UNIVERSIDADE DA CORUÑA

Facultade de Informática
Departamento de Computación

PhD Thesis

Computer aided hearing assessment: towards an automated audiometric tool

Alba Fernández Arias
March 2015

PhD advisors:
Manuel F. González Penedo
Marcos Ortega Hortas

March 2015
Universidade da Coruña

Facultade de Informática
Campus de Elviña s/n
15071, A Coruña (Spain)

When you set out on your journey to Ithaca,
then pray that the road is long,
full of adventure, full of knowledge.
The Lestrygonians and the Cyclops,
the angry Poseidon... do not fear them:
You will never find such as these on your path,
if your thoughts remain lofty, if a fine
emotion touches your spirit and your body.
Ithaca, Constantine P. Cavafy (1911)

Acknowledgments

The journey has been long, and truly full of adventure and knowledge, and it would not have been possible without the support of many people I would like to thank. Over the last years I have realized how my life has directed me to this moment. Research has always been part of my life; I can not forget the Christmas when being a 9-year-old girl my grandpa gave me a microscope or when my parent introduced me to computers through the "attractive" MS-DOS or how I was always asking how things work... Observation and curiosity are the two things that start the scientific process, they both have accompanied me through years and I am sure they will never leave me.

Several are the reasons I want to express my gratitude to my parents: they raised me with love, taught me, provided me with unfailing support throughout my years of study and through the process of researching this thesis and they allowed me to be as ambitious as I wanted. Even if being an only child was great for a while, they gave me a little sister so I can learn to share, to be patient and to care about someone.

No one knows better than Pablo all the crazy ups and downs this journey has brought me over the past few years. During these years of study and research, the best outcome was finding the better person to share my life and all the journeys it bring us with. Only one more thing to say to you: I know you will!

I also would like to extend my gratitude to the many people who helped to bring this research project to fruition. First, I would like to thank my advisors. Manuel G. Penedo for providing me the opportunity of taking part of the VARPA group and for always being available to advise me. Marcos Ortega whose valuable guidance has contributed to the success of this research.

I would also like to thank the experts from the Unit Dual Sensory Loss of the Faculty of Optics and Optometry of the University of Santiago de Compostela (Spain) who proposed us to automate their manual processes: Luz Gigirey and Covadonga

Vázquez. Thank you for bringing me closer to the clinical world and sharing your valuable experience.

My sincere thanks to my lab mates for their moral support, cheerfulness and fun. You know how important a little of crazy is to survive a PhD.

Last but not least, thanks to all the friends and family who have walked with me the path to becoming a doctor.

This journey is worth the highs and lows! Promise!

Alba Fernández Arias

Abstract

Hearing loss is a partial or full decrease in the ability to detect or understand sounds which affects a wide range of population, and has a negative impact on their daily activities. Pure Tone Audiometry is the standard test for the evaluation of the hearing capacity. During the performance of this hearing assessment the audiologist also tries to identify patients with abnormally slow responsiveness by means of their response times to the perceived sounds. This identificacion is relevant since it could be a symptom of any medical condition that should be studied. The other main target is the evaluation of patients with cognitive decline or severe communication disorders, since when evaluating this specific group of patients it is not possible to maintain a normal question-answer interaction. In these cases the expert must focus his attention on the detection of unconscious gestural reactions to the sound.

The subjective involved in the interpretation of both aims may affect the classification, introduces imprecisions, limits the reproducibility and also a high degree of inter- and also intra- observer variability can be produced. In this manner, the development of a systematic, objective computerized method for the analysis and classification of response times and gestural reactions to the sound is thus highly desirable, allowing for homogeneous diagnosis and relieving the experts from this tedious task.

The proposal of this research is the design of an automatic system to assess the gestural reactions to the sound and the patient's response times by analyzing video sequences recorded during the performance of the audiometric evaluations. On the one hand, the response times are measured by detecting the auditory stimuli delivering and the patient's hand raising (which corresponds with a positive response). On the other hand, the gestural reactions to the sound are identifed by analyzing the eye movements using two different approximations. The different automated assessments proposed save time for experts, improve the precision and provide unbiased results which are not affected by subjective factors.

Resumen

La pérdida de audición consiste en una disminución parcial o total de la capacidad para percibir sonidos que afecta a un amplio rango de población y tiene un impacto negativo en sus actividades diarias. La audiometría tonal liminar es uno de los tests estándard para la evaluación de la capacidad auditiva. Durante la realización de esta evaluación el audiólogo trata paralelamente de identificar pacientes con tiempos de respuesta anormalmente lentos. Esta identificación es relevante pues podría tratarse de un síntoma asociado a alguna patología que debiera ser estudiada. El otro objetivo principal es la evaluación de pacientes con deterioro cognitivo o trastornos graves de comunicación, puesto que no es posible mantener una interacción típica de pregunta-respuesta cuando se evalúa su audición. En estos casos, el experto debe centrar su atención en la detección de reacciones gestuales espontáneas al sonido.

La subjetividad implicada en la interpretación de ambos objetivos puede afectar a la clasificación, introducir imprecisiones, limitar la reproducibilidad y también producir un alto grado de inter e intra varibialidad del observador. En este sentido, el desarrollo de un método automatizado, objetivo y sistemático para el análisis y clasificación de los tiempos de respuesta y de las reacciones gestuales al sonido es, por tanto, altamente conveniente, permitiendo un diagnóstico homogéneo y relevando a los expertos de esta tediosa tarea.

El propósito de esta investigación es el diseño de un sistema automático para la evaluación de las reacciones gestuales y los tiempos de respuesta a través del análisis de secuencias de vídeo grabadas durante el desarrollo de la prueba audiométrica. Por una parte, los tiempos de respuesta se miden detectando el envío de estímulos y la respuesta positiva del paciente (expresada levantando la mano). Por otra, las reacciones gestuales son identificadas analizando los movimientos de la mirada usando dos aproximaciones diferentes. Las diferentes propuestas automatizadas presentadas ahorran tiempo a los expertos, mejoran la precisión y proporcionan resultados objetivos que no se ven afectados por factores subjetivos.

Resumo

A perda de audición consiste nunha disminución parcial ou total da capacidade para percibir sons que afecta a un amplo rango da poboación e ten un impacto negativo nas súas actividades diarias. A audiometría tonal liminar é un dos tests estándard para a avaliación da capacidade auditiva. Durante a realización desta avaliación o audiólogo trata paralelamente de identificar pacientes con tempos de resposta anormalmente lentos. Esta identificación é relevante pois poderá tratarse dun síntoma asociado a algunha patoloxía que debese ser estudada. O outro obxectivo principal é a avaliación de pacientes con deterioro cognitivo ou trastornos graves de comunicación, posto que non é posible manter unha interacción típica de pregunta-resposta cando se avalía a súa audición. Nestes casos, o experto debe centrar a súa atención na detección de reaccións xestuais espontáneas ao son.

A subxectividade implicada na interpretación de ambos obxectivos pode afectar á clasificación, introducir imprecisións, limitar a reproducibilidade e tamén producir un alto grao de inter e intra varibialidade do observador. Neste sentido, o desenvolvemento dun método automatizado, obxectivo e sistemático para a análise e clasificación dos tempos de resposta e das reaccións xestuais ao son é, por tanto, altamente conveniente, permitindo unha diagnose homoxénea e relevando aos expertos desta tediosa tarea.

O propósito desta investigación é o diseño dun sistema automático para a avaliación das reaccións xestuais e os tempos de resposta a través da análise de secuencias de vídeo grabadas durante o desenvolvemento da proba audiométrica. Por unha parte, os tempos de resposta mídense detectando o envío de estímulos e a resposta positiva do paciente (expresada levantando a man). Por outra, as reaccións xestuais son identificadas analizando os movementos da mirada usando dúas aproximacións diferentes. As diferentes propostas automatizadas presentadas aforran tempo aos expertos, melloran a precisión e proporcionan resultados obxectivos que no se ven afectados por factores subxectivos.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Hearing is the sense that enables the sound to be perceived. It is a major function of the ear. Any reduction in the ability to perceive sounds results in hearing loss. Hearing loss is an invisible condition implying a sudden or gradual decrease of hearing. Although hearing loss is a global phenomenon and it extends to all age groups, there is a progressive loss of sensitivity to hear high frequencies with increasing age. The latest report on Aging by the General Foundation of the CSIC (IMSERSO, 2010) states that the most common and earlier disability is the decrease of the sensory abilities: hearing and vision. Furthermore, hearing loss is the third most prevalent chronic health condition facing older adults (Collins, 1997), and it is also one of the most widely under-treated conditions. Hearing impairment commonly implies problems to understand speech and to communicate, which results in a feeling of progressive confinement. Different studies (Davis, 1989) have demonstrated the considerable negative effects that untreated hearing loss may have on the physical, social, psychological and cognitive well-being of a person.

Besides, population aging is a demographic revolution affecting the entire world (IMSERSO, 2008; Kalache, Barreto, & Keller, 2005). This increase in longevity involves a parallel increase of the years lived with incapacity and invalidity. Related with hearing, the age distribution changes due to aging population have as consequence not only a higher prevalence of hearing problems, but also a greater severity of their effects. Age-related hearing loss, also called presbycusis, is characterized by elevated hearing thresholds, difficulties to understand speech in noisy and reverberant environments and interferences in the perception of rapid changes of speech. In turn, with age increases the possibility of emergence of neurodegenerative disorders and communication problems. This problematic also implies limitations on the evaluation of the hearing capacity that are going to be considered later.

The World Health Organization (WHO) has set the criteria for hearing loss above 25dB, i.e., a person with hearing thresholds above 25dB in both ears is not able to hear as well as someone with normal hearing (with hearing thresholds of 25dB or better), so it is said that he or she suffers from hearing loss. Hearing loss may be mild, moderate, severe or profound. It can affect to one or both ears. Hard of hearing refers to people with hearing loss ranging from mild to severe. Profound range corresponds with deafness and it implies very little or no hearing. Disabling hearing loss (which corresponds with the profound range) refers to hearing loss greater than 40dB for adults and a hearing loss greater that 30dB in children. People with hard of hearing need to be proper diagnosed so they can benefit from hearing aids, assistive listening devices or cochlear implants.

Over the 5% of the world's population has disabling hearing loss (328 million adults and 32 million children). Recent studies indicate that hearing problems are at an increasingly early age (Agrawal, Platz, & Niparko, 2008). Davis (1989), Director of the MRC Hearing & Communication Group at the University of Manchester, estimates that more than 700 million people all over the world will suffer hearing loss over 25dB for 2015, increasing this number to 900 million people worldwide for 2025.

