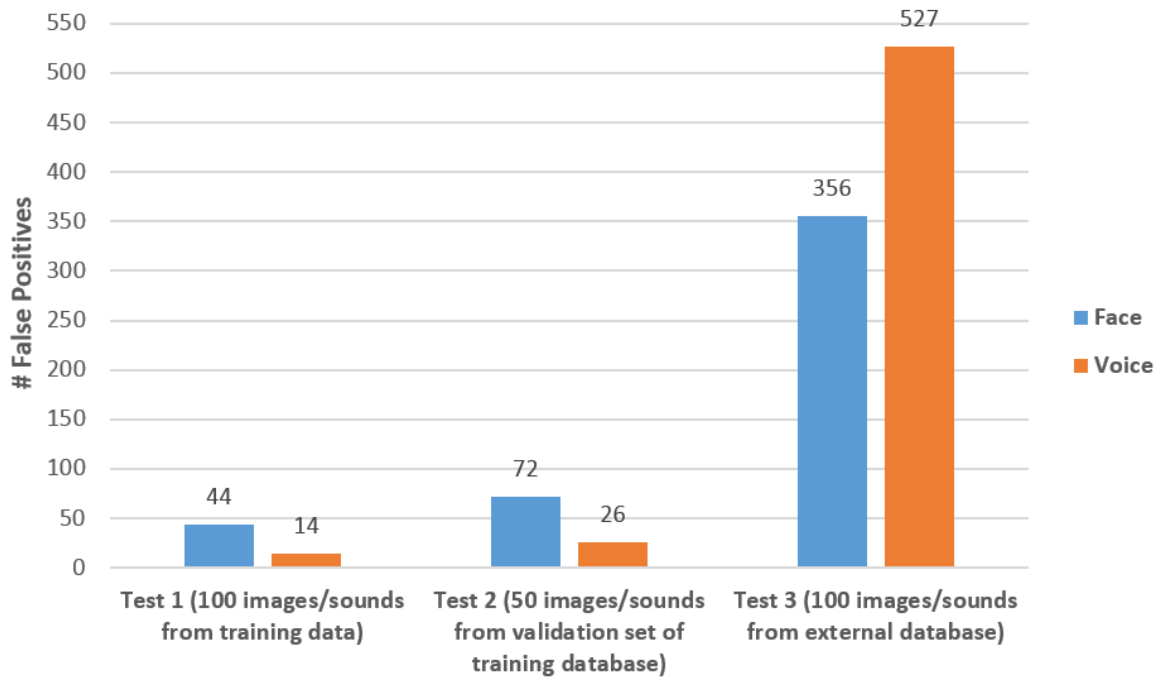

Figure 5.2: Number of false positive results.

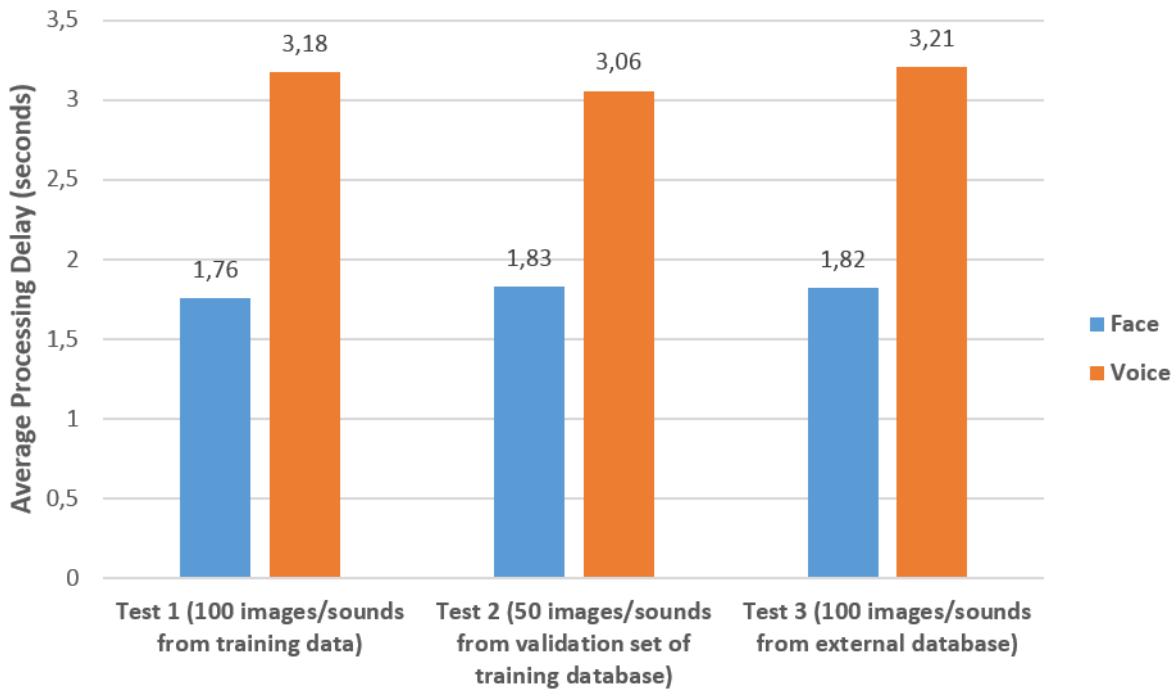

Figure 5.3: Average processing delay, in seconds.

Figure 5.1 compares the percentage of matches from both classifiers in all tests. As we can see, both the facial emotion classifier and the vocal emotion classifier had excellent results in the first test: above 90%. This proves that when the data used for evaluation is the same data that was used for training the accuracy is very high. It is important to note that, albeit by a small margin, the vocal classifier obtained a better result than the facial classifier, which shows that this classifier has a deeper connection with

the training data, being able to extract relevant properties during the learning process. As expected, although the percentage of matches is still very high, the second test shows a reduction in the match percentage of both classifiers. In this case the voice classifier has even better results when comparing with the facial classifier, which shows once again that the sound properties used for training are very useful for the classifier when evaluating and training data have similar properties. The bad results of the third test are debatable. Starting with the facial classifier, we believe that the bad results are mainly due to the noticeable difference between the images used in the training and evaluation phases. For the vocal classifier, the explanation resides on the presence of a distinguishable background noise. This is a problem identified by our algorithm that could be alleviated by a background noise removal phase before the classification of the sound. Figure 5.2 shows the number of false positives for each classifier in each test.

Figure 5.3 shows the analysis to the performance of RRHE-Server and the classifiers. Firstly, it is important to note that the Processing Delays are relatively close to each other in all tests, which shows that the performance of the classifiers is not affected by the data. We can see that the facial classifier is faster than the vocal classifier by a factor of 2 which shows that the analysis of a sound file involves a lot more operations than the analysis of an image file. Finally, it is important to state that the timings of RRHE-Server are acceptable in a system with the capacity to work remotely. Considering that the client-server and server-actuator exchange of messages was made over the Internet, a maximum of 3 seconds of delay until the emotion replication is a good result.

5.2 Manual Annotation With Questionnaire

This experiment involved a total of 30 people whom individually answered the questionnaire. The number and profile of people was chosen so that there are as many answers of people coming from distinct professional areas as possible.

The procedure of this experiment was the same for all participants so that answers are not influenced by non-controllable variables. The age of the respondents varied between 19 and 40 years. All respondents have higher education and their professional areas are as follows:

1. 3.33% working in agriculture;
2. 13.33% are computer engineers;
3. 10% work in health;
4. 60% are students (in computer science, psychology and medicine);
5. 6.67% are high-school teachers;
6. 6.67% are architects.

Out of all users, 63.33% were male and 36.67% were female. Regarding nationality, all users were Portuguese and all understand English quite well. All sessions happened in the same room and using the same computer. Each session consisted of:

1. 10 minutes of presentation of the work and the goal of the questionnaires – this presentation was verbal with the help of a set of slides to illustrate the most important concepts;
2. 5 minutes to fill out a pre-session questionnaire;
3. 35 minutes for the session itself and filling the questionnaire;
4. 10 minutes for demonstration and interaction with RRHE.

5.2.1 Pre-Session Questionnaire

The pre-session questionnaire (Appendix A) had the goal of collecting information about the respondent, understand their level of knowledge about the matter and understand their decision making process in the following questionnaire. The results of this questionnaire are as follows:

What is your gender?	Male	Female				
	19	11				
How old are you?	<20	20-25	26-30	31-35	36-40	>40
	4	16	5	2	3	0
What is your job area?	Agriculture	Engineering	Health	Student	Teaching	Other
	1	4	3	18	2	2
Have you ever had any practice detecting emotions?	No	Yes				
	18	12				
Are you able to identify emotions looking to a face?	Yes, for sure	Only the most basic ones	I'm not sure	No, I don't		
	17	5	8	0		
Are you able to identify emotions listening to a human speech?	Yes, for sure	Only the most basic ones	I'm not sure	No, I don't		
	13	7	10	0		
Are you able to identify an irony when looking and listening to other people?	Yes, for sure	I'm not sure	No, I don't			
	13	17	0			
How do you think you can extract more emotional information?	Biometrics data	Emotional Speech	Facial Expressions	Text Analysis		
	5	7	16	2		
Classify your knowledge on each one of the following emotions (1 to 5, where 1 means poor and 5 means good knowledge).	1	2	3	4	5	
	Happiness: 0	Happiness: 0	Happiness: 0	Happiness: 8	Happiness: 22	
	Surprise: 0	Surprise: 0	Surprise: 3	Surprise: 18	Surprise: 9	
	Fear: 0	Fear: 0	Fear: 3	Fear: 21	Fear: 6	
	Disgust: 0	Disgust: 0	Disgust: 0	Disgust: 18	Disgust: 12	
	Anger: 0	Anger: 0	Anger: 3	Anger: 21	Anger: 6	
	Sadness: 0	Sadness: 0	Sadness: 0	Sadness: 8	Sadness: 22	
	Neutral: 0	Neutral: 5	Neutral: 21	Neutral: 4	Neutral: 0	
When classifying an emotion what are the metrics with more impact on your choice?	According to the situation I measure each metric	Always choose the most obvious one	Always use the same technique			
	17	10	3			

5.2.2 Session Questionnaire

During the session the respondents faced a set of 30 videos for observation. Each of these videos had an image with a facial expression and a human voice also expressing an emotion. These videos were

compiled from the validation sets in the database used for the classifier training and had the duration of 1 second. Each video was repeated 5 times with a time interval of 10 seconds, so that the respondents had time to correctly understand the transmitted emotions – facial and vocal.

In each video the users were asked to classify the interpreted emotions in 3 categories: Facial, Vocal and Overall video emotions.

In most cases all respondents were able to immediately classify the emotions exposed by the facial expression and voice separately. On the contrary, the respondents have shown more difficulties in classifying the videos as a whole. We concluded that the respondents were able to quickly classify the video when the voice and face emotions matched each-other. When the emotions were distinct, in 84.33% instances the respondents gave more weight to the facial expression while 12.33% were able to identify the voice and facial expression and only 3.33% classified the video using only the voice emotion.

Session Questionnaire

*Required

i. What emotions do you recognize on segment i? *

Choose only one option per row.

Mark only one oval per row.

	Happiness	Sadness	Fear	Anger	Disgust	Surprise	Neutral
Face							
Voice							
Combined							

Powered by Google Forms

Figure 5.4: Session Questionnaire - where 'i' goes from 1 to 30

The questionnaire has the structure presented on Figure 5.4 where 'i' is the question number, which goes from 1 to 30. The results are as follows:

1. What emotions do you recognize on the first segment?

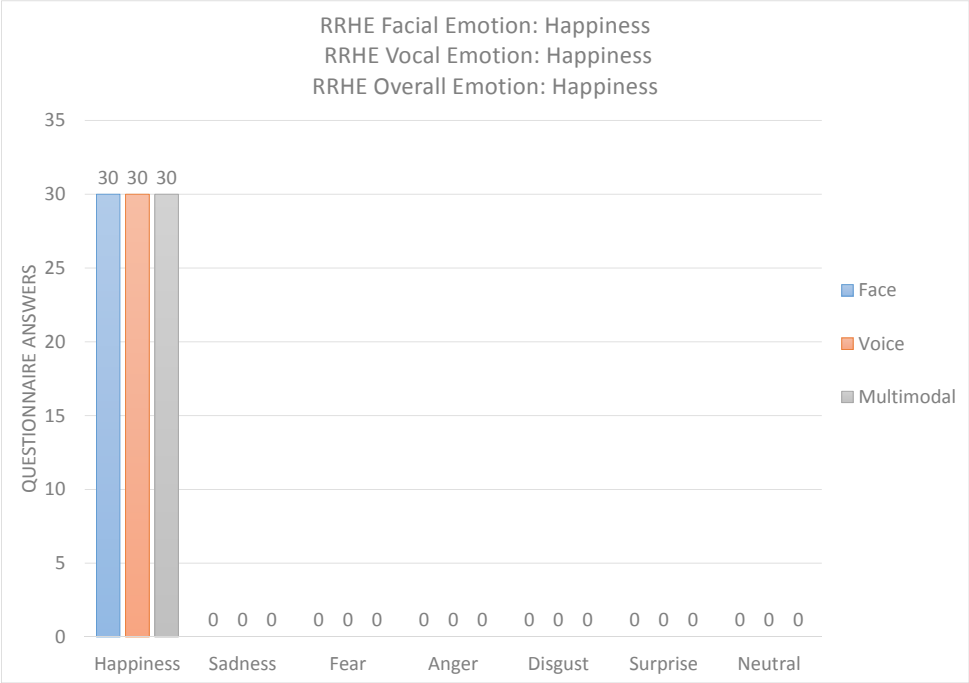

Figure 5.5: Session Questionnaire - Question 1 results.

Thanks to the clear expression of the emotions in this video, all respondents identified the facial, vocal and video emotion with ease. All respondents classified the emotion in accordance to RRHE in all categories.