Lack of hearing is one of the most frequent sensory deficits among elder population. It certainly extends to all age ranges, but it is among elder people where it has a higher incidence. According to the National Institute of Deafness and Other Communication Disorders (of Deafness & Disorders, 2009) about 2% of North American adults aged 45 to 54 have disabling hearing loss, the rate increases to 8.5% for adults aged 55 to 64, nearly 25% of those aged 65 to 74 and 50% of those who are 75 and older have disabling hearing loss. Meanwhile, the Australian Hearing Annual Report (Hearing, 2009) states that more than half of individuals aged between 60-70 years have some hearing deficit, increasing to 70% for people with 70 years old or more. In general terms, approximately one-third of people over 65 years of age are affected by disabling hearing loss. The prevalence in this age group is greatest in South Asia, Asia Pacific and sub-Saharan Africa. In general terms, population aging is a global reality (IMSERSO, 2010). Its onset is usually insidious but it gradually worsens.

Different studies have demonstrated the considerable negative effects that untreated hearing loss may have on the physical, social, psychological and cognitive well-being of a person (Davis, 1989; Ciorba, Bianchini, Pelucchi, & Pastore, 2012; Kochkin & Rogin, 2000). Impaired hearing results in distorted or incomplete communication. In fact, those who suffer from hearing loss can experience an incomplete communication which will impact negatively to their social lives, at times leading to isolation, withdrawal and lack of independence.

Hearing loss is the disability more closely related to aging (Davis, 1989; Mulrow et al., 1990). In turn, with age it also increases the possibility of emergence of neurodegenerative disorders and communication problems. One of the most common demonstrations of neurodegenerative disorders is the Alzheimer's disease, which tends to affect people over the age of 65 (Acton, 2013). The prevalence of neurodegenerative disorders and specially of Alzheimer's disease is increasingly significant. Worldwide, nearly 44 million people are believed to be living with Alzheimer's disease or other dementias. By 2030, if breakthroughs are not discovered, we will see an increase to nearly 76 million. By 2050, rates could exceed 135 million of people affected by Alzheimer's disease.

Furthermore, recent investigations show that hearing loss is a potential risk factor for cognitive impairment (Lin, 2011). In turn, there is scientific evidence of a possible association between decreased hearing and an increase in Alzheimer's disease (Lin, 2011). Older adults with hearing loss have a rate of cognitive decline that is up to 40% faster than the rate in those with normal hearing. Lin et al. (2013) state that rates of cognitive decline and the risk for incident cognitive impairment are linearly associated with the severity of an individual's baseline hearing loss. This study also estimates that for 2050, 100 million people may suffer different problems related with cognitive decline.

The reasons of this association between hearing loss and cognitive decline might be due to the social isolation that suffer those individuals with degraded hearing, since this social isolation has long-term consequences to healthy brain functioning. Besides, hearing loss may also force the brain to devote too much energy on processing sound, reducing the energy spent on memory or thinking. Co-pathology is a major complication for the diagnosis of hearing problems. Almost all elderly adult will develop some degree in cognitive decline capacity as time progresses. Since aging is highly related to both hearing loss an age related cognitive decline, the coexistence of these two conditions is substantially likely.

The use of hearing aids and hearing rehabilitation process is closely related to the improvement of the social, emotional, psychological, and physical well being of people with hearing disabilities (Gatehouse, Naylor, & Elberling, 2003; of Deafness & Disorders, 2009). Modern hearing aids improve speech intelligibility and therefore communication. The benefits of hearing aids have been demonstrated throughout different scientific research. All these considerations highlight the importance of the conduction of regular hearing checks, specially among the elder population or in case of doubt about the ability of hearing at any age.

Pure-Tone Audiometry (PTA) (addressed throughout Chapter I) has been unequivocally described as the gold standard test for the clinical evaluation of the hearing sensitivity. It determines the faintest tones a person can hear at selected frequencies. This test allows the audiologists to evaluate the hearing capacity and also to determine the prevalence of hearing problems. It is a subjective behavioral measurement of the hearing thresholds, as it relies on the patient response to the pure tone stimuli. Therefore, patient's cooperation is required during the test procedure, which may involve certain operational constraints that will be discussed soon.

During the audiometric evaluation, pure tones are delivered to the patient via earphones and the patient must indicate when he perceives the stimulus (typically, by raising his hand). The performance of this test is typically a completely manual process, which entails certain problems. The measurement of the response times is an additional evaluation that the audiologists accomplish during the hearing assessment. This measure is relevant for the identification of patients with abnormally slow response times, which can be a symptom of any medical condition that should be studied. The problem here is that the expert must have treated and studied a large number of patients in order to be able to know what the average behavior is. Despite the audiologist's skills this is a subjective task making it prone to errors and imprecisions. Even if the audiometric evaluations were recorded in video for a later analysis, the measurement of the response times would be really time consuming for the expert. This is why an automated solution would be very helpful so it would speed up the process and would provide accurate and reproducible measurements.

However, in the case of patients with cognitive decline (or other severe communication disorders), the standard protocol becomes unenforceable since no active interaction audiologist-patient is possible. This specific group of patients has limitations when it comes to maintaining a normal interaction, limitations that are aggravated as the cognitive decline worsens. Even though the evaluation of these patients becomes much more complex, it is still possible is the audiologist focus his attention on the detection of spontaneous subtle facial reactions. The subjectivity involved in the gesture interpretation and the subtlety of the facial reactions makes of this task an imprecise problem, prone to errors and difficult to reproduce. All these reasons make clear the improvements that an automated solution could offer by assisting the audiologists in the detection and interpretation of these unconscious gestural reactions.

The main goal of this thesis is to provide an automatic tool for assisting the audiologist in the evaluation of the hearing capacity. To that end, the detection of the positive responses and the measurement of the reaction times should be accomplished. For overall patients, the typical positive response to the sound that must be detected is the hand raising. In order to extend this automatic tool for the particular case of patients with cognitive decline or severe communication disorders it is necessary to analyze the unconscious gestural reactions to the sound as well. Although they already exist manual solutions for these problematics, they are highly affected by the subjective interpretation of the observer. Training and experience also affect the interpretation of the gestural reactions, while a manual measurement of the response times from a recorded video sequences is very time consuming for the expert. Different computer vision techniques are studied, proposed and evaluated in order to provide an automated methodology which assists the audiologists in the two mentioned problematics.

1.1 Overview

The proposal of this research is to design a system to perform different automatic assessments in the evaluation of the hearing capacity. This system is based on the interpretation of the images acquired with a conventional video camera during the performance of the audiometric evaluations. Different image processing techniques and machine learning algorithms are applied in the development and validation of the automated assessments following presented.

This chapter has introduced the main topics to be presented in this work an a general description of our domain.

Part I initially describes the methodology traditionally conducted by the audiologist to assess the hearing capacity. Throughout this Part a methodology for the automatic assessment of general patients is proposed by assuming the hand raising as the expected positive response. This methodology also addresses the measurement of the patient's response times, and it also evaluates the separability between “normal” and “slow” patients. At the end of this Part, in Chapter 3, a web application designed for the audiologists providing an interface that facilitates the use of the automatic methodology is presented.

The main operational problem arrives when the audiologist tries to evaluate patients with cognitive decline or severe communication disorders and a typical question-answer interaction can not be maintained. Throughout Part II different approaches for the detection of unconscious eye-based gestural reactions to the sound are presented. The first approach is carried out using optical flow and machine learning algorithms. Then, a second approach addresses an alternative solution for the detection these unconscious eye-based gestural reactions to the sound based on the color distribution of the sclera (the white area in the eye). The previous approaches provide accurate (but not optimum) results and they are based on different but complementary foundations, for this reason, a final combination of both techniques is proposed in order to improve the global accuracy. At the end of this Chapter a final improvement in order to increase the classification accuracy of the relevant categories by the use of machine learning techniques is presented.

Finally, Chapter 7 provides a brief overview of some concluding remarks and proposes some future lines of research.

Appendices from A to D address different issues that complement the information depicted throughout the different chapters. Notice that Appendix E reports the author's key publications and mentions.

Part I

Automatizing the hearing assessment

Chapter 2

Hearing assessment

This Part introduces the methodology for automatizing the hearing assessment by the analysis of video sequences recorded during the perfomance of the audiometric evaluations. In order to properly understand the domain and the related circumstances it is necessary to provide first a general description of the protocol for hearing assessment.

Pure-Tone Audiometry (hereinafter PTA) is the standard test to identify the hearing threshold levels of an individual. It determines the faintest tones a person can hear at selected frequencies. It allows the audiologist to diagnose the presence or absence of hearing loss by determining the softest sound that can be perceived in a controlled environment. PTA is a subjective, behavioral measurement of hearing threshold, as it relies on patient response to pure tone stimuli. Behavioral hearing tests require the participant to reliably demonstrate a change in behaviour when a test sound is heard. Therefore, PTA is mainly used on adults and children old enough to cooperate with the test procedure. PTA provides ear specific thresholds, and uses frequency specific pure tones to give place specific responses, so that the configuration of a hearing loss can be identified.

The test should be conducted in a specific soundproof room or a quiet place with no noises. The patient wears earphones connected to a device called audiometer (a sample of this kind of devices can be seen in Figure 2.1), so auditory stimuli can be delivered to each ear separately. These auditory stimuli are pure-tone sounds at different frequencies and intensities. The audiologist will test if the patient is able to hear a variety of different pitches. Cooperation is needed during the test procedure, since the patient taking this test is typically asked to raise his hand or to show some kind of positive reaction when he perceives the sound.

Figure 2.1: Madsen Xeta by Otometrics audiometer.

The results of hearing sensitivity are plotted on a graph called audiogram (see Figure 2.2), which is a graph displaying intensity as a function of frequency. The frequency in hertz (Hz) is displayed over the horizontal axis (with low frequencies on the left increasing to high frequencies on the right), and a linear dBHl scale on the vertical axis. There will be a series of symbols across the chart (see Figure 2.3). The position of these symbols on the chart indicates the quietest sounds the patient can hear at different frequencies.

The frequencies more commonly tested are 100, 250 and 500 Hz, and 1, 2, 4, and 8 KHz; and the intensities are commonly plotted range from -10dB to 110dB, in multiples of 10. The range between 100Hz and 8KHz represents the most important levels for clear understanding of speech. The results of the audiometric test determine the subject's hearing levels. Normal conversation speech is about 45dB. Normal hearing is expected to be between -10dB and 20dB. According to the obtained results hearing can be classified in (see Figure 2.2): normal hearing, mild hearing loss, moderate hearing loss, moderately severe hearing loss, severe hearing loss and profound hearing loss.

Prior to any exploration, an otoscopic examination is performed to check the absence of any obstruction (e.g, earwax) in the outer ear canal, since obstructions can interfere to the hearing capacity; for this reason, they must be solved prior to any audiometric evaluation. For this exploration, audiologist typically use a manual otoscope. In the case of excessive cerumen in the ear canal (inability to see more that the 50% of the tympanic membrane), the patient must be referred to the appropriate specialist and the hearing assessment must be rescheduled. This otoscopic

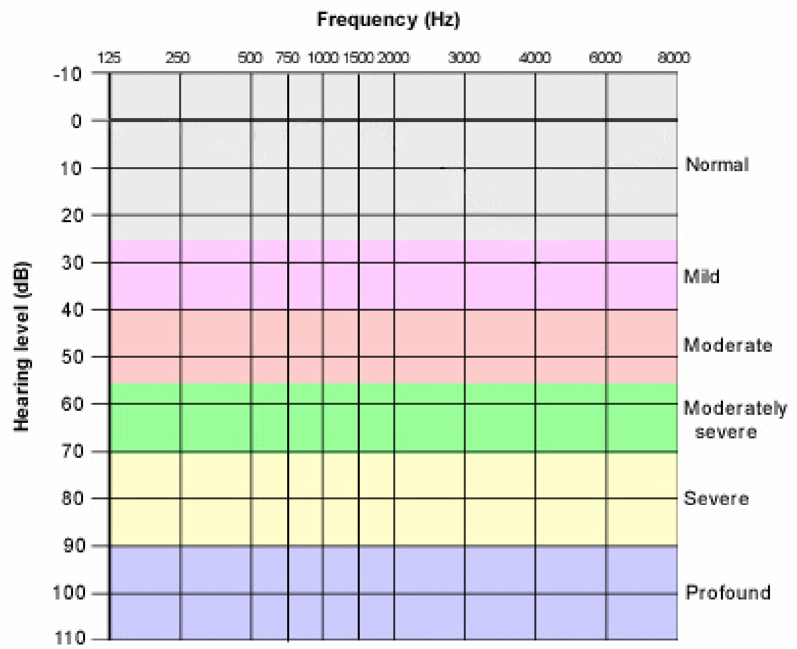


Figure 2.2: Audiogram sample. According to the results charted here, hearing can be classified in: normal, mild, moderate, moderately severe, severe or profound.

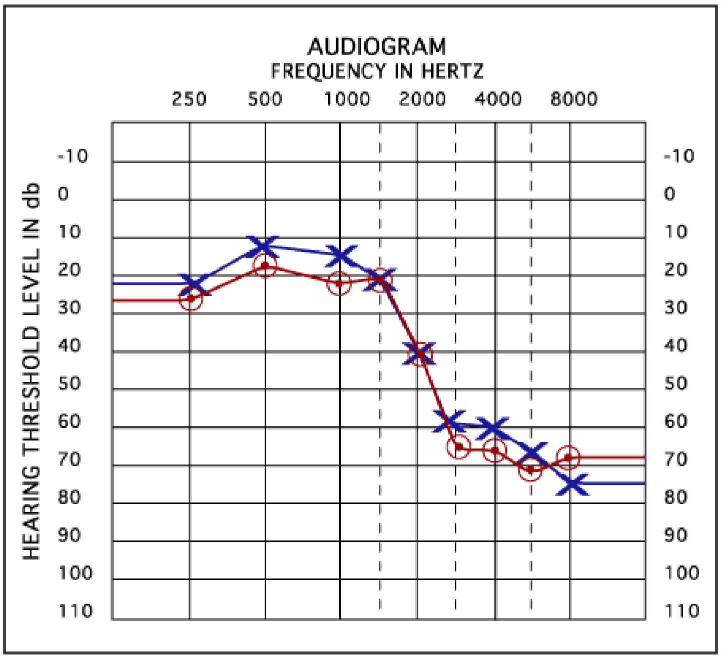


Figure 2.3: Audiogram sample. The X's and blue lines are responses for the left ear and the 0's and red lines correspond with the right ear.

examination allows also to determine if the eardrum presents any damage that can reduce hearing, such as perforations in the eardrum of congenital malformations, circumstances that should be considered in the interpretation of the hearing assessment results.

Next, the audiologist explains to the patient the protocol for the audiometric test. Since it is a behavioral test, it is highly important that the patient understands the instructions given. This need of understanding is what makes difficult the assessment of patients with cognitive decline or patients with a profound hearing loss without hearing aids. For patients without impairments, the audiologist indicates to them that they are going to receive different types of auditory stimuli via earphones and that they must respond to them affirmatively, usually by raising his hand, when they perceive them. Since each ear is evaluated separately, patient must respond consistently, by raising his right hand when he perceives the auditory stimuli on the right ear, and equivalently for the left ear.

The performance of the PTA allows the audiologist to evaluate air and bone conduction, this way, the type of hearing loss can also be identified via the air-bone gap. For the air conduction audiometry, the patient wears conventional earphones and the results establish the extent of sound transmission through the bones of the middle ear. For bone conduction, patient wears a vibrating ear-piece placed behind the ear next to the mastoid bone. This bone vibrator uses the skull to transfer the sound vibrations to the cochlea (the hearing organ of the inner ear), by-passing the ear canal and middle ear. Thresholds obtained with the bone conductor are called bone conduction thresholds. Occasionally a wind-like noise called masking is used to occupy one ear while the tester determines in which ear the beeps are being heard. Results of bone conduction determine the extent to which there is sensorineural hearing loss. If a hearing loss exists, bone conduction helps the audiologists to determine whether the problem is in the outer, middle, or inner ear. If hearing difficulty is due to a problem with the middle ear it can be due to ear infection; however, if the problem is in the inner ear it may be due to aging, noise exposure, or a variety of other causes.

The auditory stimuli are sent through an audiometer, where the audiologist sets the different frequencies and intensities that he wants to test. Once the expert has selected the frequency and the intensity, he delivers the auditory stimuli to the patient and waits for his reply. This way, the examiner systematically finds the softest sounds the patient can hear across a range of frequencies and determines the hearing thresholds (the softest sounds that the patient is able to perceive). If the patient is able to perceive the stimulus he raises his hand (see Figure 2.4); in the event the patient does not respond to the delivered stimulus, the audiologist can try to deliver the same stimulus again of to increase the intensity and send a new stimulus. In the case of patients with cognitive decline, the communication process will be more complex, but the handling of the audiometer and the delivery of the auditory stimuli remain the same.

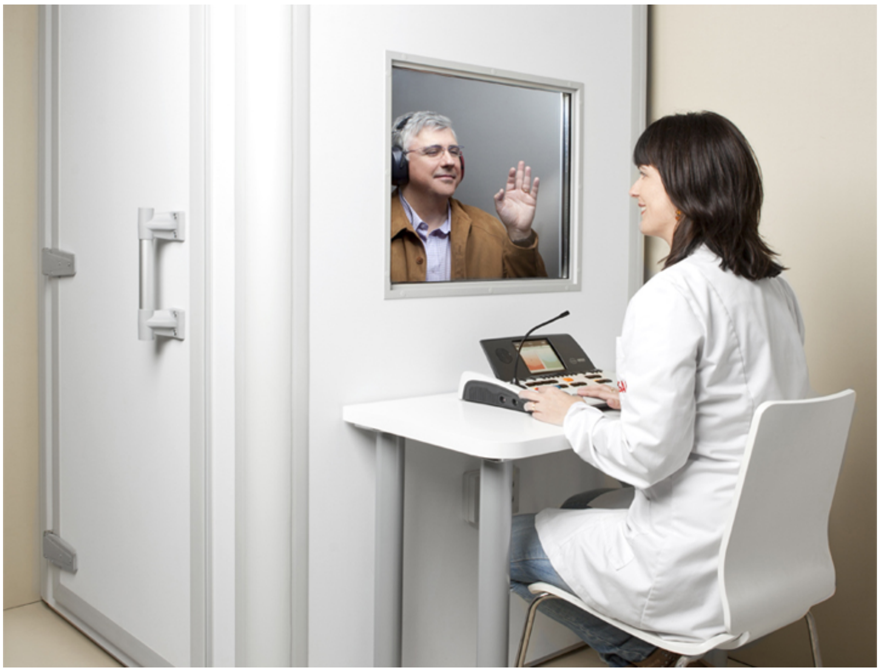


Figure 2.4: Audiometry sample into a soundproof room. The patient raises his hand to indicate a positive response.

It must be noted here that it is highly important for all this kind of procedures to maintain this clinical protocol as maximum stable as possible, without involving significant changes in the behavior of the patient or the audiologist. A quiet and silent environment is required for the proper development of audiometric evaluations. Besides, a high level of concentration is required from the patient, since forgetfulness or lack of concentration could lead to inaccurate diagnosis. It has been observed by the experts that significant changes over the traditional protocol affect to the patient's concentration, for this reason, it is required to modify this protocol as minimum as possible, which adds difficulties to provide automated technical solutions.

As mentioned before, the typical expected response during an audiometric evaluation is the hand raising, so this is the behavior that we are going to automatically detect with the proposed methodology. Besides providing a system able to analyze the video sequences recorded during the performance of the audiometric evaluations and automatically detect the patient's positive responses as the audiologist does, it is also important for us to measure the response times. A standard psychometric measurement is response time, the interval elapsing between a stimulus and a response. A substantial part of the response time reflects a central delay due to the brain processing the input and elaborating a response. Inefficient central processing, lack of motivation or low stimulus intensity (which raises doubts in the patient) are among factors that increase reaction times.

The classification of patients according to how quickly or not they respond to the auditory stimuli requires a certain degree of expertise to the audiologist. This classification allows the experts to detect those patients who have a response time considerably slow compared to the average. The importance of the detection of patients with response times abnormally slow relies on the possible association of this slowness as a symptom of some type of cognitive problems or other pathologies that should be studied. In the case of response times abnormally slow, patients are referred to an specialist. Response times in audiometries had been also under study for clinical evaluations as the one presented in (Kelly, Walsh, Norman, & Cunningham, 1999), where a study about the influences on the reaction times under the effects of an specific drug was conducted.