2. What emotions do you recognize on the second segment?

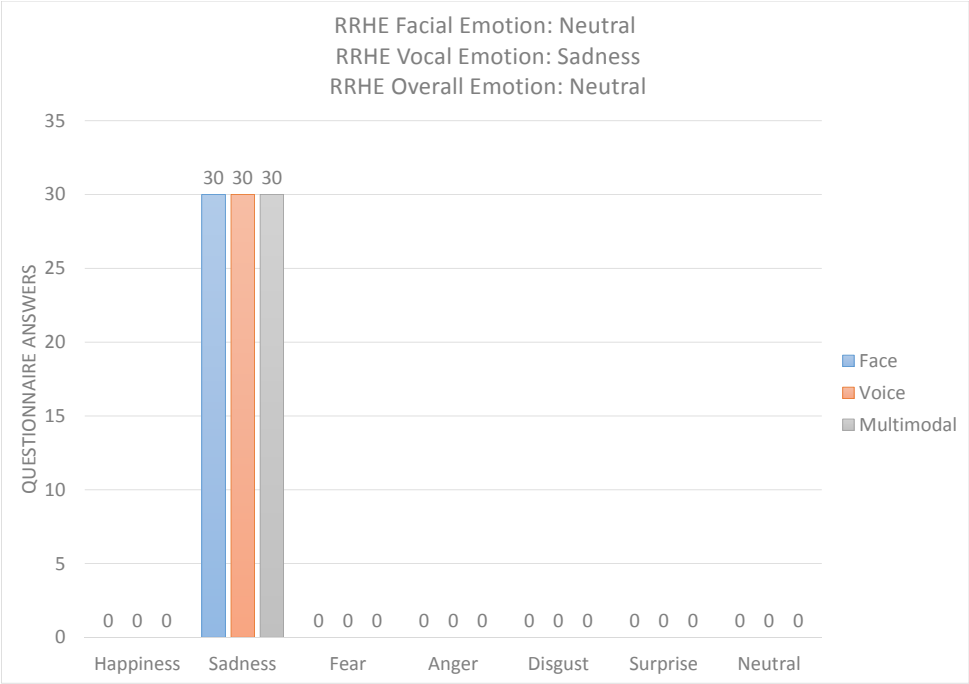

Figure 5.6: Session Questionnaire - Question 2 results.

Once again, users had no doubt when it came to classifying this video as sadness. In this video,

the answers of the respondents diverged from those of RRHE which classified face and overall video as neutral. RRHE agreed with the respondents only in the voice emotion. Since RRHE gave a bigger weight to the emotion extracted from the facial recognition, its final verdict was not correct.

3. What emotions do you recognize on the third segment?

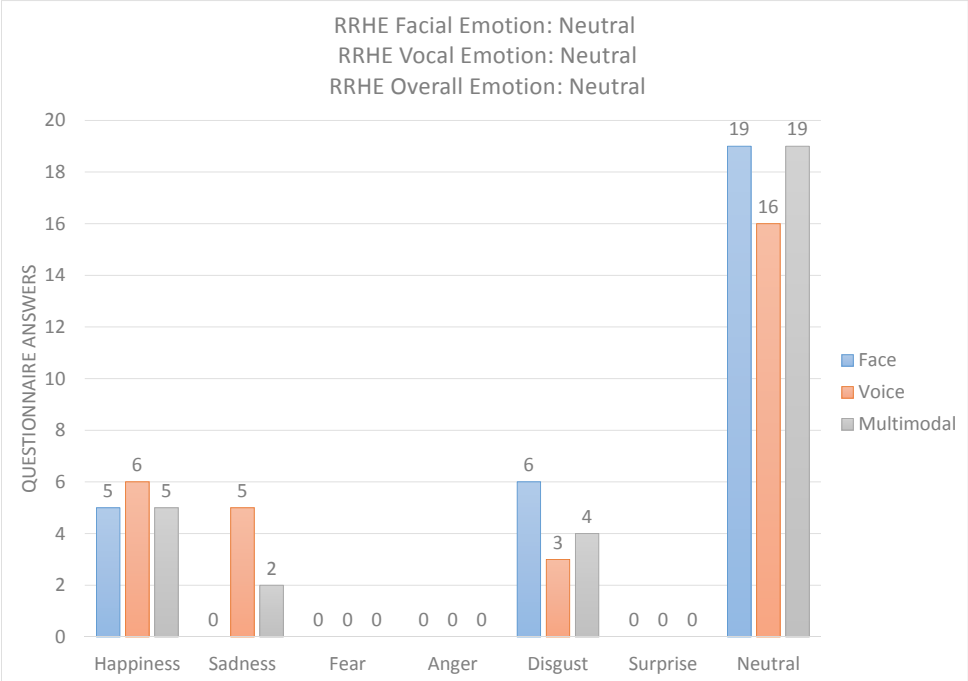

Figure 5.7: Session Questionnaire - Question 3 results.

The weak expression of emotions in this video caused difficulties to the respondents. We can see that the facial expression had a bigger weight in the choice of the final emotion with most respondents agreeing with RRHE at all levels.

4. What emotions do you recognize on the fourth segment?

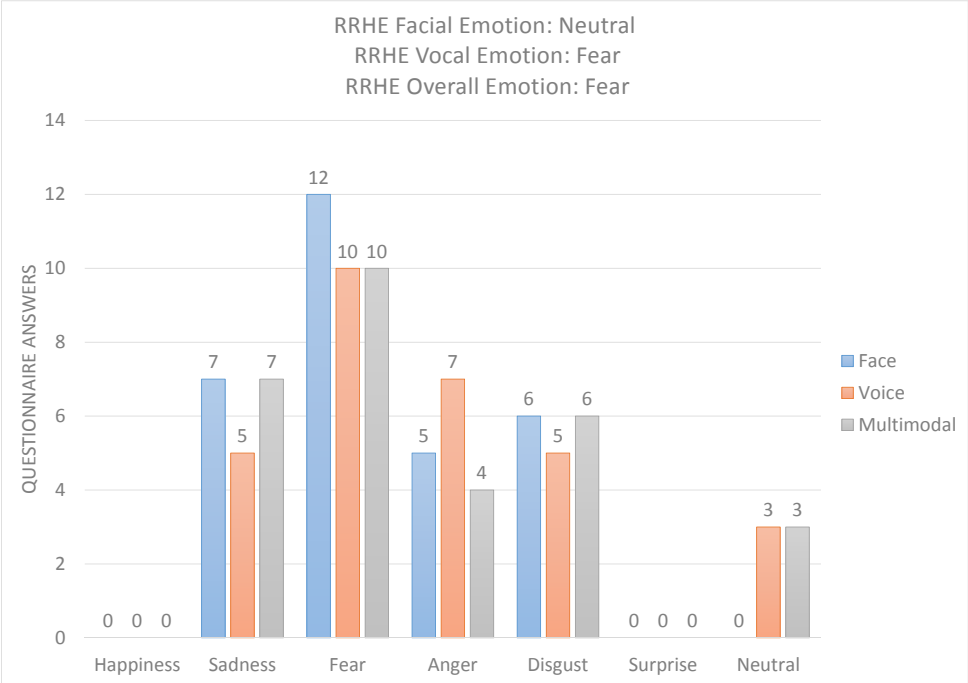

Figure 5.8: Session Questionnaire - Question 4 results.

The results in this video are spread, but most respondents gave more weight to the vocal expression. Although the facial expression chosen by RRHE was not in accordance to that of the respondents, a bigger weight was given to the voice and the final answer was in accordance to that of the respondents.

5. What emotions do you recognize on the fifth segment?

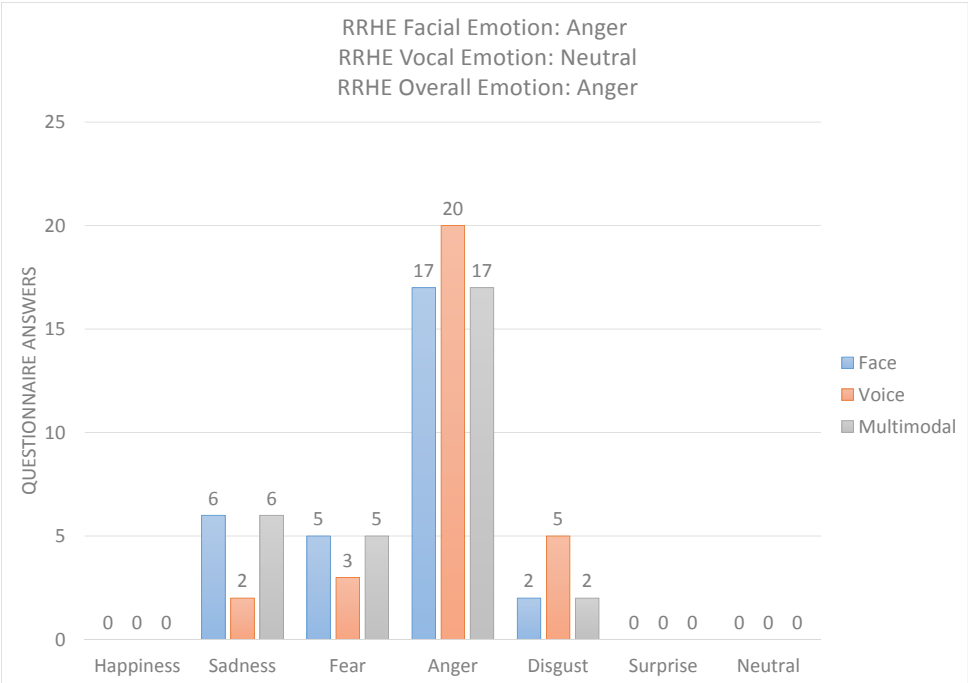

Figure 5.9: Session Questionnaire - Question 5 results.

In this video the facial expression was the key to the classification, both for the respondents and for RRHE. Most of the respondents found the voice more expressive than the face as opposed to RRHE who gave a bigger weight to the face.

6. What emotions do you recognize on the sixth segment?

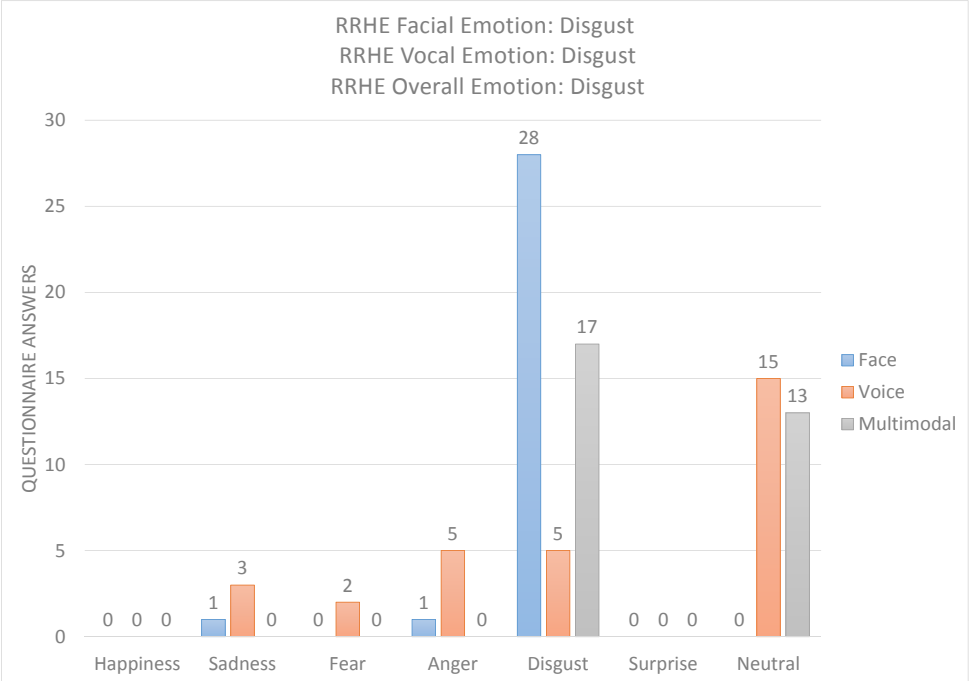

Figure 5.10: Session Questionnaire - Question 6 results.

Respondents had trouble to decide on the voice emotion of this video. The key for the classification was the facial expression which was clear, leading to the same overall classification both from respondents and for RRHE which had no doubts choosing the expression.

7. What emotions do you recognize on the seventh segment?

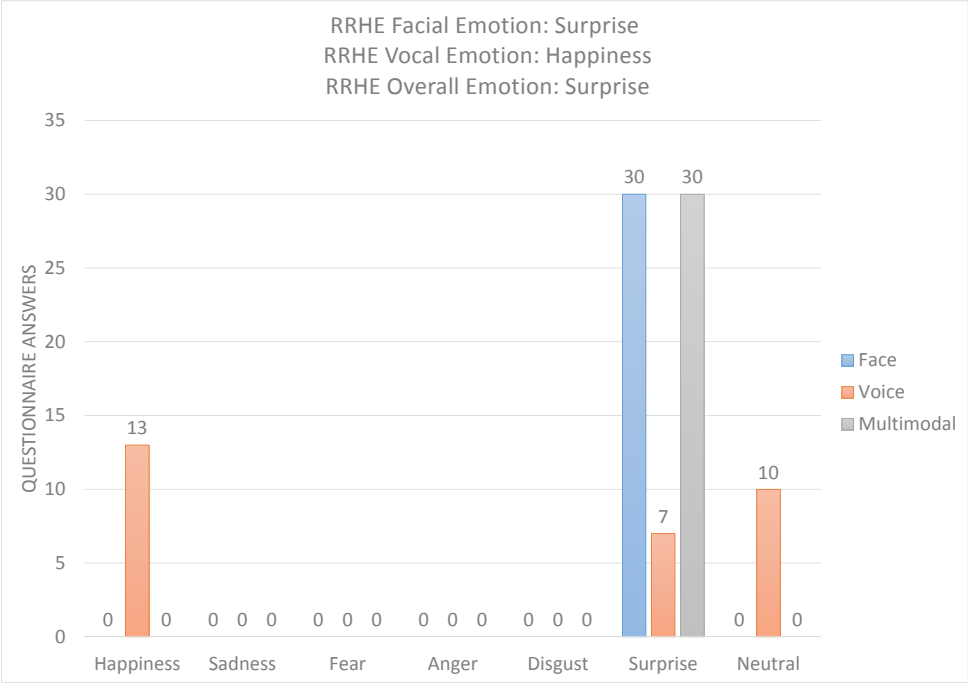

Figure 5.11: Session Questionnaire - Question 7 results.