To carry out this discrimination between normal and slow patients, the expert must have treated and studied a large number of patients in order to determine what the average behavior is, and thus, to be able to correctly classify the patients according to their response times. Still, even for experts with enough experience, this is a subjective task, which makes it very prone to errors. It is important to note that, sometimes, slowness on the patient's movements might confuse the expert's interpretation. Furthermore, the main target of the audiologist it to be focused on the hearing assessment (handling the audiometer, waiting the patient's answer and registering results); for this reason, the fact of having to be focused on the response times too might be lead to imprecisions.

2.1 The automatic proposal

During the audiometric evaluation the patient is asked by the audiologist to raise his hand when he perceives the auditory stimuli that are delivered to him through the audiometer. Prior to this proposal there was not any tool to assist the audiologist during this analysis, so the only conclusions that the expert could draw from the test about the patient's behavior were qualitative interpretations based on his experience. That is the main reason why the expert's conclusions could be sometimes subjective or influenced by previous experience with the patient. By automating the detection of the hand raising and automatically measuring the response times we allow the experts to extract quantitative information about the patient's response making possible the performance of new surveys and statistical studies. Besides, since the hearing assessment are recorded in video, experts can evaluate the tests after they have been conducted and compare their impressions with the objective measurements provided by the methodology

The possibility of registering the tests on video allows the expert to assess the results, and even to be interpreted by more than one expert in order to obtain a classification more robust and less biased, but this solution has the problem of being a slow process that requires a lot of time by the expert contributors, which is an additional problem of the manual approach.

For all these reasons, an automated solution is very relevant in this domain in order to speed up the process and to obtain accurate and reproducible measurements allowing an objective classification. This proposal will also pave the way for a totally automated hearing assessment system. A novel method is proposed for human body part detection focused on face and hands or arms detection for automatizing the audiometric process (Fernández et al., 2012). This justifies the introduction of computer vision techniques in order to automate a number of processes of the problem by analyzing video sequences: detection of the stimuli sending, detection of the patient, detection of the hand raising and measurement of the response times.

The proposed methodology needs to detect the patient and to be able to measure how long he takes to positively react to the auditory stimuli, which in this case is manifested by a hand raising. A schematic representation of the methodology is depicted in Figure 2.5. The stimuli detection stage is accomplished due to the need of correlating the patient's responses and the auditory stimuli that he receives and it is dependent on the device used by the audiologist. The response reaction corresponds with the part aimed to detect the patient's hand raising. And finally, the analysis of responses to the stimuli correlates both information and provides a measurement of the response times. Each one of these steps is going to be addressed through the next sections, but first, we are going to detail the video sequence acquisition.

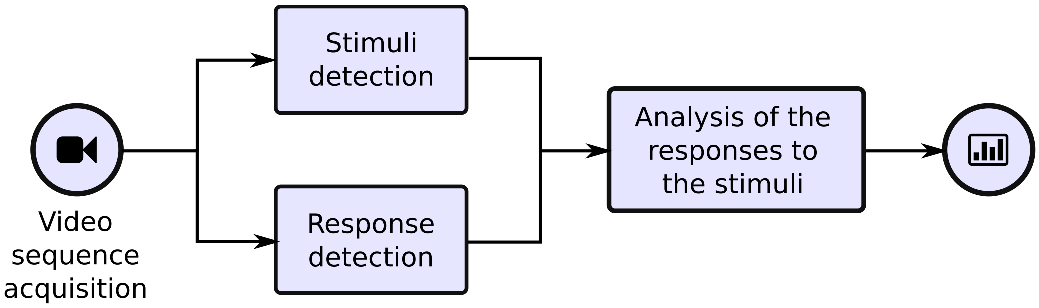

Stimuli detection
Video sequence acquisition
Analysis of the responses to the stimuli
Response detection

Figure 2.5: Schematic representation of the methodology.

2.2 Video sequence datasets

The procedure for the acquisition of the recorded scenes, and the different video sequence datasets used in this research are subsequently described. These datasets were acquired at different situations, with different illumination conditions and from a wide range of patients. They were annotated by audiologists in order to test the proposed automated assessments. All images have been acquired and annotated by audiologists from the Faculty of Optics and Optometry, University of Santiago de Compostela (Spain).

2.2.1 Video sequences acquisition

The recorded video sequences were obtained using a conventional video camera. The only requirements for the video sequences used with this methodology are that: they must have high resolution to cover with enough definition the whole scene and a frame rate of 25 FPS (frames per second).

With the purpose of maintaining the clinical protocol as maximum stable as possible, without disturbing the patient's concentration, the video camera will be located in a discrete location behind the audiologist (the audiologist will be seated in front of the patient). This location will allow us to record not only the patient (which will be seated at a determined position) but also the audiometer, through which the audiologist will be delivering the auditory stimuli that the patient perceives via earphones. A schematic representation of this scene can be observed in 2.6.

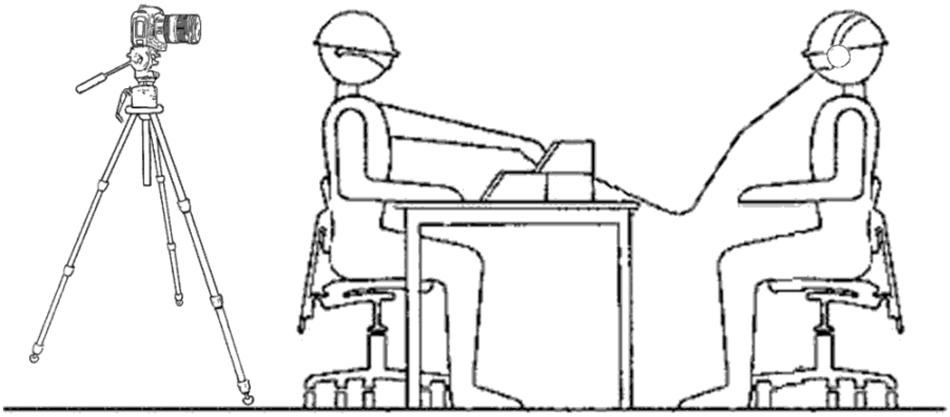


Figure 2.6: Schematic representation of the scene

The reason for such a general scene is that it is necessary to study the patient's responses (whether they are expressed as a hand raising or they are eye-based gestural reactions) and also the delivery of the stimuli that the experts sends by performing the audiometer. The necessity for such a general scene in conjunction with the precision required to evaluate eye gestural reactions is the reason why Full HD is required for the development of this methodology. A sample of the recorded scenes is presented in Figure 2.7.

2.2.2 Illumination conditions

The influence of the illumination conditions is specially relevant in those cases where color information is used. For the measurement of the response times skin color detection needs to be accomplished, so the methodology must take into account the influence of the illumination conditions in order to guarantee the proper behavior of the method regardless of the situation.

Changes in the illumination conditions during the performance of the audiometric evaluation can cause shades or other situations that substantially modify the appearance of the scene and can induce to detection inaccuracies. For this reason, it is suggested to try to maintain favorable and constant lighting conditions when recording the scenes, in order to improve the quality of the recorded images and to avoid shades or occlusions.

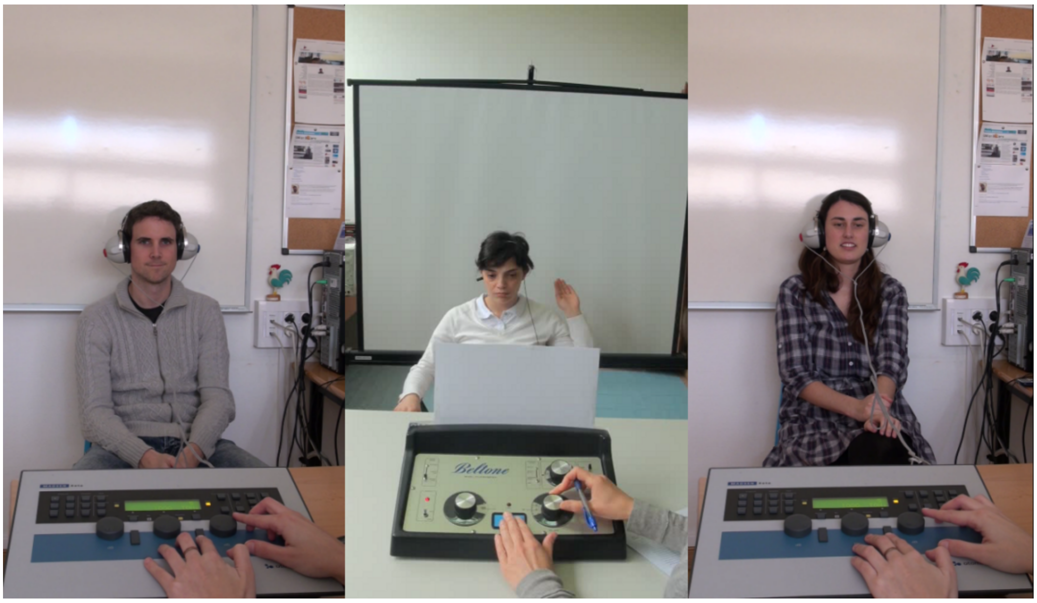


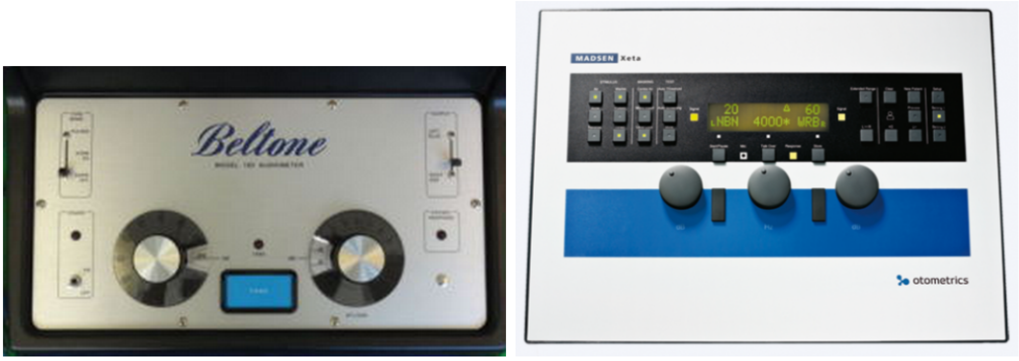
Figure 2.7: Sample of the recorded images.

2.3 Stimuli detection

As pointed out by the audiologists, a response time should be measured from the moment the auditory stimulus starts until the moment when the patient's reaction starts. Considering this, it is necessary to have information about the stimuli delivery so it can be correlated with the information about the patient's response.