The respondents and RRHE had no doubt when classifying this video as a surprise emotion due to the clear facial expression. Indeed, even some of the test subjects attributed a surprise or neutral emotion to the voice, although most votes were for the happiness emotion on the audio signal. But all subjects were in agreement with RRHE concerning the recognized emotion when considering both modalities.

8. What emotions do you recognize on the eighth segment?

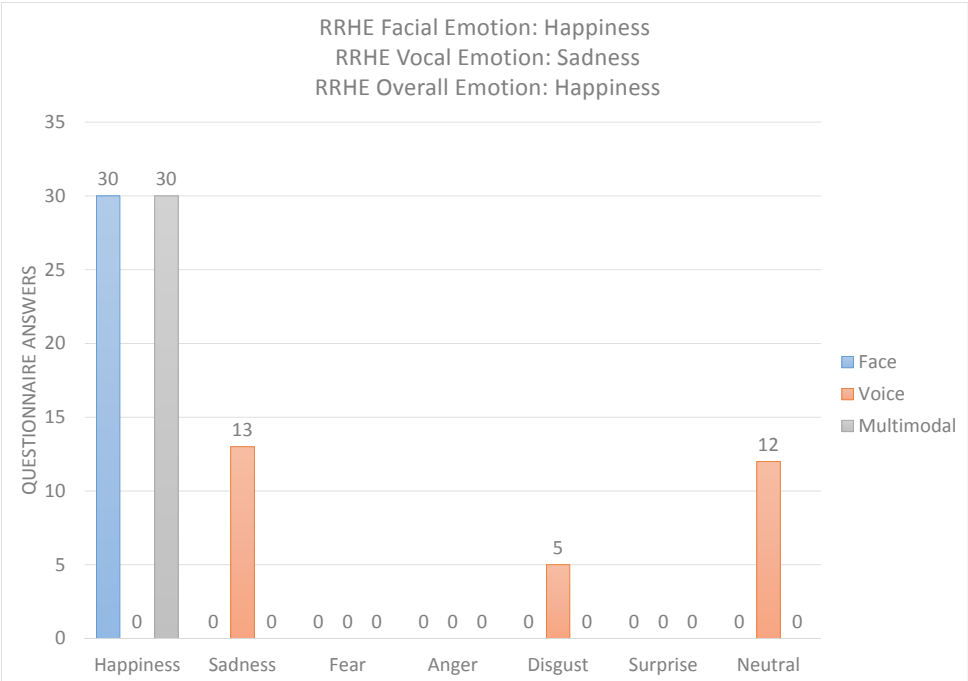

Figure 5.12: Session Questionnaire - Question 8 results.

This is a clear example of irony. The video shows a happy face and a sad voice. Some respondents did not consider the voice expressive enough and classified it as neutral. Overall, since the voice expression was not so clear, both respondents and RRHE attributed an identical emotion classification for the video.

9. What emotions do you recognize on the ninth segment?

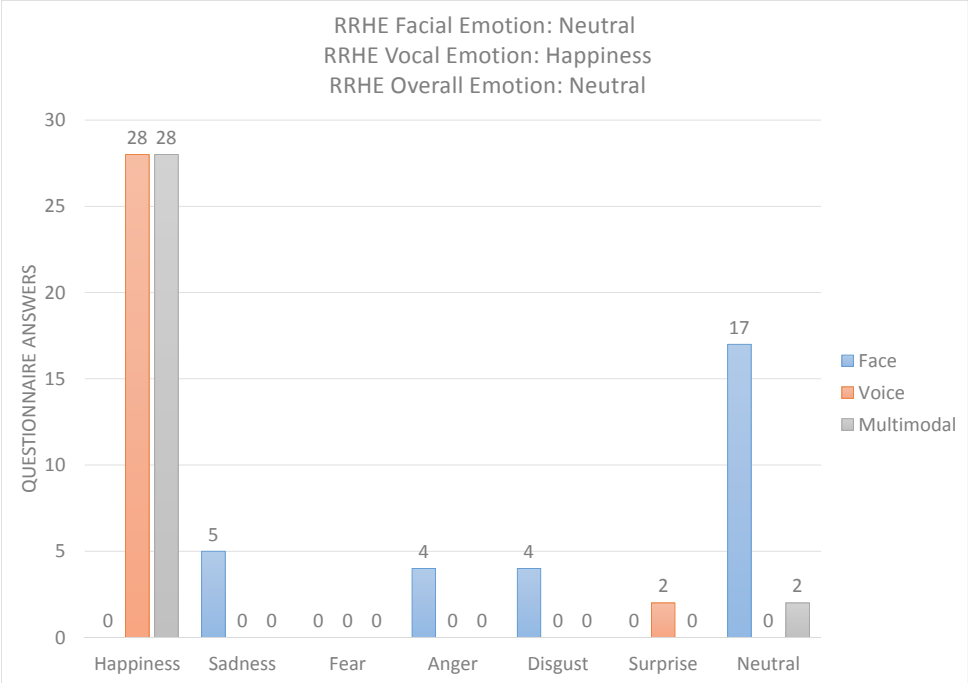

Figure 5.13: Session Questionnaire - Question 9 results.

Contrarily to the previous video, the respondents considered the voice the most important input when classifying the video. Although most respondents – and RRHE - classified the facial expression as neutral, this was not enough to overcome the vocal expression. With these videos we can conclude that the respondents take into account both the vocal and the facial expressions.

10. What emotions do you recognize on the tenth segment?

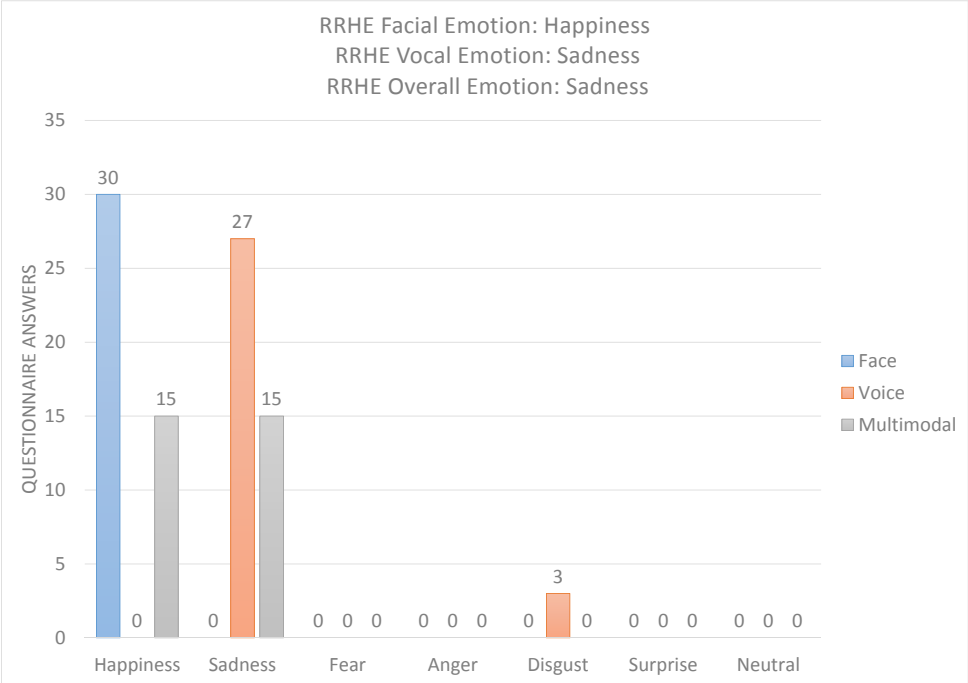

Figure 5.14: Session Questionnaire - Question 10 results.

This video shows another irony, with a happy face and a sad voice. The respondents were divided in the final classification, while RRHE gave a bigger emphasis to the voice and classified the video as Sadness.

11. What emotions do you recognize on the eleventh segment?

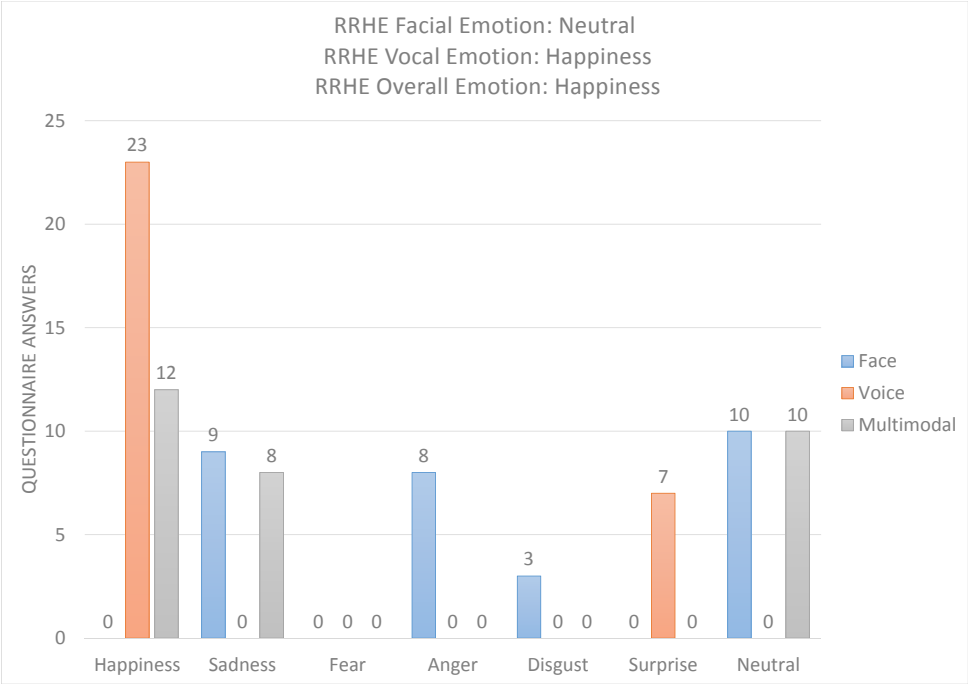

Figure 5.15: Session Questionnaire - Question 11 results.

In this video the vocal expression was obvious for RRHE and for the respondents. The facial

expression divided the respondents but most agreed with RRHE and classified the video as Happiness.

12. What emotions do you recognize on the twelfth segment?

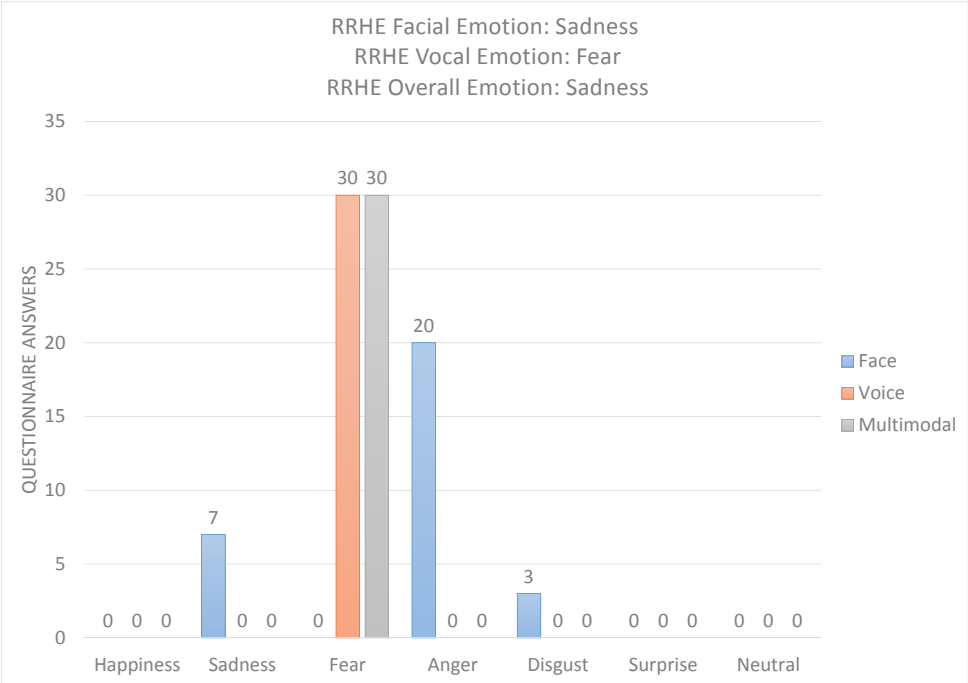

Figure 5.16: Session Questionnaire - Question 12 results.

This video shows once again that the respondents do not always follow the same principles. Although most of the respondents considered the face to be expressing anger, all agreed that the overall expression was dictated by the voice. RRHE agreed with the respondents on the vocal and facial expressions but gave a bigger weight to the facial expression.

13. What emotions do you recognize on the thirteenth segment?

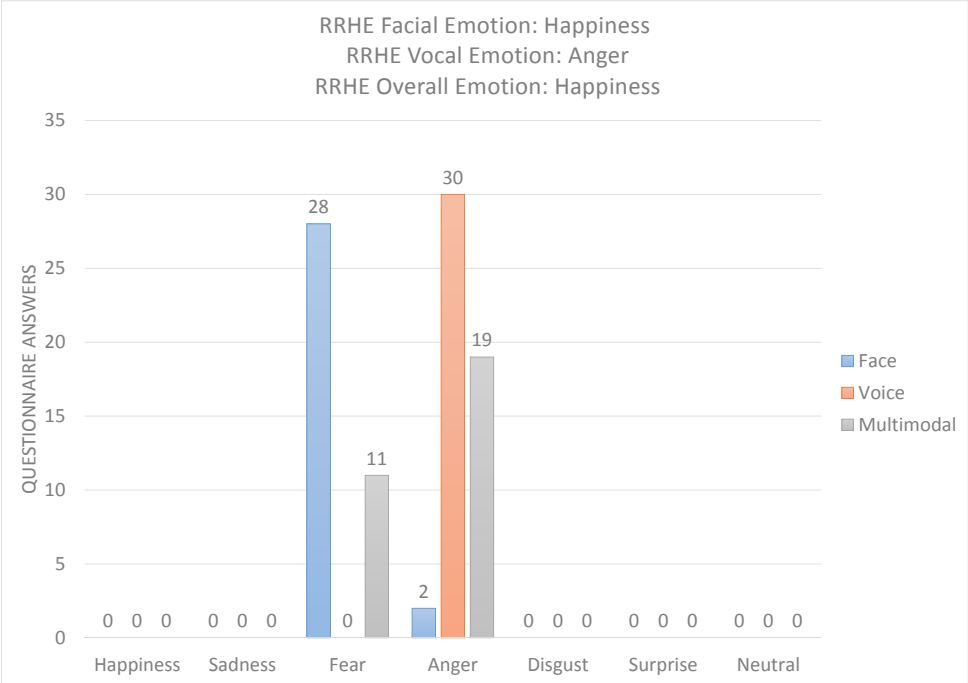

Figure 5.17: Session Questionnaire - Question 13 results.

In this video the respondents and RRHE clearly diverged in their classifications. Once again, the respondents gave more weight to the vocal expression.

14. What emotions do you recognize on the fourteenth segment?

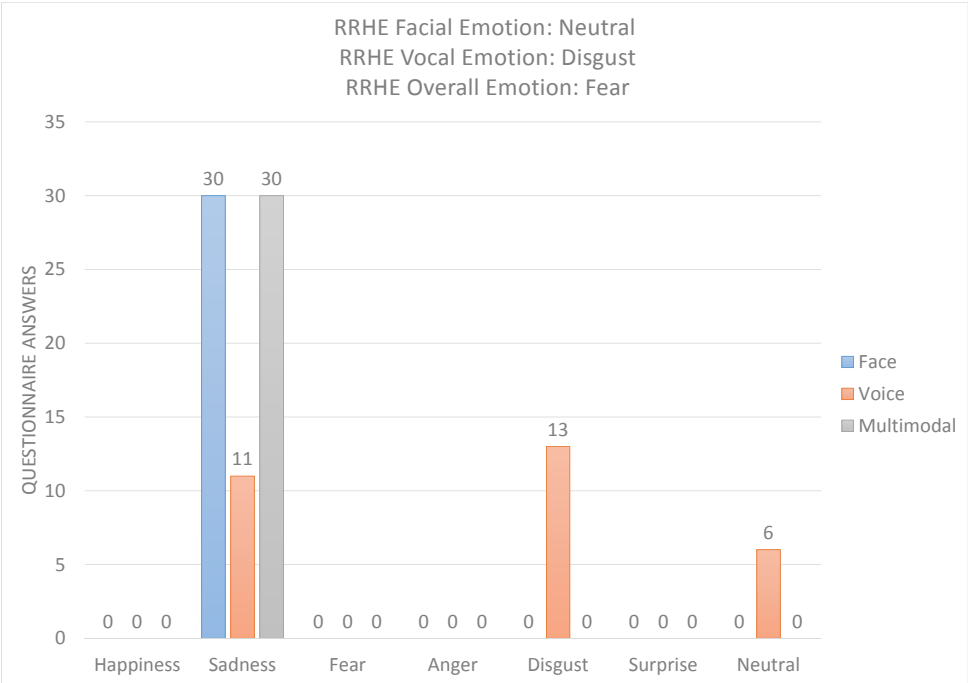

Figure 5.18: Session Questionnaire - Question 14 results.

In this video we can see that although RRHE detected a neutral facial expression and disgust in the vocal expression, the overall result was Fear. This happens because when taking the final

verdict, RRHE takes into account the weights of all possible expressions.

15. What emotions do you recognize on the fifteenth segment?

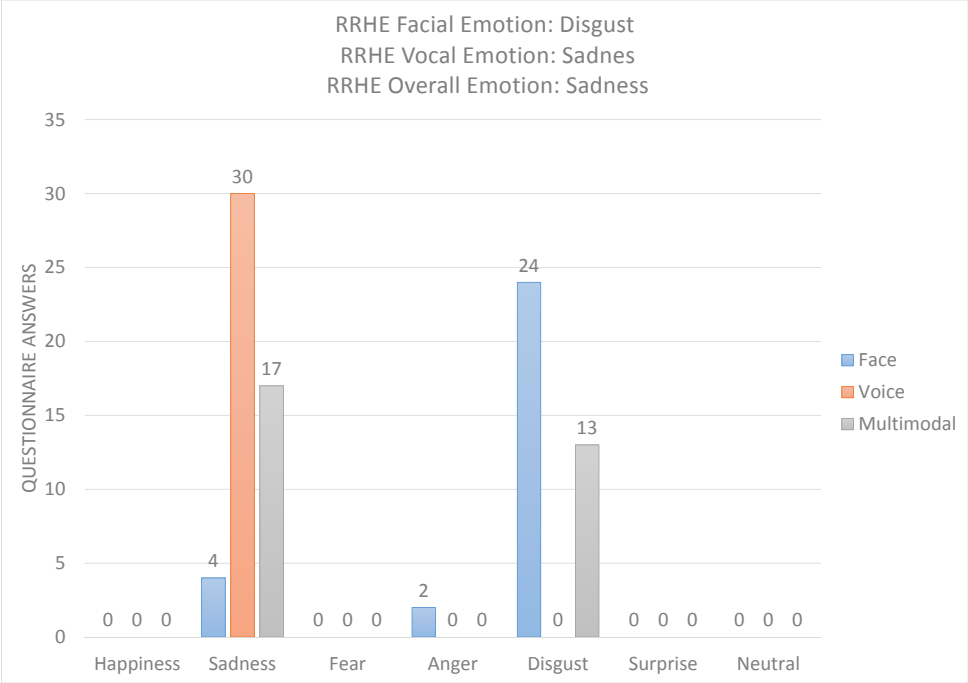

Figure 5.19: Session Questionnaire - Question 15 results.

Respondents and RRHE agreed on the classification of this video, both giving a bigger weight to the voice.

16. What emotions do you recognize on the sixteenth segment?

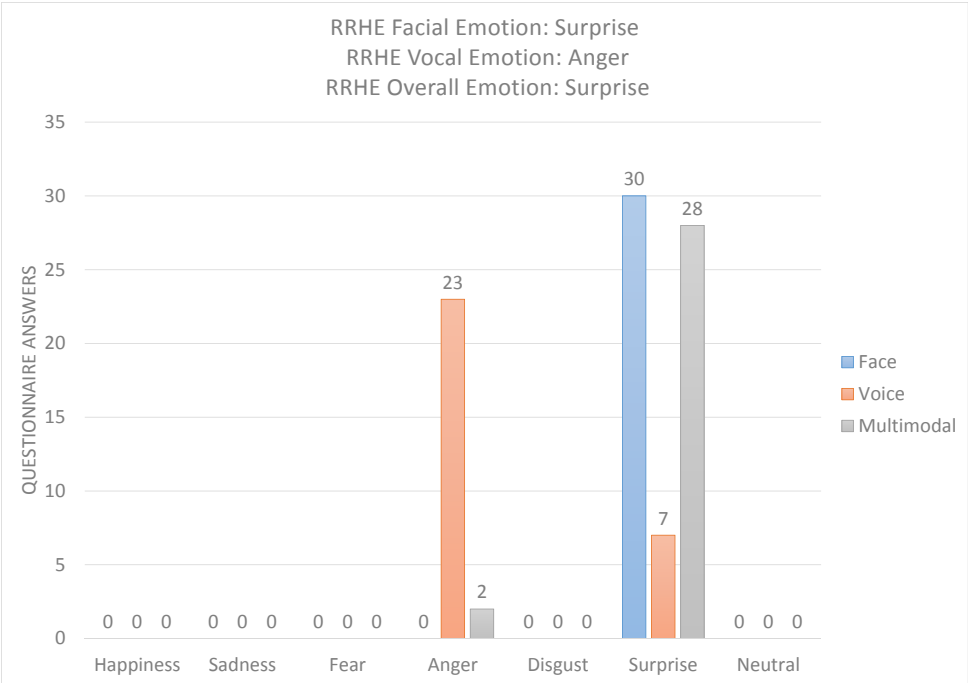

Figure 5.20: Session Questionnaire - Question 16 results.

Another irony detected by RRHE as well as by the respondents. Although the voice showed anger for both, the face was very expressive and as such, decisive in the final decision.

17. What emotions do you recognize on the seventeenth segment?

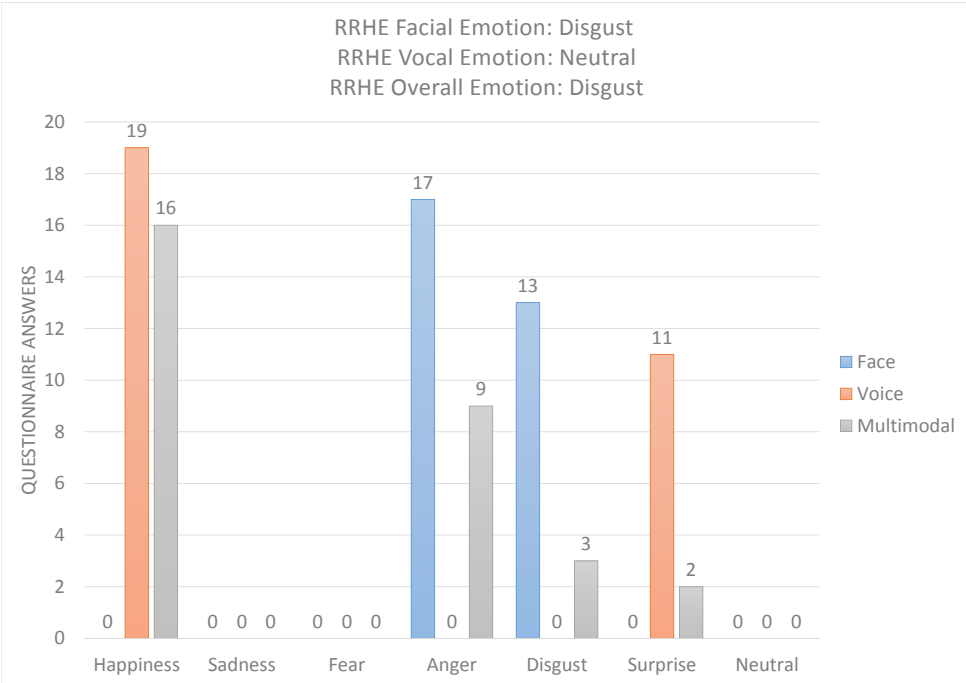

Figure 5.21: Session Questionnaire - Question 17 results.

RRHE could not detect the irony in this video. Respondents also found it hard to decide, leading to diverse answers. Most respondents found the face to be the most expressive element and classified the video as Happiness.

18. What emotions do you recognize on the eighteenth segment?

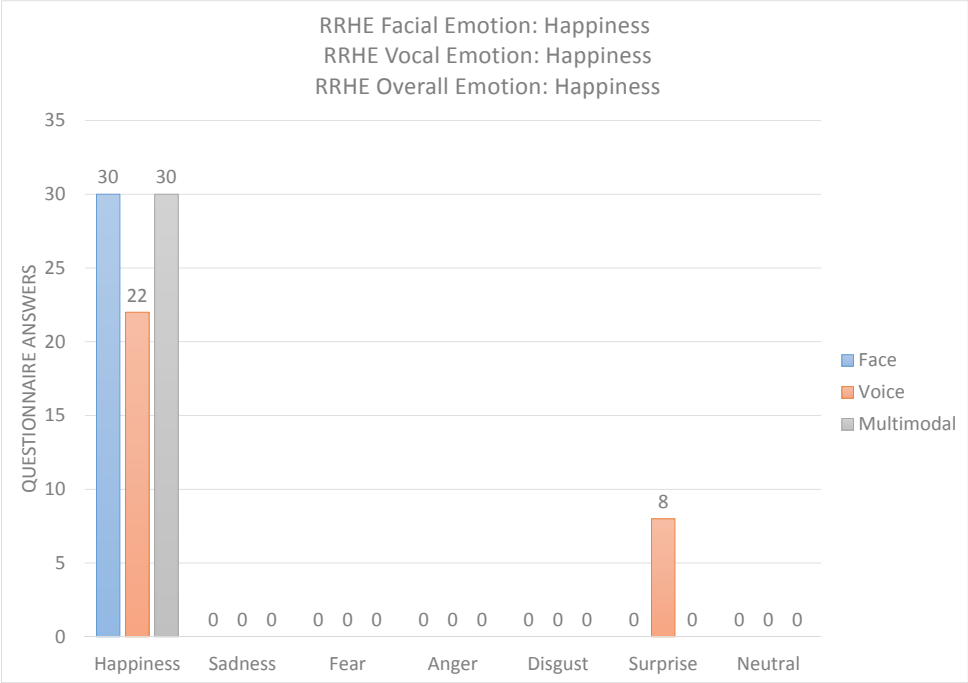

Figure 5.22: Session Questionnaire - Question 18 results.

This video was clear and RRHE and the respondents had the same classification.

19. What emotions do you recognize on the nineteenth segment?

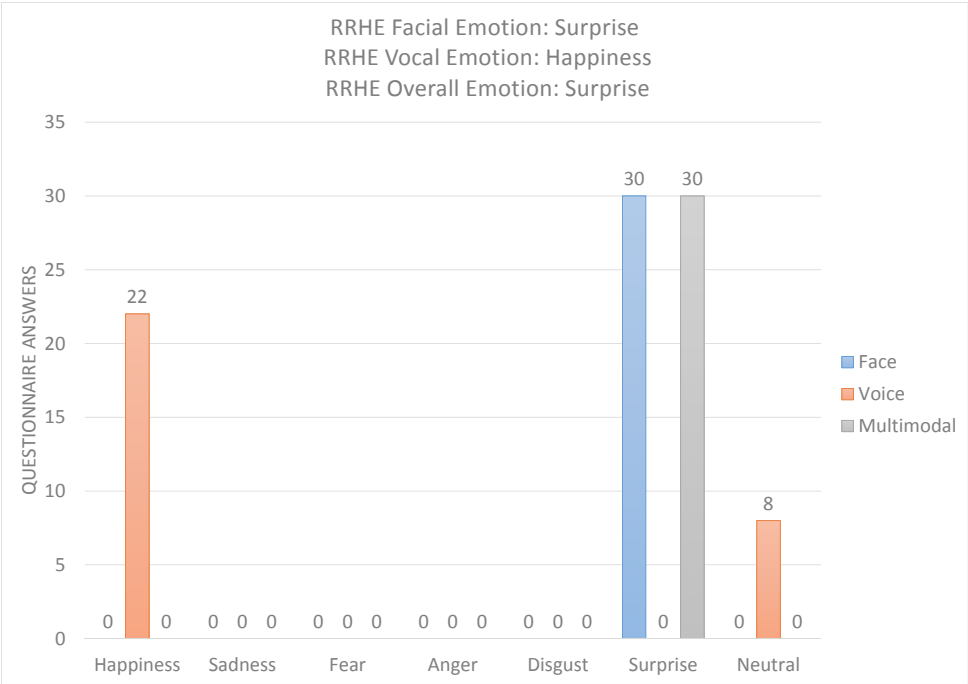

Figure 5.23: Session Questionnaire - Question 19 results.