Although there are digital audiometers in the audiological market, most of audiologists still work with analogical devices. In the case of digital devices, the correlation of the positive reactions with the stimuli delivery could be more easily addressed (anyway, it would be necessary to find a way to correlate and synchronize that information with the captured video sequence). However, in the case of analogical devices it is necessary to automate somehow the detection of the moments when the stimuli are being sent. Without this step, it would not be possible to automatically correlate stimuli and responses using only the video sequences recorded.

This stage of the methodology is ad-hoc to the device. In our particular case, audiologists work with two different analogical audiometers: the Beltone Electronics 190 audiometer and the Madsen Xeta by Otometrics audiometer (see Figure 2.8), so the proposed method was adapted to both of them. Since analogical audiometers usually have a similar design, the adaptation to new devices should be easily addressed.



(a)

(b)

Figure 2.8: (a) Beltone Electronics 109 audiometer. (b) Madsen Xeta by Otometrics audiometer.

In this specific domain, we know that when the expert delivers the auditory stimuli a light turns on on the device (the location and color of that light depends on the specific device). This way, this task is going to be accomplished by defining areas of interest and seeking for image templates using normalized cross correlation. These assumptions can be made since we are working in a specific domain and the experiments are going to be recorded always following a similar layout as the one shown in Figure 2.9, where the patient is seated in front of the camera, and in front of the audiologist performing the test, and the audiometer is located on the lower part of the image.

We are going to exemplify here the steps followed in the case of the Beltone Electronics audiometer (Figure 2.8(a)). Anyway, these proposed steps are almost analogous for the Madsen device (Figure 2.8(b)).

The first step consists on defining a region of interest (hereinafter ROI), and within this region locating the light of the audiometer that indicates the stimuli emission. In the case of the Beltone Electronics audiometer the stimuli indicator is a red light located just above the touchpad. In the case of the Madsen audiometer there are two stimuli indicators, one at each side of the digital screen, where the left light indicator corresponds with the delivery of pure tone sound to the left ear, and the right light indicates the delivery of stimuli to the right ear. Is it possible to guarantee that in our domain this ROI is going to be located in the lower third of the image (since the audiometer is always located at the bottom of the image); and, considering that the stimuli indicators are in the central part of the two models of

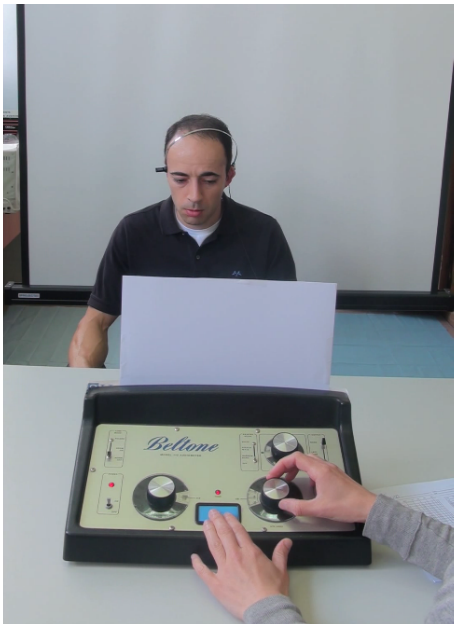


Figure 2.9: Screenshot during the audiometric evaluation

audiometers, we can center the ROI on the central third of the lower third of the image (the highlighted area in Figure 2.10(a)).

Since the stimuli indicator is a small light located whitin the audiometer, we reduce the search area by reducing the first ROI. To that end, we locate the blue touchpad (in the case of the Madsen device, we locate the green digital screen). This decision is motivated by two reasons: first, it is easier to find a larger region which is also completely different from other items in the device; and second, the light that indicates the stimuli emission is not the only light present in the panel, so we need to establish a reliable criteria that allow us to distinguish the specific light in which we are interested in. Thus, using a template image of the touchpad (see Figure 2.11(a)) and normalized cross correlation we locate the position of this touchpad panel, and based on this location, we define a second ROI.

To ensure the presence and the correct location of the touchpad, a normalized cross correlation with a high value (greater than 0.95) for five frames (not necessarily consecutive) is required. In order to prevent that a frame with a very different location, but also with a high correlation, could alter the results, the four locations with the most similar positions to each other are chosen, and their average position is computed and taken as reference for the next steps. This way, we try to ensure that the location that we obtain for the touchpad is stable and consistent.

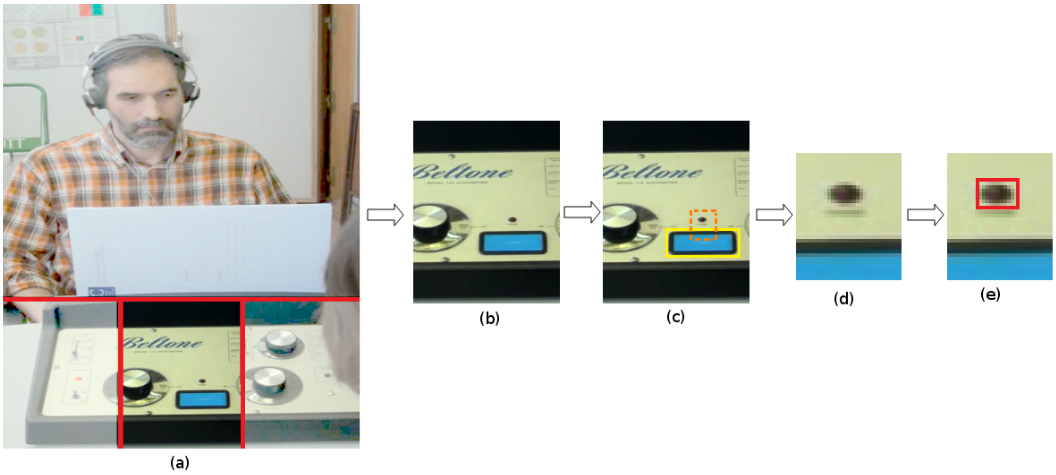


Figure 2.10: Steps towards the location of the stimuli indicator. (a) General scene with the search area highlighted. (b) Initial search area. (c) Touchpad located and definition of the new search area. (d) Secondary search area. (e) Location of the red light.

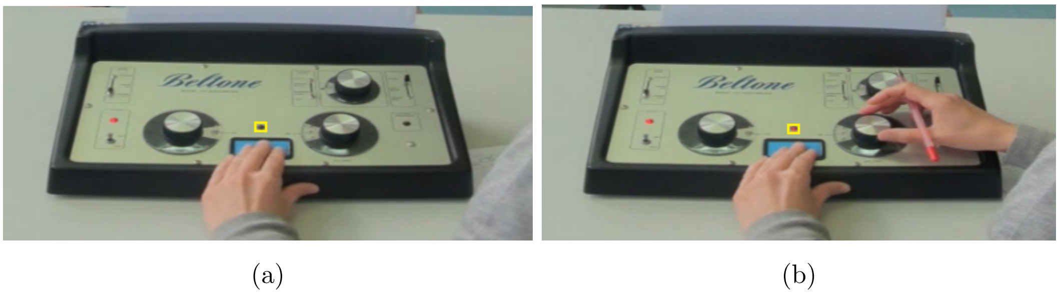
(a) (b)

Figure 2.11: (a) Blue touchpad template. (b) Stimuli indicator template

The new region of interest is defined from this average position of the touchpad as it can be seen in Figure 2.10(c): centered in the central third of the touchpad, with a height equal to the height of the touchpad and starting 1.5 times above the center of the region.

Within this narrower ROI (corresponding to Figure 2.10(d)) we use a new template for the red light (see Figure 2.11(b)) in order to finally determine the location of the stimuli indicator. After locating this position (by using normalized cross correlation again), it will be stored by the system in order to check at each frame of the video sequence if the light is on or if it is off, namely, if a stimulus is being delivered by the audiologist or not.

Towards checking the on/off status of emission we assume that: during the initial frames of the test the expert is not sending stimuli, and as consequence, during the initial detections the light is off. So, in the HSV color space H and S components are considered during two frames and their average is stored as the *off-value*. This way, a relative threshold is established requiring that, to consider the light as *on*, the value of the H and S components must be higher than 1.2 times the value stored as *off-value*. In Figure 2.12 we can see two samples, Figure 2.12(a) when the stimuli indicator is off and Figure 2.12(b) when the stimuli indicator is on. It was tested that the establishment of a relative threshold is more appropriate than a fixed threshold since the illumination conditions may affect the appearance of the stimuli indicator.



(a) (b)

Figure 2.12: (a) Stimuli indicator off. (b) Stimuli indicator on

Applying this criteria to all the frames across the video sequence, we can determine the precise moments when the auditory stimuli are delivered to the patient, being able to obtain a stimuli signal as the one in Figure 2.13, where the pulse is up when a stimulus is being sent and down otherwise.

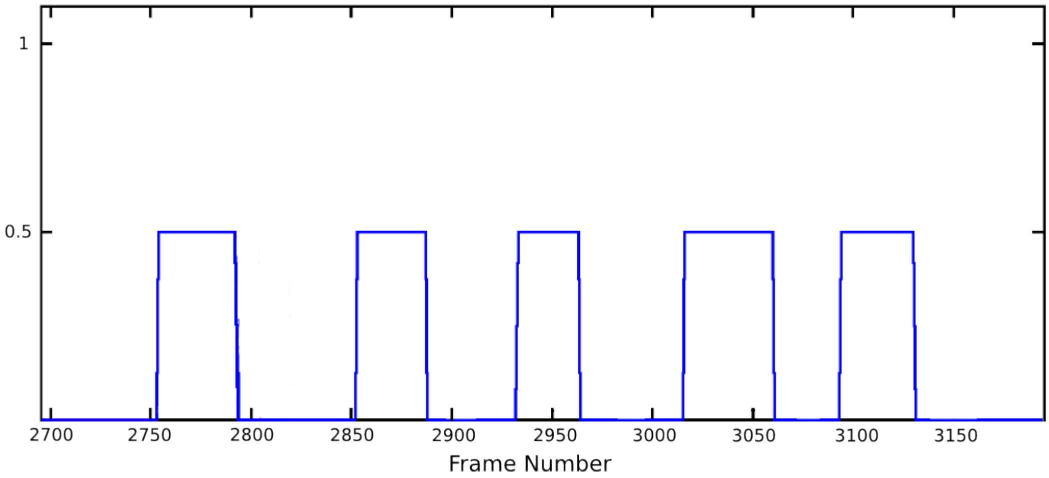


Figure 2.13: Auditory stimuli signal: the signal is up when a stimulus is being sent and down otherwise.