This video did not show an irony, but the combination of the two expressions was very hard to understand. The facial expression was once again predominant in the video classification.

20. What emotions do you recognize on the twentieth segment?

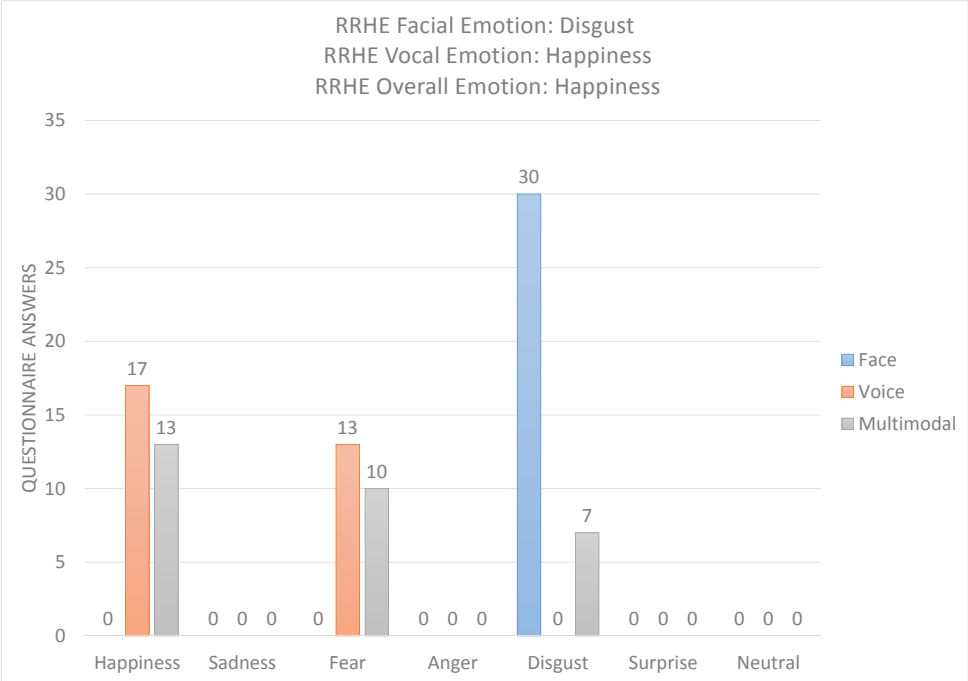

Figure 5.24: Session Questionnaire - Question 20 results.

Another video with an irony, detected both by RRHE and by the respondents. The respondents separately identified the facial and vocal expressions but were influenced but the facial expression when classifying the vocal expression.

21. What emotions do you recognize on the twenty-first segment?

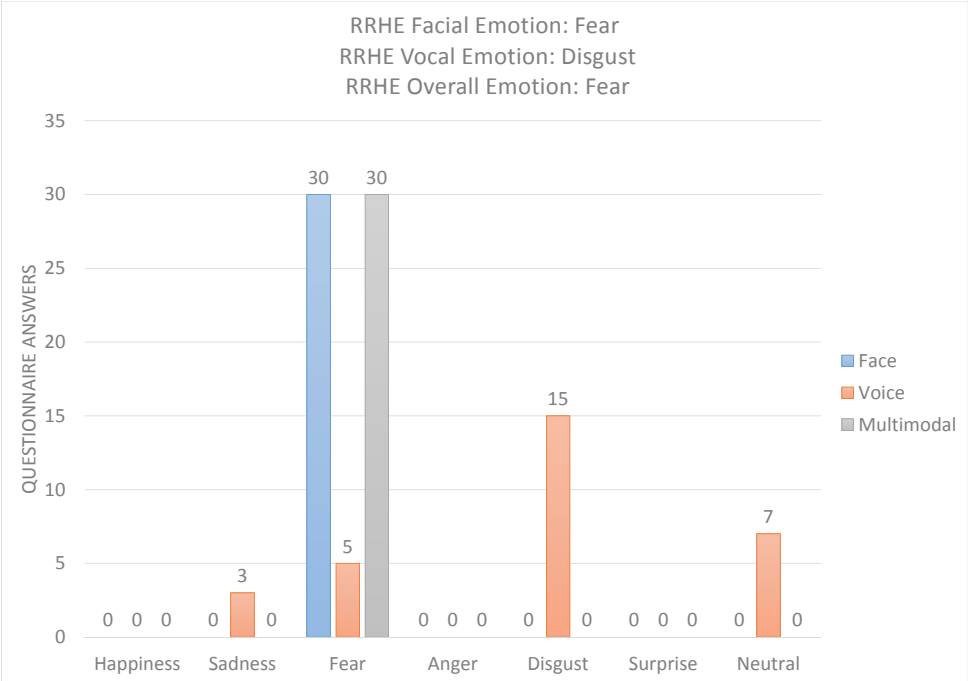

Figure 5.25: Session Questionnaire - Question 21 results.

The fear emotion was very clear and both the respondents and RRHE easily identified it.

22. What emotions do you recognize on the twenty-second segment?

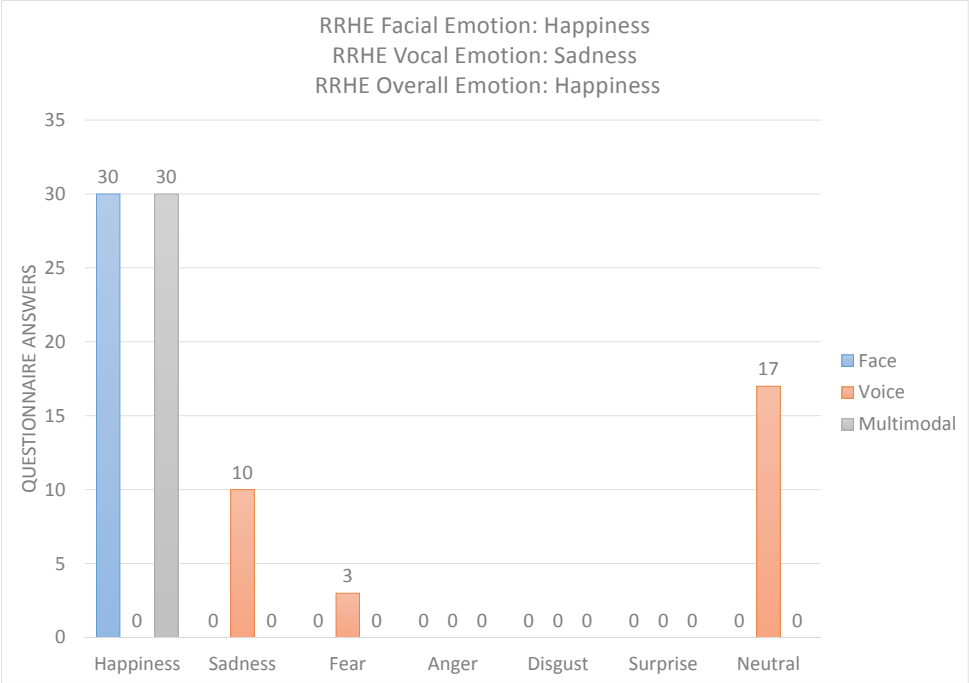

Figure 5.26: Session Questionnaire - Question 22 results.

Once again, this video showed an irony with Happiness and Sadness. Although some respondents failed to classify the vocal expression due to the influence of the facial expression, most correctly identified both expressions. The same happened with RRHE, which once again showed that the decision is not based always on the same component.

23. What emotions do you recognize on the twenty-third segment?

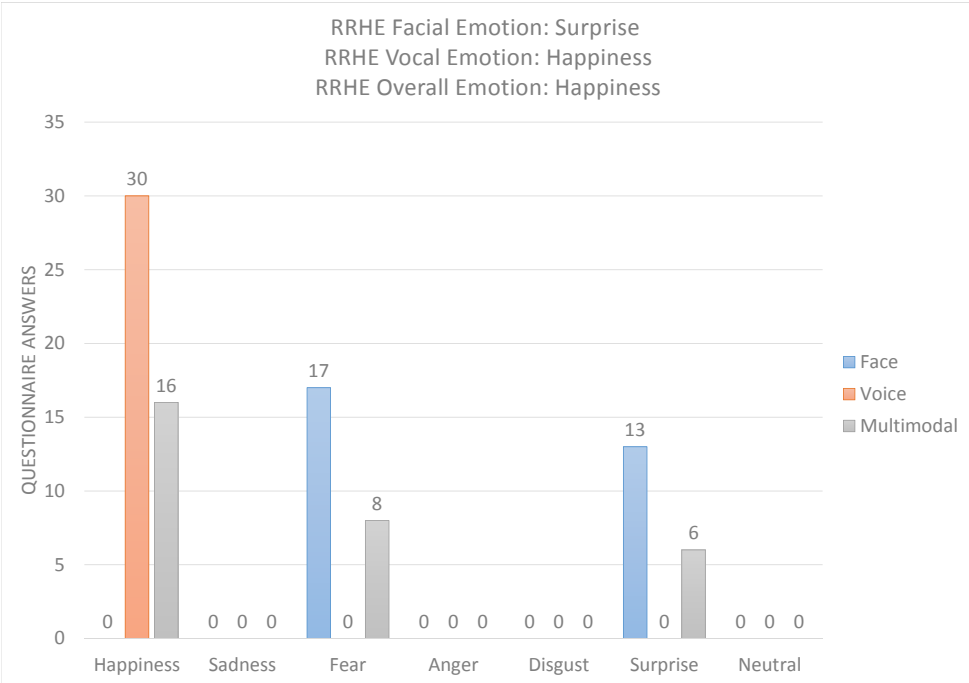

Figure 5.27: Session Questionnaire - Question 23 results.

The respondents were undecided between Fear and Surprise. These emotions are hard to distinguish without more information. It this case, the vocal expression had a bigger impact and the respondents used it for the final decision, just like RRHE.

24. What emotions do you recognize on the twenty-fourth segment?

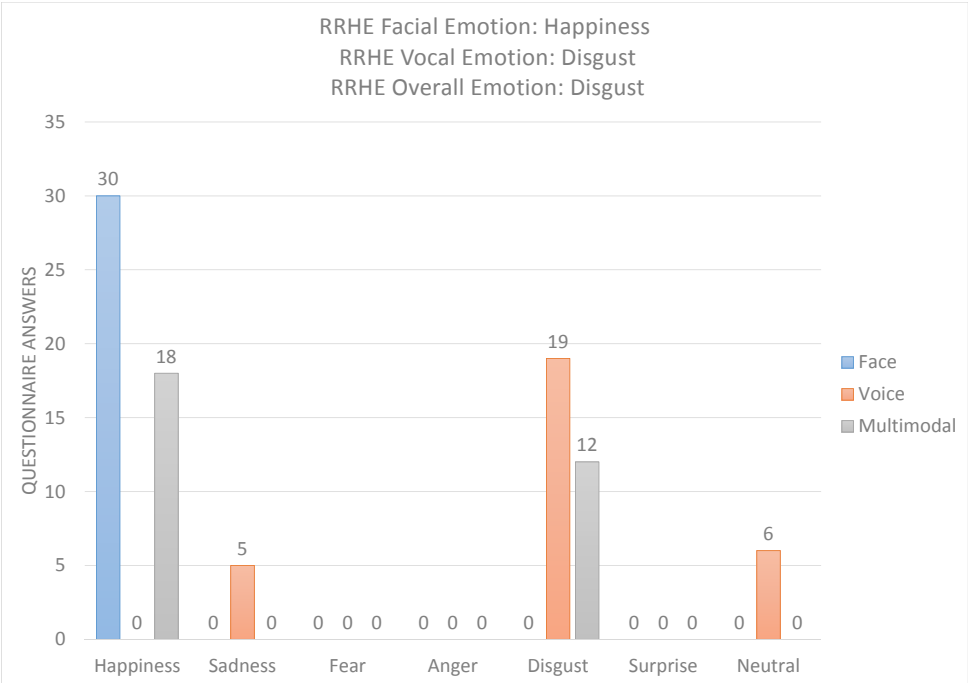

Figure 5.28: Session Questionnaire - Question 24 results.

This video contained another irony and RRHE and the respondents disagreed on the classification. Although the facial and vocal expression individual classifications were the same, RRHE gave more weight to the vocal expression while the majority of the respondents chose the facial expression.

25. What emotions do you recognize on the twenty-fifth segment?

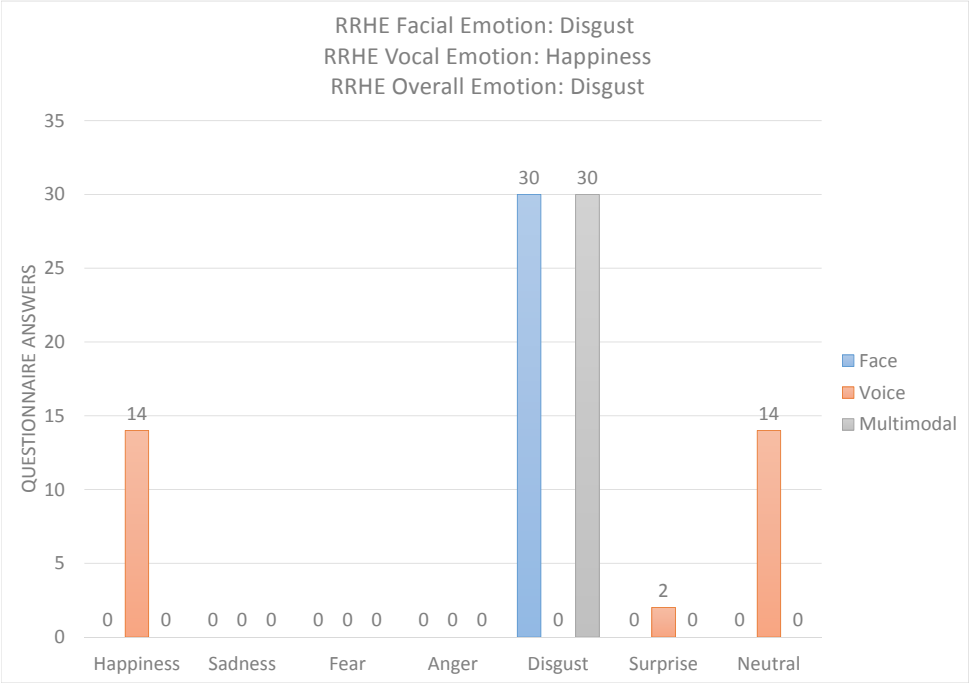

Figure 5.29: Session Questionnaire - Question 25 results.

Once again, this video showed an irony with RRHE and the respondents choosing the vocal over the facial expression and classifying the video as Disgust.

26. What emotions do you recognize on the twenty-sixth segment?

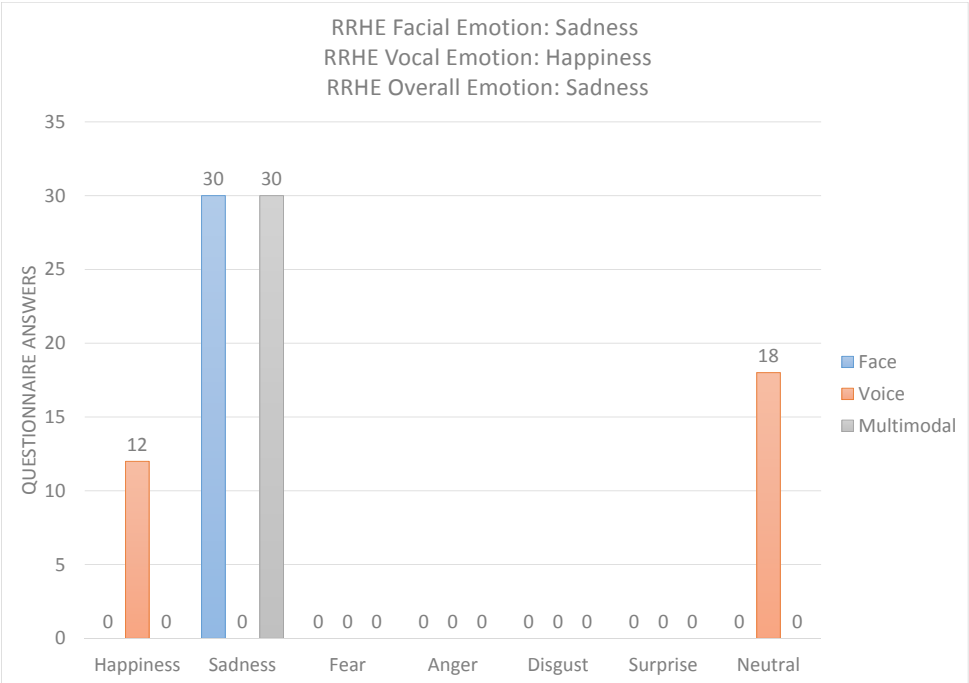

Figure 5.30: Session Questionnaire - Question 26 results.

This video showed one more time the tendency of the respondents to relate the voice with the face expression. Despite the voice in this video clearly showing Happiness, the majority of the

respondents classified it as neutral, influenced by the vocal expression. The facial expression was still very clear and the respondents all agreed on the classification of this video.

27. What emotions do you recognize on the twenty-seventh segment?

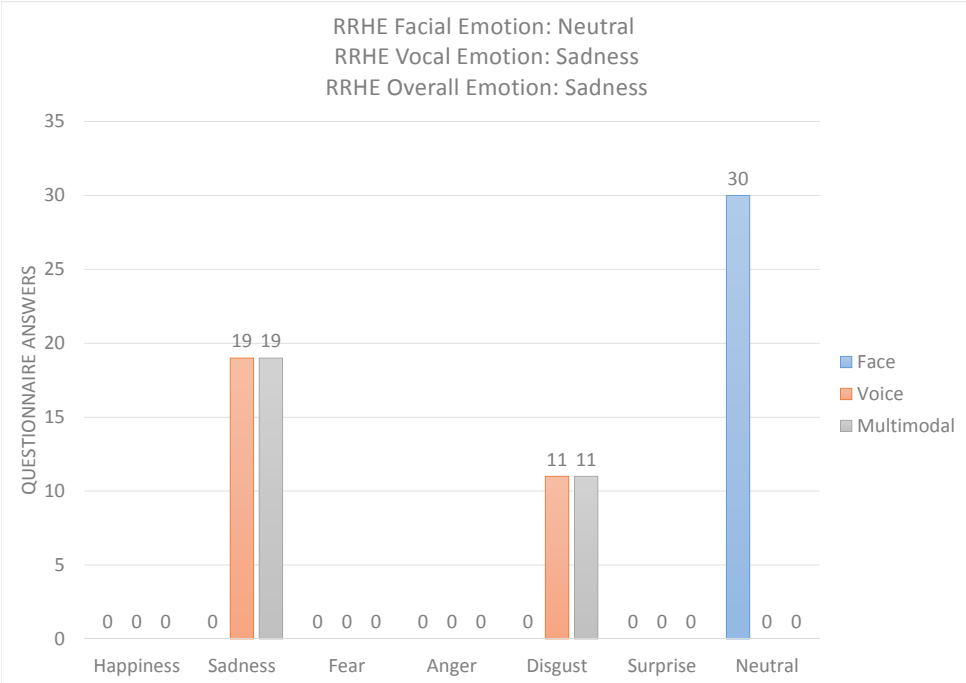

Figure 5.31: Session Questionnaire - Question 27 results.

This video shows a combination of emotions that is hard to comprehend. Respondents hesitated between Sadness and Disgust, which is understandable given the similarities between the two. The facial expression was clearly neutral and the respondents classified the video according to the vocal expression.

28. What emotions do you recognize on the twenty-eighth segment?

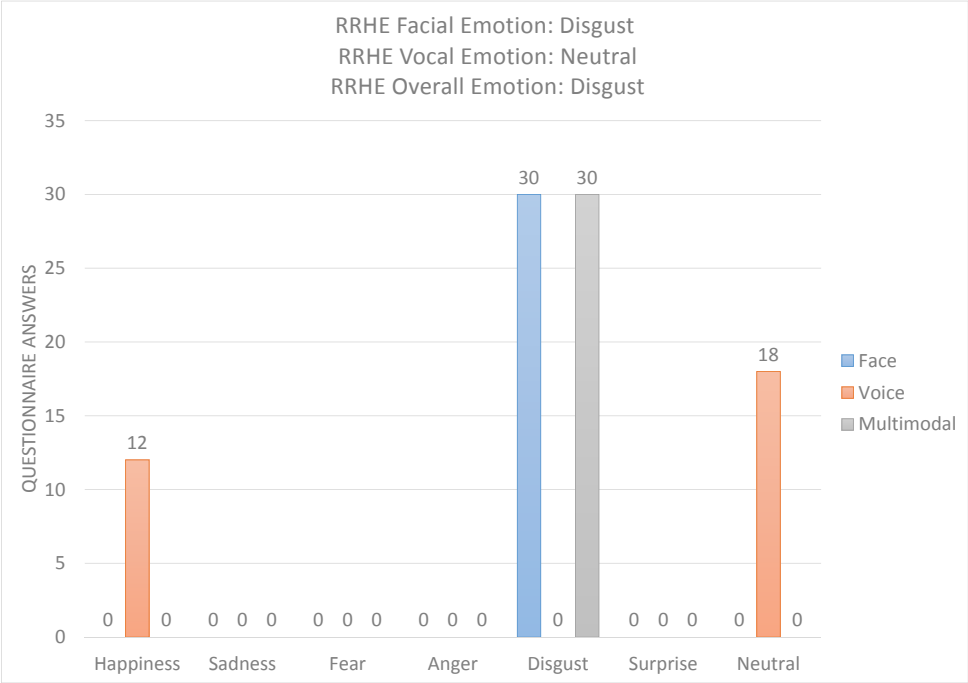

Figure 5.32: Session Questionnaire - Question 28 results.

This is another case that shows the influence of the facial expression over the vocal expression classification. Despite the vocal expression being clearly happiness, most respondents classified it as neutral, due to the negativity of the facial expression. In the end, the very expressive face classification was chosen as the video classification by RRHE and the respondents.

29. What emotions do you recognize on the twenty-ninth segment?

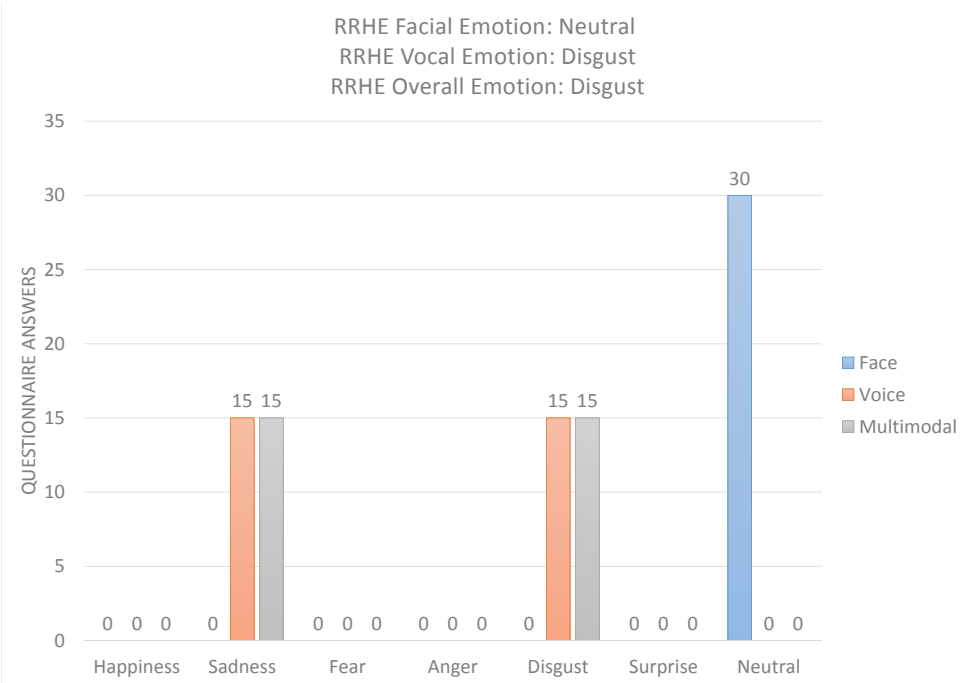

Figure 5.33: Session Questionnaire - Question 29 results.

The face shown in this video is clearly neutral and the respondents had no doubt about that but the voice classification was not so easy. In fact, this was the video where the respondents had more trouble, in part due to the similarity between the emotions expressed, but the vocal emotion dictated the final verdict.

30. What emotions do you recognize on the thirtieth segment?

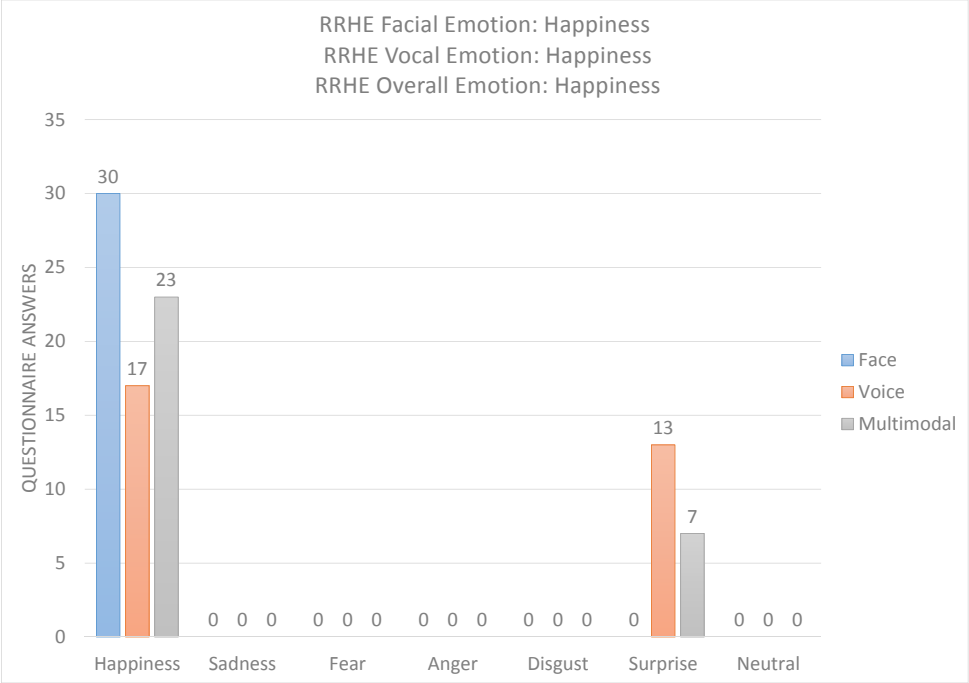

Figure 5.34: Session Questionnaire - Question 30 results.

Finally we have a video where the emotion is clearly happiness. Although some of the respondents considered the vocal emotion to be surprise, the majority agreed with RRHE and classified the video with the most evident emotion.