2.4 Response detection

In Figure 2.14 a schematic representation of the different steps of this part of the methodology can be observed. This approach receives as input the recorded video sequence (which is going to be processed frame by frame) and, as a result, it provides the detection of the positive responses of the patient (expressed as a hand raising). To that end, we first determine the patient's location, and after that, we detect the hand raising.

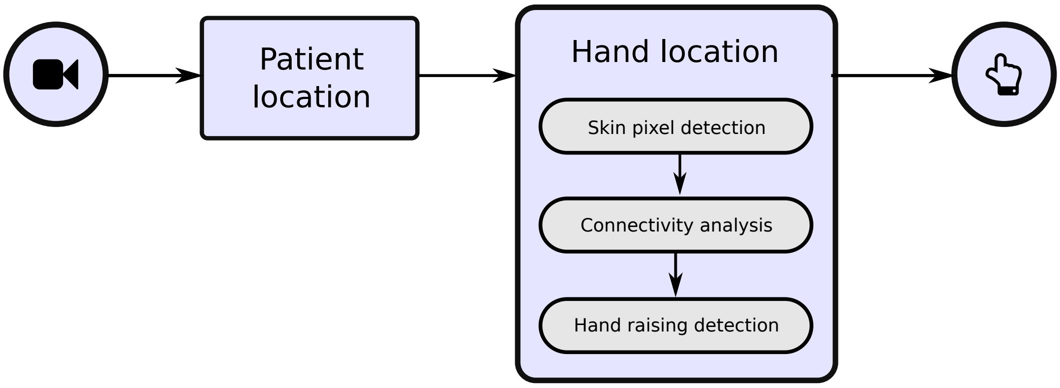


Figure 2.14: Schematic representation of the hand raising detection steps.

2.4.1 Patient location

The location of the patient is determined by the location of the face. Patients will be seated in front of the camera, so the location of the face provides a consistent location of the patient in the recorded scene. Since during the audiometric evaluation patients will remain seated it is possible to ensure that his position will not vary substantially during the test.

As we are working on a pre-fixed domain where the conditions in which the test is performed are stable and already known, we can ensure that faces will always be recorded in frontal position. The location of the face will serve as reference for the subsequent steps of the methodology, since this location facilitates the location of the hand when it is raised nearby the face. Due to the stability of the domain, we can apply a Viola-Jones approach (Viola & Jones, 2001) for the location of the face. Since face location is a recurrent procedure across this thesis, it will be addressed in detail on the Appendix A, where we also present a couple of samples of the face location provided by this method.

2.4.2 Hand detection

Although the hand raising is not the only positive reaction that a patient may show, experts have pointed out that it is the more valid in their domain. There are other alternatives where the patient must press a button in order to indicate that he has perceived the stimulus, however, our experts consider that this alternative could induce to inaccurate assessments. To press a button is not a big effort for the patient, so, in some cases, he could start to automatically press the button without paying his full attention to the assessment. Nevertheless, the fact of raising a hand is a more conscious gesture, which implies that the patient needs to put more attention on the assessment. For this reason, audiologists prefer to ask the patient to raise his hands as sign of perception, and it is the only conscious reaction admitted by our experts.

There is not an established and recognized technique along the literature for the hand detection task. There are different solutions based on different approaches but there is not a clear reference method. Our proposal here is to use skin color information, since both hands and arms are skin regions and we can ensure that, at least the hands, will always be exposed.

Some of the literature related to hand detection takes as input images in which it only appears a hand on a more o less complex background. This is the case of the proposals of Peng, Wattanachote, Lin, and Li (2011) or Dawod, Abdullah, and Alam (2010), where color information is used for the skin color detection; however, these two methods have the limitation that they do not perform correctly with more than one skin region in the scene. In (Caglar & Lobo, 2006) the authors avoid the consequences of using skin color information by exploiting the geometric properties of the hand, identifying parallel finger edges and curved fingertips. Using color information in combination with motion information we can find some hand tracking approaches such as (Spruyt, Ledda, & Geerts, 2010) or (Asaari & Suandi, 2010).

Since both hands and arms are skin areas, color information is a useful feature for their detection. Color provides robust information against rotation, scaling and partial occlusions (Kakumanu, Makrogiannis, & Bourbakis, 2007). The performance of a suitable skin color segmentation is an essential task for the proper behavior of the rest of the process. Prior studies have demonstrated that different skin colors from different races fall in a compact region of the color spaces (Yoo & Oh, 1999; Yang & Waibel, 1996).

In this section, the different stages for the detection of the hand raising are addressed. To achieve this purpose it is necessary to identify skin areas; therefore, we start by introducing a study on the suitability of different color spaces for the skin detection task. After this initial study, the different stages addressed in this section are: skin pixel detection, connectivity analysis and, finally, the hand raising detection. The sequence of this steps is represented on Figure 2.15.

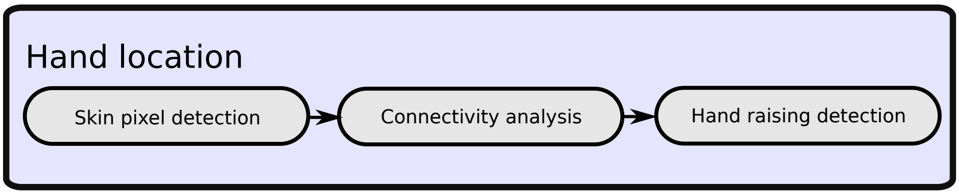


Figure 2.15: Steps of the hand location.

Color space analysis

When working with color filtering, the correct choice of the color spaces becomes very relevant. The selection of an appropriate model depends on the shape of the skin distribution in a given chrominance space, which in turn depends on the space that is chosen. The color spaces proposed to be considered in this formal study for skin detection were: RGB, HSV, CIE L*a*b* and TSL (Terrillon & Akamatsu, 2000) (see Appendix B for more information). Since we have a special interest in having good results regardless of the lighting conditions, we decided to work only with those color spaces that have a specific component for brightness. This implied to discard the RGB color space(which do not have a specific component for brightness), considering only HSV, CIE L*a*b* and TSL color spaces for our analysis.

In the proposal of Sandeep and Rajagopalan (2002), the HSV color space is chosen for skin segmentation tasks. According to the authors, it gives a good performance for the skin pixel detection, providing clearly superior results to the ones obtained with RGB or YCrCb color spaces. As well, the CIE L*a*b color space was considered because it was designed to approximate human vision. Finally, TSL is a color space specified in terms of tint (T), saturation (S) and luminance (L); it was designed having as target to provide an efficient detection and location of human faces in static images; their authors have demonstrated its robustness in the face detection task under different lighting conditions, points of view or scale.

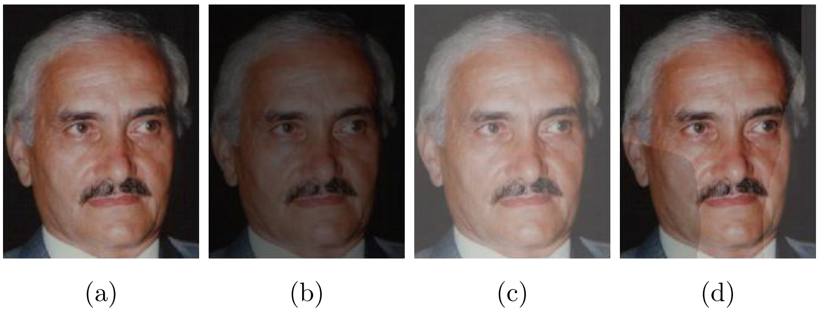
Factors such as the illumination conditions can spoil the performance of color filtering based applications. Our goal is to obtain as much invariance as possible regarding to the lighting conditions. We will show how, discarding the brightness component from the considered color spaces, and working only with the two remaining components, the skin tones are going to fall into similar areas with independence of the illumination conditions.

In order to evince this particularity of the chosen color spaces, we took a face image (Figure 2.16(a)) and we applied artificial light changes on it, obtaining this way different illuminations over the original image: Figure 2.16(b) a darker image, Figure 2.16(c) a lighter image and Figures 2.16(d) an image with heterogeneous changes of illumination.

Samples of skin and non-skin pixel were taken from these images, discarding the brightness component and considering only the two remaining components. As shown in Figures 2.17(a)-(c), if the two remaining components are represented in a 2D space, the skin tones (red dots) fall broadly on similar areas.

For the color space analysis, color images from a human face database (*FG-NET Aging Database*, n.d.) were used to extract skin pixels from them using manually defined masks. The resulting training set consisted of more than 2700000 skin pixels from 60 different faces.

Histograms for the skin and non-skin pixels of each color space were computed.



(a) (b) (c) (d)

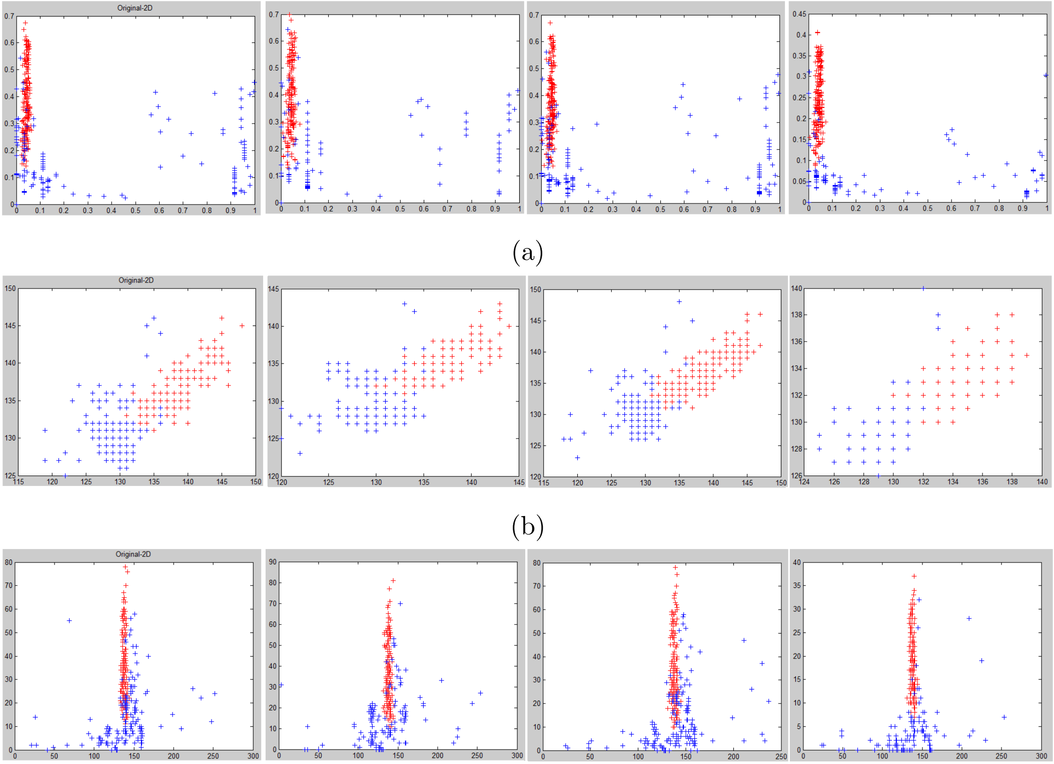
Figure 2.16: Illumination changes over a face image. (a) Original image. (b) Darker image. (c) Lighter image. (d) Image with heterogeneous changes of illumination.

The considered components for the computation of these histograms were: H and S for HSV color space, a and b for CIE L*a*b color space and T and S for TSL color space. Values of these components are normalized to 100 bins. Finally, when the histogram is totally computed, it can be observed in Figure 2.18 how one region is very much apart, with high skin probability, which shows the separability of the skin pixels.

Based on these computed histograms, and in order to classify a pixel as skin or non-skin, we can establish a threshold defining a minimum probability (height) in the histogram. The height of a bin in the histogram is proportional to the probability that the color represented by that bin corresponds with a skin color, so pixels corresponding with a higher value than the threshold are classified as skin pixels. Otherwise, they are classified as non-skin pixels. In order to compare the three proposed color spaces (HS, ab and TS) we have calculated the false negative (FN) and false positive (FP) rates for each one as a function of the threshold value in a 60 image test set. Figure 2.19 shows the obtained results.

According to the results represented in Figure 2.19, the CIE L*a*b color space is the first one to be rejected. It showed to be clearly worst than the other color spaces since it has a behavior much more abrupt than desired. Although the minimum error rate correspond to HS(V), TS(L) seems like a more suitable color space because it provides a much more stable error rate on the threshold range thus decreasing the dependence of the method quality to threshold selection. In particular, even if the false positive rate of the TS(L) color space takes longer to decrease, we have a much larger interval where the behavior will be reasonably good since the TSL false negative rate grows more slowly.

The main interest is the proper skin identification, rather than the misidentifi-



(a)

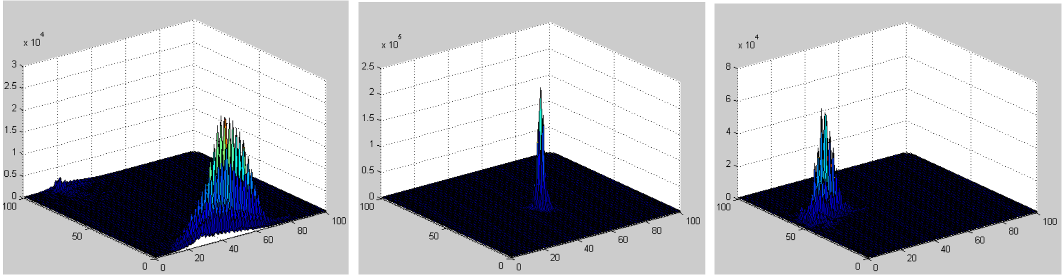
(b)

(c)

Figure 2.17: Distribution of the two considered components for different lighting conditions. First image: original, second image: darker, third image: heterogeneous and fourth image: lighter. (a) H and S components for HSV color space. (b) *a* and *b* components for CIE L*a*b* color space. (c) T and S components for TSL color space.

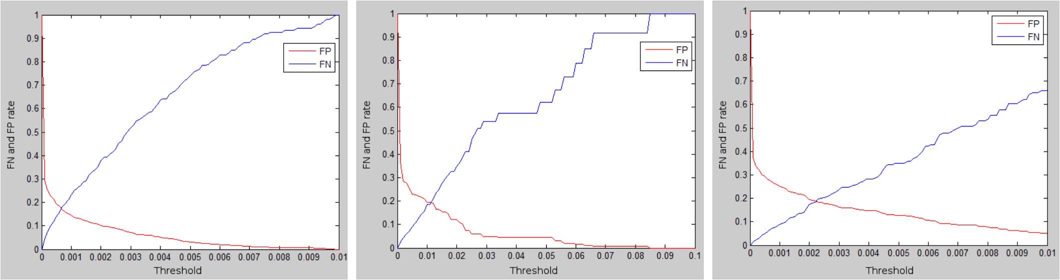
cation of some skin pixels as non-skin pixels. Since in our methodology connectivity analysis is going to be considered for creating skin regions, it is more important not to produce many false negatives in this step.

After the choice of the TS(L) color space, the threshold for the skin pixel classification needs to be established. If the threshold value is too high, all the non-skin pixels are going to be correctly classified, but some of the skin pixels are going to be classified as non-skin pixels too. By contrast, if the threshold value is too low all the skin pixels are going to be correctly classified while some of the non-skin pixels are going to be incorrectly identified as skin pixels. The goal is to find an optimum threshold that allows the detection of most of the skin pixels and the rejection of most of the non-skin pixels. Searching for a trade-off between the false positive and the false negative rates, and also trying to find the center of the interval of the curve



(a) (b) (c)

Figure 2.18: Skin histograms: (a) for HS(V), (b) for (L)ab and (c) for TS(L).



(a) (b) (c)

Figure 2.19: False negative (FN) and false positive (FP) rates as a function of the different thresholds applied to the histograms of Figure 2.18. (a) For HS(V). (b) For (L)ab. (c) For TS(L).

showed in Fig. 2.19(c), the 0.002 value was established as our threshold (called from this moment on *skinthreshold*). Of course this is valid in a 100 bin scenario.

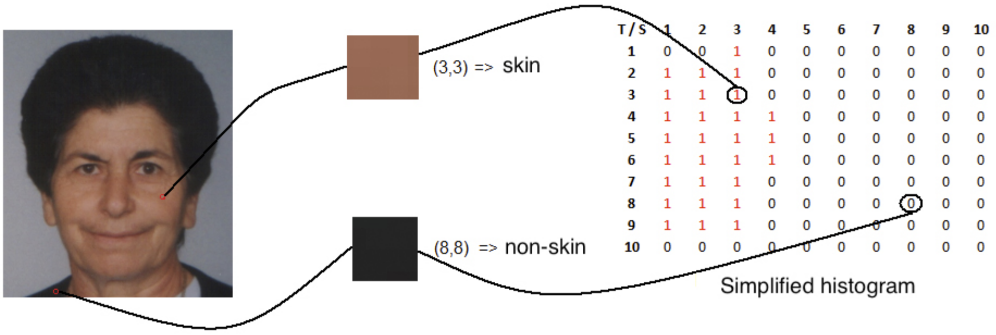
Once the color space is selected and the *skintreshold* is established, we are going to be able to detect skin-like regions and use this information in order to detect different human body parts. To that end, we resume the sequences detailed in Figure 2.15, which were: skin pixel detection, connectivity analysis and, finally, hand raising detection.

Skin pixel detection

Color filtering is a powerful tool in computer vision applications, specially for the detection and tracking of human body parts. Color processing has a low computational cost and it is robust against geometrical transformation (e.g., rotation, scaling and shape changes). However, it is necessary to be careful against illumination conditions or other factors that can spoil the performance of color filtering based applications.

The detection of the different body parts is accomplished using color information in order to detect skin-like regions. The combination of color information with other types of information and an appropriate set of rules will allow us to detect the hands (hands and exposed arms) in our audiological domain.

When working with color filtering, the correct choice of the color space becomes very relevant. We conducted a formal study of some of the most well-known color spaces in order to find the most suited one for the skin color detection task. As a result of this previous analysis the TSL (Terrillon & Akamatsu, 2000) color space has been chosen. In order to avoid the influence of the lighting conditions only the T and S components are considered, while the luminance component (L) is discarded. For each pixel, their T and S components are localized in the TS(L) histogram obtained in the previous section. If the corresponding value is higher than the established *skinthreshold*, the pixel is classified as skin pixel; otherwise, it is classified as non-skin pixel. A simplified sample of this behavior can be observed on Figure 2.20, where a quantized binary map in TS space is used for pixel classification.



T / S	1	2	3	4	5	6	7	8	9	10
1	0	0	1	0	0	0	0	0	0	0
2	1	1	1	0	0	0	0	0	0	0
3	1	1	1	0	0	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0	0
5	1	1	1	1	0	0	0	0	0	0
6	1	1	1	1	0	0	0	0	0	0
7	1	1	1	0	0	0	0	0	0	0
8	1	1	1	0	0	0	0	0	0	0
9	1	1	1	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0

Figure 2.20: Simplified example of the skin pixel classification. Each possible value on the quantized TS map is assigned to a binary value for skin (1) or non-skin (0) classification according to *skinthreshold*.

Connectivity analysis

In the previous step each pixel on the image is classified as skin or non-skin pixel. However, the information on the pixel level is not enough for identifying human body parts. At this stage we know whether a pixel corresponds to skin or not, but we need to go into a higher level in order to group the connected skin pixels so they can represent something meaningful as a group, for example, a face or a hand.

The skin pixels geometrically connected to other skin pixels are classified into the same group of pixels, this is done in our case by applying 8-connectivity neighborhood. This way, if one skin pixel has got another skin pixel located in any of its 8 neighboring places, then, both skin pixels belong to the same region. After this region growing step, we obtain different skin regions to be classified as the different body parts.

However, color information can be insufficient in certain cases. For example, when nearby body parts are present, they can be so close to each other that they can be considered as a single skin region. Something similar can happen if we have in the image colors that are similar to skin tones but that do not really correspond with skin; different skin regions can be merged into one if those skin-like pixels (that are not skin) connect the two real skin regions.

After studying those cases, we found out that many of the points located close to a body part with similar tonalities to skin may be at the same time, edges. This way, the skin pixel classification can be improved by considering in addition edge information and modifying the way that skin pixels are classified. Pixels with high gradients (like pixels located at the boundary of the body part) are removed from the skin pixel classification in order to avoid the connecting path. The goal is to remove only the stronger edges in the image; for this reason, soft edges such as scars are not affected by this new rule. Sobel operator (Sobel & Feldman, 1968) is going to be applied for detecting edges. It is a fast detector which detects edges at finest scales and it also has got a smoothing along the edge direction, which avoids noise edges.

This way, the skin pixel classification is modified as follows: to consider a pixel as a skin pixel it is necessary that it exceeds the *skinthreshold* and also that it has got a gradient lower than a certain threshold called *edgethreshold*, empirically computed too (established on the 65% of the maximum edge on the image). The rest of the process remains the same.