Summary

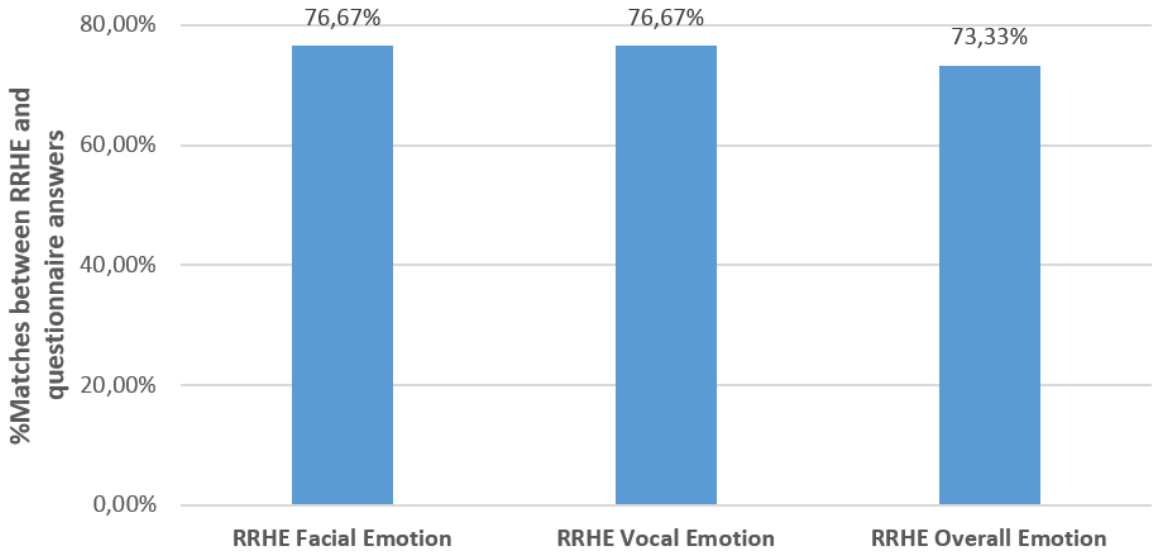

Figure 5.35: Total percentage of matches between RRHE and questionnaire results.

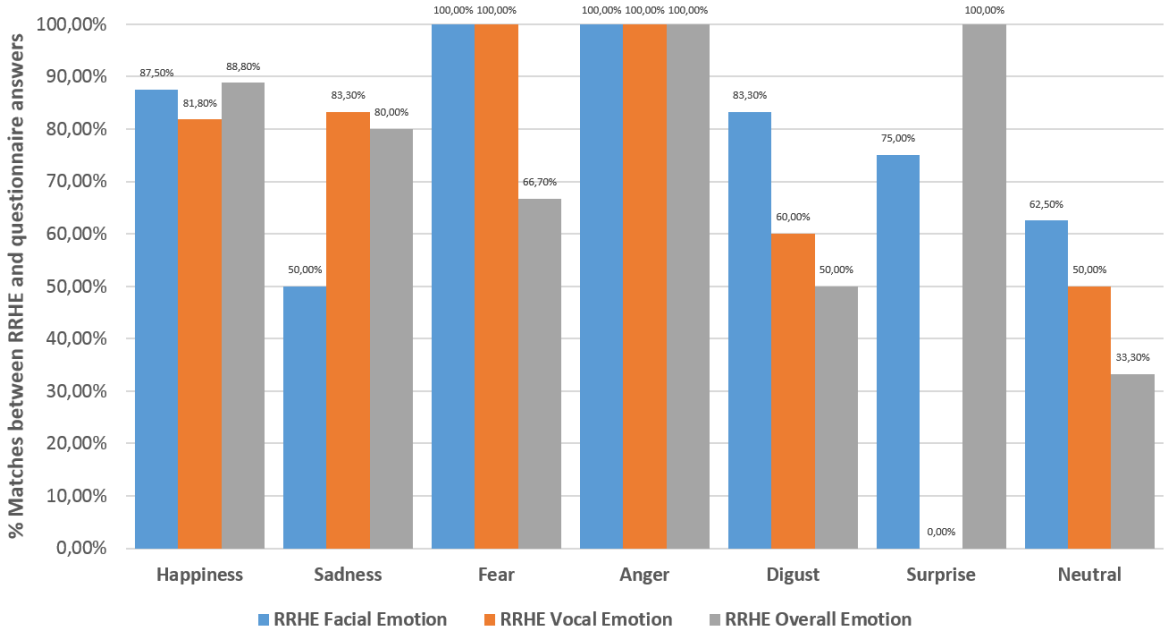

Figure 5.36: Percentage of matches, by emotion, between RRHE and questionnaire results.

Figure 5.35 shows a global view over the results obtained in the questionnaire, comparing them with RRHE results. The percentages show the proportion between the number of correct guesses of RRHE and the respondents. Curiously, the results reveal that the percentage of matches obtained by the facial emotion classifier is exactly the same as the vocal classifier. This data shows that both classifiers have the same level of accuracy which lends credibility to both algorithms. The overall classification has in some cases a lower match percentage, which can be explained by the difference in criteria of RRHE and the respondents when choosing the final classification. Nevertheless, the percentage of matches reveals that most of the times RRHE is correct according to the classification of the respondents. Given

the high number of ironies shown in the video set, we can say that RRHE was able to correctly identify them, assigning the same classification as the respondents in the majority of the videos. This constitutes a very promising result for future improvements on irony recognition systems.

Figure 5.36 presents the results regarding the different emotional states recognized by RRHE. It is important to take into consideration that we only have a sample of 30 videos, which explains scenarios such as the 0% in the voice classifier for the emotional state of Surprise that was not classified in any video by RRHE or the respondents. It is worthy to notice that, for the emotional states with the higher correct classification rate, the match percentage between the different classifiers is very similar and shows considerable values, reaching 100% on 4 occasions. Another relevant fact is that the facial emotion classifier shows a higher accuracy when compared to the vocal classifier. The facial emotions are usually more expressive and thus having a higher impact in the emotional analysis. This was one of the motives that lead us to give a bigger weight to this classifier. Finally, looking at the overall emotion values, it is important to note the high accuracy in certain emotions. The emotions more easily detected by RRHE are clear: Happiness, sadness, anger and surprise. We believe that these results are directly connected to the drastic facial and vocal changes that these emotional states cause in the human body.

5.3 Experiences with Kinect V2

We decided to test the new emotion recognition feature included in Kinect V2 SDK so that we could compare with the results obtained by RRHE. Since the Kinect only recognizes emotions from facial expressions, we could only compare the results with our facial emotion extraction algorithm.

Figure 5.37 shows the result of the implementation that was made to test Kinect V2 – based on the samples made available by Microsoft's SDK. As shown in the image, the outputs of Kinect are: Happy, Fear, Wearing Glasses, Left Eye Closed, Right Eye Closed, Surprise, Mouth Moved and Looking Away

We could only compare 3 emotions with those existing in our project – Happy, Fear and Surprise – which does not make this experiment very conclusive. For this comparative study we decided to use the Kinect camera as a capture source, since this camera is required to provide the inputs to the Kinect SDK. Afterwards we have used the images captured by Kinect in RRHE to obtain its classification. Table 5.13 shows the confusion matrix that resulted from the experiment – for the 3 common emotions – in a total of 60 frames.

We can conclude, according to this experiment's evidence, that in 38.33% of the cases the Kinect and RRHE are in accordance to each-other, returning the same classification. However, the Kinect does not present the neutral state, for each most emotions are mapped on RRHE. This may correspond to smaller degrees of happiness, fear or surprise that are very close to neutral. As such, RRHE will map them as neutral, while the Kinect as to map them into one of the other 3 available emotional states. Although in a smaller extent, there are anyway some significant classification differences remaining mapping Kinect emotions to other RRHE different emotional states.

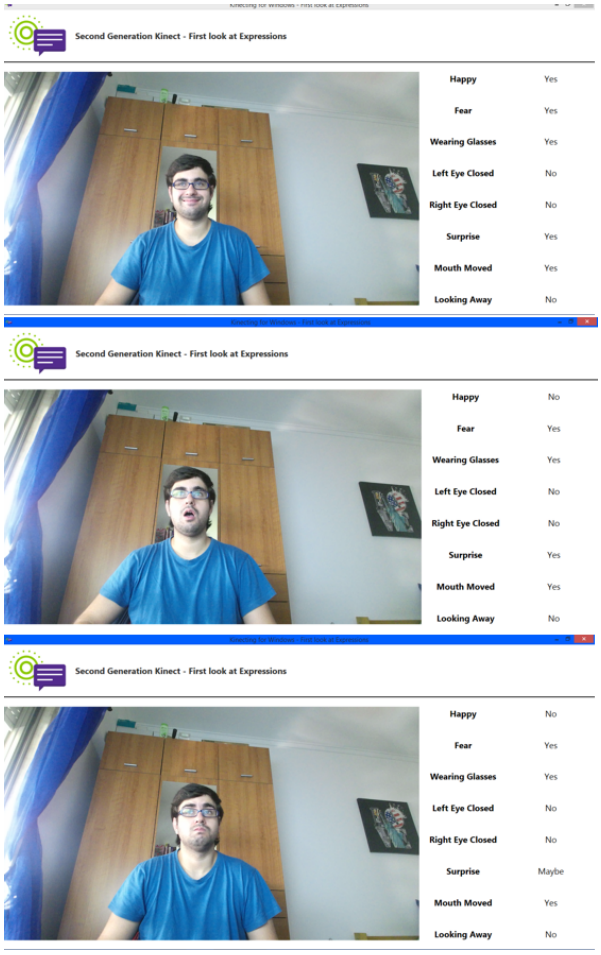

Figure 5.37: Kinect V2 Implementation

	Kinect - Happiness	**Kinect - Fear**	**Kinect - Surprise**
RRHE – Happiness	7	0	0
RRHE – Fear	0	4	2
RRHE – Surprise	3	0	12
RRHE – Disgust	0	0	0
RRHE – Anger	0	3	3
RRHE – Sadness	0	0	0
RRHE – Neutral	10	13	3

Table 5.13: Confusion matrix for Kinect V2 experiment.

Chapter 6

Conclusions

We hypothesized that the application of multimodal approaches for emotion detection could result in classification improvements, while enabling the detection of ironies, hardly possible using individual sensing modalities. This thesis also aimed to exploit the remote replication of such emotions in robots, virtual agents and social webs.

We found more interesting to develop an actuator simulator that can demonstrate the application of this technology to different areas. The commitment to have a ubiquitous system forced the design of an architecture that supports it. Another big commitment in this project was to have a client with a low memory footprint and low CPU usage so that it can be executed in any mobile device. Hence we decided to have only in the client side the capture of images and sound. The data is sent to the server for its immediate classification, corresponding to the heavier processing.

Another problem we faced was the definition of the algorithm for human emotion extraction. After investigation and analysis of the State of the Art algorithms, we concluded that single-modal techniques are more used and explored but still show several deficiencies. A clear example where single-modal algorithms are still lacking is in irony detection. We concluded that it would be preferable to use a multimodal algorithm for extraction of emotions. Among the available options, based on the objective of this work, we decided to use facial and voice expressions as an information source by implementing a function that merges both sources.

With this work we developed a client application, highly portable for any mobile device capable of image and sound capture for later transmission to the central server. The central server was developed on a multi-thread architecture for bigger scalability.

The experiments [Chapter 5] show the performance of the solution both in terms of accuracy and processing speed. As expected, the tests made with data extracted from the database used to train the classifiers show good results whenever the samples are known to the classifiers. On the other hand, when using a different database – not the one used for classifier training –, the results show that the system reacted well to inputs captured in controlled environments, even without having been used in classifier training. Relative to the results of the questionnaires we conclude that, in most cases, the classification of the respondents matches that of RRHE. Finally, we compared RRHE with Kinect

2, using the emotion detection features provided by Microsoft's SDK. The results are not relevant due to the big difference between the available labels in both systems, although both systems had similar classification results in the matching labels.

6.1 Contributions

We have shown that it is possible to replicate human emotions remotely with good results, both at the level of the correct classifications and the performance of the system. With this solution we replicate human emotions in approximately real-time, which is a requisite of most projects nowadays.

RRHE can be integrated in multiple use-cases, for example robots, call centers and conference systems. In this kind of systems one of the requisites is the small delay between data extraction and emotion representation, which we managed to implement with very good results.

Another contribution of this thesis is the fusion between two algorithms of emotion recognition, facial expressions and voice. Albeit not being able yet to achieve perfect results, this algorithm is a demonstration that it is possible to do it and solve deficiencies in single-modal implementations, as is the case of ironies.

6.2 Future Work

Concerning security, the overall solution may be both subject to in-place terminal security attacks, as well as networked attacks. Although security is outside this thesis scope, a commercial implementation should in the future implement the protection mechanisms for the solution vulnerabilities, such as authentication and data encryption.

Regarding facial and vocal emotion extraction algorithms, we suggest the implementation of a gender-dependent algorithm with different classifiers depending on the user sex. We also suggest to investigate different algorithms for both facial and vocal classifiers - for example, neural networks, HMM and AdaBoost Classifiers. We also suggest, as a potential improvement factor, to train both classifiers with multiple databases, containing data recorded with people of different races and different background conditions.

The algorithm to merge both emotion extraction techniques can be significantly improved. Instead of using weights for each of the techniques, some intelligence can be added. Instead of using only the confidence level of each classifier, other algorithm execution parameters can be used such as pitch, energy and some FACS extracted during facial analysis to feed a classifier trained with this kind of data.

Finally, the way the client and actuator functionality are provided are not perfect for eventual integration in other projects. Although an implementation detail, the most correct approach would be the development of an SDK without UI or logical process binding.

Bibliography

[1] Waleed Abdulla and Nikola Kasabov. Improving speech recognition performance through gender separation, 2001.

[2] T. Ahmad, H. Akhtar, A. Chopra, and M.W. Akhtar. Satire detection from web documents using machine learning methods. In *Soft Computing and Machine Intelligence (ISCMI), 2014 International Conference on*, pages 102–105, Sept 2014. doi: 10.1109/ISCMI.2014.34.

[3] Michael Argyle. *Bodily Communication*. University paperbacks. Methuen, 1988. ISBN 9780416381504. URL `http://books.google.pt/books?id=crYOAAAAQAAJ`.

[4] Devi Arumugam and Sun Purushothaman. Emotion Classification Using Facial Expression. *International Journal of Advanced Computer Science and Applications*, 2, 2011.

[5] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Fully automatic facial action recognition in spontaneous behavior. In *in 7th International Conference on Automatic Face and Gesture Recognition*, pages 223–230, 2006.

[6] Michael Black and Yaser Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *In ICCV*, pages 374–381, 1995.

[7] Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W14/W14-2608`.

[8] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. *Analysis of emotion recognition using facial expressions, speech and multimodal information*. ACM Press, 2004.

[9] Tanja Bänziger and Klaus Scherer. The role of intonation in emotional expressions. *Speech Communication*, 46(3-4):252–267, 2005. URL `http://dblp.uni-trier.de/db/journals/speech/speech46.html#BanzigerS05`.

[10] Justine Cassell and Kristinn Thórisson. The power of a nod and a glance: Envelope vs. emotional

feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4-5):519–538, 1999. URL http://dblp.uni-trier.de/db/journals/aai/aai13.html#CassellT99.

[11] Ira Cohen, Nicu Sebe, Larry Chen, Ashutosh Garg, and Thomas Huang. Facial expression recognition from video sequences: Temporal and static modelling. In *Computer Vision and Image Understanding*, pages 160–187, 2003.

[12] Zoran Duric, Wayne Gray, Ric Heishman, Fayin Li, Azriel Rosenfeld, Michael Schoelles, Christian Schunn, and Harry Wechsler. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. In *PROC. OF THE IEEE*, pages 1272–1289, 2002.

[13] Paul Ekman and Wallace Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971. URL http://psycnet.apa.org/journals/psp/17/2/124/.

[14] Paul Ekman and Wallace Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Prentice-Hall, Oxford, 1975.

[15] Paul Ekman and Wallace Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1978.

[16] Elena Filatova. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *LREC*, pages 392–398. European Language Resources Association (ELRA), 2012. ISBN 978-2-9517408-7-7. URL http://dblp.uni-trier.de/db/conf/lrec/lrec2012.html#Filatova12.

[17] Nickolaos Fragopanagos and John Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389–405, 2005. URL http://dblp.uni-trier.de/db/journals/nn/nn18.html#FragopanagosT05.

[18] Alan Fridlund. *Human Facial Expression: An Evolutionary View*. Acad. Press, 1994. ISBN 9780122676307. URL http://books.google.pt/books?id=ofnaAAAAMAAJ.

[19] Vladimir Hozjan and Zdravko Kacic. Context-independent multilingual emotion recognition from speech signals. 6:311–320, 2003.

[20] Carroll Izard. *The face of emotion / Carroll E. Izard*. Appleton-Century-Crofts, New York :, 1971.

[21] Ashish Kapoor, Winslow Burleson, and Rosalind Picard. Automatic prediction of frustration. *Int. J. Hum.-Comput. Stud.*, 65(8):724–736, August 2007. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2007.02.003. URL http://dx.doi.org/10.1016/j.ijhcs.2007.02.003.

[22] Theodoros Kostoulas and Nikos Fakotakis. A speaker dependent emotion recognition framework. In *Proc. 5th International Symposium, Communication Systems, Networks and Digital Signal Processing(CSNDSP)*, pages 305–309, 2006.

[23] Jun Lang. The emotion probe: Studies of motivation and attention. *American psychologist*, 50: 372–372, 1995.

[24] Stan Li and Anil Jain. *Handbook of Face Recognition*. Springer, 2005. ISBN 9780387405957. URL http://books.google.pt/books?id=amVDaTdgKYcC.

[25] Mark Liberman, Kelly Davis, Murray Grossman, Nii Martey, and John Bell. Emotional Prosody Speech and Transcripts. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2002S28, 2002. [Online; accessed 19-July-2015].

[26] Christine Lisetti and Fatma Nasoz. Maui: A multimodal affective user interface. In *Proceedings of the Tenth ACM International Conference on Multimedia*, MULTIMEDIA '02, pages 161–170, New York, NY, USA, 2002. ACM. ISBN 1-58113-620-X. doi: 10.1145/641007.641038. URL http://doi.acm.org/10.1145/641007.641038.

[27] Gwen Littlewort, Ian Fasel, Marian Stewart Bartlett, and Javier Movellan. Fully automatic coding of basic expressions from video. Technical report, Tech. rep.(2002) U of Calif., S.Diego, INC MPLab, 2002.

[28] Gwen Littlewort, Marian Stewart Bartlett, Ian R. Fasel, Joshua Susskind, and Javier R. Movellan. Dynamics of facial expression extracted automatically from video. *Image Vision Comput.*, 24(6):615–625, 2006. URL http://dblp.uni-trier.de/db/journals/ivc/ivc24.html#LittlewortBFSM06.

[29] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression.

[30] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, FG '98, pages 200–205, Washington, DC, USA, 1998. IEEE Computer Society. ISBN 0-8186-8344-9. URL http://dl.acm.org/citation.cfm?id=520809.796143.

[31] Ludo Maat and Maja Pantic. Gaze-x: Adaptive affective multimodal interface for single-user office scenarios. In *Proceedings of the 8th International Conference on Multimodal Interfaces*, pages 171–178, New York, NY, USA, 2006. ACM. URL http://doc.utwente.nl/62064/.

[32] Aleix Martinez. Face image retrieval using hmms, 1999.

[33] Keqji Mase. Recognition of facial expression from optical flow. pages 3474–3483, 1991.

[34] Albert Mehrabian. *Communication without words*, pages 51–52. 2 edition, 1968.

[35] Donn Morrison, Ruili Wang, and Liyanage De Silva. Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2):98–112, 2007. URL http://dblp.uni-trier.de/db/journals/speech/speech49.html#MorrisonWS07.

[36] Clifford Nass and Scott Brave. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. MIT Press, Cambridge, MA, 2005. ISBN 978-0-262-64065-7.

[37] Nuria Oliver, Alex Pentland, and François Bérard. Lafter: Lips and face real time tracker. pages 123–129, 1997.

[38] Andrew Ortony and Terence Turner. What's basic about basic emotions? *Psychological Review*, 97(3):315–331, 1990.

[39] Takahiro Otsuka and Juan Ohya. Recognizing multiple persons' facial expressions using hmm based on automatic extraction of significant frames from image sequences. In *In Proc. Int. Conf. on Image Processing (ICIP-97*, pages 546–549, 1997.

[40] Takahiro Otsuka and Juan Ohya. A study of transformation of facial expressions based on expression recognition from temporal image sequences. In *Technical report, Institute of Electronic, Information, and Communications Engineers (IEICE)*, 1997.

[41] Pritam Pal. *Emotion Detection from Infant Facial Expressions and Cries*. Temple University, 2006. URL `http://books.google.pt/books?id=GrEGtwAACAAJ`.

[42] Shrikanth Narayanan Pantelis Georgiou, Sungbok Lee. Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In *Proceedings of IEEE Multimedia Signal Processing Workshop, Chania, Greece*, 2007.

[43] Maja Pantic and Leon Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. In *Proceedings of the IEEE*, pages 1370–1390, 2003.

[44] Stavros Petridis and Maja Pantic. Audiovisual discrimination between laughter and speech. In *ICASSP*, pages 5117–5120. IEEE, 2008. ISBN 1-4244-1484-9. URL `http://dblp.uni-trier.de/db/conf/icassp/icassp2008.html#PetridisP08`.

[45] Valery Petrushin. Emotion recognition in speech signal: experimental study, development, and application. In *In: Proc. ICSLP 2000*, pages 222–225, 2000.

[46] Rosalind Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997. ISBN 0-262-16170-2.

[47] Robert Plutchik. The Nature of Emotions. *American Scientist*, 89(4):344+, 2001. ISSN 0003-0996. doi: 10.1511/2001.4.344. URL `http://dx.doi.org/10.1511/2001.4.344`.

[48] Mark Rosenblum, Yaser Yacoob, and Larry Davis. Human emotion recognition from motion using a radial basis function network architecture, 1994.

[49] Nicu Sebe, Michael Lew, Ira Cohen, Ashutosh Garg, and Thomas Huang. Emotion recognition using a cauchy naive bayes classifier. In *ICPR (1)*, pages 17–, 2002. URL `http://dblp.uni-trier.de/db/conf/icpr/icpr2002-1.html#SebeLCGH02`.

[50] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas Huang. Emotion recognition based on joint visual and audio cues. In *Proceedings of the 18th International Conference on Pattern Recognition - Volume 01*, ICPR '06, pages 1136–1139, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2521-0. doi: 10.1109/ICPR.2006.489. URL http://dx.doi.org/10.1109/ICPR.2006.489.

[51] Raunaq Shah and Michelle Hewlett. Emotion detection from speech. In *CS 229 Machine Learning Final Projects*, 2007.

[52] Ben Shneiderman and Catherine Plaisant. *Designing the User Interface: Strategies for Effective Human-Computer Interaction (4th Edition)*. Pearson Addison Wesley, 2004. ISBN 0321197860.

[53] Mingli Song, Jiajun Bu, Chun Chen, and Nan Li. Audio-visual based emotion recognition - a new approach. In *CVPR (2)*, pages 1020–1025, 2004. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2004-2.html#SongBCL04.

[54] Huang Sui, You Jianping, Zhang Hongxian, and Zhou Wei. Sentiment analysis of chinese micro-blog using semantic sentiment space model. In *Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on*, pages 1443–1447, Dec 2012. doi: 10.1109/ICCSNT.2012.6526192.

[55] Dov Te'eni, Jane Carey, and Ping Zhang. *Human Computer Interaction: Developing Effective Organizational Information Systems*. John Wiley & Sons, Hoboken, 2007. ISBN 0471677655.

[56] Michel Valstar, Ioannis Patras, and Maja Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *CVPR-V4HCI*, jun 2005. URL http://mmi.tudelft.nl/pub/michel/CVPR2005_ValstarPatrasPantic-final.pdf.

[57] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. pages 511–518, 2001.

[58] Thurid Vogt, Elisabeth André, and Nikolaus Bee. Emovoice – a framework for online recognition of emotions from voice. In *Proceedings of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems: Perception in Multimodal Dialogue Systems*, PIT '08, pages 188–199, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-69368-0. doi: 10.1007/978-3-540-69369-7_21. URL http://dx.doi.org/10.1007/978-3-540-69369-7_21.

[59] Yongjin Wang and Ling Guan. Recognizing human emotional state from audiovisual signals. *Trans. Multi.*, 10(4):659–668, June 2008. ISSN 1520-9210. doi: 10.1109/TMM.2008.921734. URL http://dx.doi.org/10.1109/TMM.2008.921734.

[60] Mark Weiser. Ubiquitous computing. *Computer*, 26(10):71–72, October 1993. ISSN 0018-9162. doi: 10.1109/2.237456. URL http://dx.doi.org/10.1109/2.237456.

[61] Yaser Yacoob and Larry Davis. Recognizing human facial expressions from long image sequences using optical flow. pages 636–642, 1996.

[62] N. Yang, R. Muraleedharan, J. Kohl, I. Demirkol, W. Heinzelman, and M. Sturge-Apple. Speech-based emotion classification using multiclass svm with hybrid kernel and thresholding fusion. In *SLT*, pages 455–460. IEEE, 2012. ISBN 978-1-4673-5125-6. URL `http://dblp.uni-trier.de/db/conf/slt/slt2012.html#YangMKDHS12`.

[63] Serdar Yildirim, Sungbok Lee, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Carlos Busso, and Shrikanth Narayanan. Study of acoustic correlates associated with emotional speech. In *148th ASA Meeting, San Diego, CA*, 2004.

[64] Zhihong Zeng, Jilin Tu, Brian Pianfetti, Ming Liu, Tong Zhang, ZhenQiu Zhang, Thomas Huang, and Stephen Levinson. Audio-visual affect recognition through multi-stream fused hmm for hci. In *CVPR (2)*, pages 967–972. IEEE Computer Society, 2005. ISBN 0-7695-2372-2. URL `http://dblp.uni-trier.de/db/conf/cvpr/cvpr2005-2.html#ZengTPLZZHL05`.

[65] Zhihong Zeng, Yuxiao Hu, Ming Liu, Yun Fu, and Thomas Huang. Training combination strategy of multi-stream fused hidden markov model for audio-visual affect recognition. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, MULTIMEDIA '06, pages 65–68, New York, NY, USA, 2006. ACM. ISBN 1-59593-447-2. doi: 10.1145/1180639.1180661. URL `http://doi.acm.org/10.1145/1180639.1180661`.

[66] Zhihong Zeng, Maja Pantic, Glenn Roisman, and Thomas Huang. A survey of affect recognition methods: Audio, visual and spontaneous expressions. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, ICMI '07, pages 126–133, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-817-6. doi: 10.1145/1322192.1322216. URL `http://doi.acm.org/10.1145/1322192.1322216`.

Appendix A

Pre-Session Questionnaire

Pre-Session Questionnaire

*Required

1. **What is your gender?** *
   Choose only one option.
   Mark only one oval.
   - Male
   - Female

2. **How old are you?** *
   Choose only one option.
   Mark only one oval.
   - Less than 20
   - Between 20 and 25 (including)
   - Between 26 and 30 (including)
   - Between 31 and 35 (including)
   - Between 36 and 40 (including)
   - Over 40

3. **What is your job area?** *
   Mark only one oval.
   - Student
   - Teaching
   - Engineering
   - Health
   - Agriculture
   - Other: _____

4. **Have you ever had any practice detecting emotions?** *
   Choose only one option.
   Mark only one oval.
   - Yes
   - No

Figure A.1: Pre-session questionnaire, page 1.

5. **Are you able to identify emotions looking to a face?** *
Choose only one option.
Mark only one oval.

- Yes, for sure
- Only the most basic ones
- I'm not sure
- No, I don't

6. **Are you able to identify emotions listening to a human speech?** *
Choose only one option.
Mark only one oval.

- Yes, for sure
- Only the most basic ones
- I'm not sure
- No, I don't

7. **Are you able to identify an irony when looking and listening to other people?** *
Choose only one option.
Mark only one oval.

- Yes, for sure
- I'm not sure
- No, I don't

8. **How do you think you can extract more emotional information?** *
Choose only one option.
Mark only one oval.

- Facial Expressions
- Emotional Speech
- Text analysis
- Biometrics data

9. **Classify your knowledge on each one of the following emotions: (1 - low knowledge; 5 - high knowledge)** *
Choose only one option per row.
Mark only one oval per row.

	1	2	3	4	5
Happiness					
Surprise					
Fear					
Disgust					
Anger					
Sadness					
Neutral					

Figure A.2: Pre-session questionnaire, page 2.

10. **When classifying an emotion what are the metrics with more impact on your choice?** *

Choose only one option.

Mark only one oval.

- Always use the same technic (same weight for all the metrics)
- According to the situation I measure each metric
- Always choose the most obvious one

Powered by Google Forms

Figure A.3: Pre-session questionnaire, page 3.