Once we have the skin region, a little optimization is accomplished: those skin regions with a negligible size compared to the total size of the image are discarded in order to avoid unnecessary processing. In most cases, these small regions correspond to slight inaccuracies in the skin detection task or to tiny skin areas that were disconnected from the rest of the region. Specifically, we discard those regions having an area smaller than the 0.1% of the total area of the image. Finally, we obtain skin regions to be classified as the different human body parts. All these steps are exemplified over a face image in Figure 2.21.

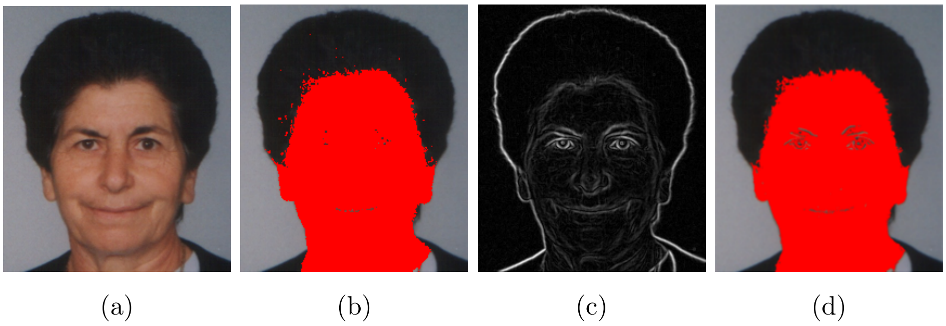


Figure 2.21: Skin pixel classification. (a) Original image. (b) Classification using only the *skinthreshold*. (c) Edge information. (d) Final classification considering the edge information and discarding the regions with negligible size.

Hand raising detection

The first step towards the hand raising detection is the detection of valid skin regions. Since we are focused in a specific domain and the conditions under which the test is performed are quite stable, we can define search areas with high probability of appearance of the hand within them in order to facilitate the detection. As it can be seen in Figure 2.22, in the lower third of the image is where the audiometer is located, so this is not a skin search area. Furthermore, since we know the location of the head (addressed in the first step of the methodology) we can use this information as reference. Hands are going to be raised to one side or another of the head, but never over or under it. This way, we can define the search areas *a* and *b* (both highlighted in Figure 2.22) which are defined relative to the face location. This delimitation of the search areas reduces the computational cost and it also avoids incorrect detections.

Skin pixel classification and connectivity analysis previously explained are applied into the two search areas (*a* and *b*). Once we obtain the skin regions, we consider domain knowledge information that allows the establishment of a set of conditions that a raised hand should fulfill. The first one is related to the required number of skin pixels inside the region to be considered as a hand candidate: at least 40% of the pixels contained inside the region must be skin pixels. For the second rule we define:

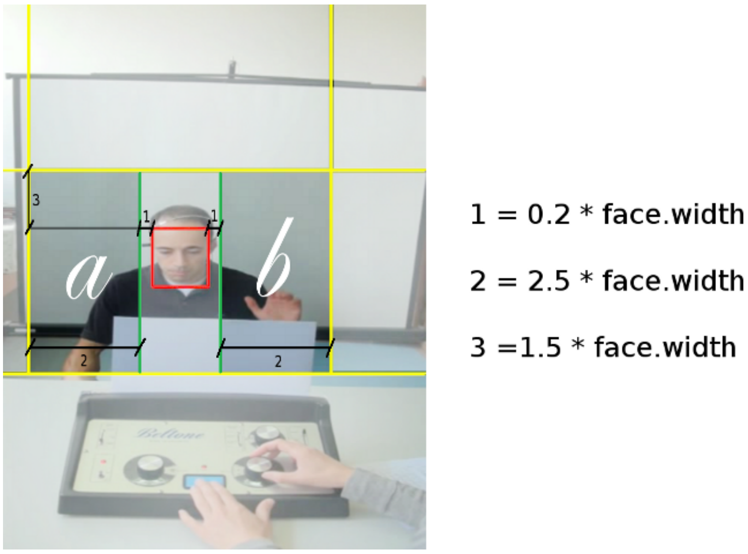


Figure 2.22: Search regions for skin detection: *a* and *b*.

$$\begin{aligned}&\text{limSup = head.y - (0.25 * head.y);}\\&\text{limInf = head.y + head.height + (0.25 * head.y);}\end{aligned} \tag{2.1}$$

where *limSup* is the upper limit of the hand position and *limInf* is the bottom limit of the hand position, where *head.y* represents the *y* coordinate of the upper end of the head and *head.height* represents the height of the head. Next, a region is discarded if:

$$\text{(skinRegion.y} < \text{limSup) or (skinRegion.y} > \text{limInf)} \tag{2.2}$$

where skinRegion.y represents the *y* coordinate of the upper end of the skin region evaluated at the moment. This means that, if the upper end of the considered skin region is located above or below the location of the face, most likely it will not correspond with a hand.

The size of the skin region is also taken into account. A region is also discarded if some of these conditions apply:

$$\begin{aligned}&\text{(a)}\quad \text{skinRegion.height} < 0.42 * \text{head.height}\\&\text{(b)}\quad \text{skinRegion.width} < 0.3 * \text{head.width}\\&\text{(c)}\quad \text{skinRegion.width} > \text{head.width} + 0.25\text{*head.width}\end{aligned} \tag{2.3}$$

meaning that if the skin region is very small regarding to the size of the head, or also in the case it is very much wider than the head, then, it can not correspond to a hand.

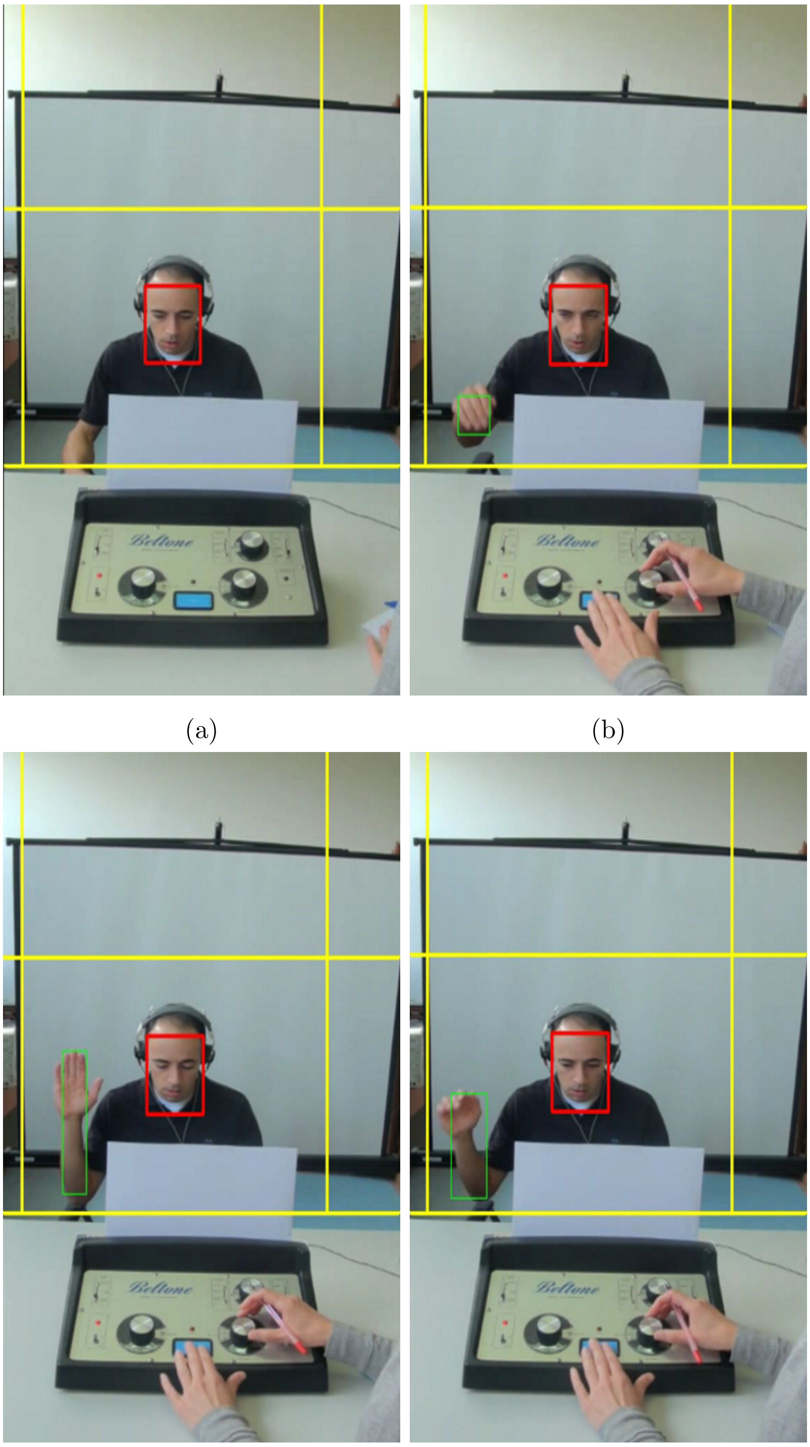
Finally, the remaining regions are classified as hands, and we can proceed to analyze the responses. In Figure 2.23 some examples during the hand raising detection can be observed. They show how the proposed method detects all the stages during the hand raising procedure: prior to the perception of an auditory stimulus the patient is waiting for the sound with his hand down (Figure 2.23(a)), then, when he perceives the auditory stimulus he starts to raise his hand (Figure 2.23(b)), going through the moments when the hand is fully raised (Figure 2.23(c)), and ending with the descent of the hand (Figure 2.23(d)) until returning to the initial position (Figure 2.23(a)).

2.5 Analysis of the responses to the stimuli

Once both the stimuli and the hand raising are detected, the next step according to the schema shown in Figure 2.5 is the analysis of the responses to the auditory stimuli. Combining the information obtained in the previous stages of the methodology we can analyze the response and some properties of interest about this patient's response.

For each auditory stimulus delivered by the expert we know: the moment when it starts, the moment when it ends, and accordingly, its duration (Figure 2.24(a) represents the signal for the auditory stimuli). The length of a stimulus is a variable parameter that depends on the expert. Through the evaluation of the different audiometric video sequences recorded it was observed that in cases where the expert already knew the patient and this patient was a person with slow responsiveness, the expert sent longer stimuli. On the opposite case, if the expert knew that the patient is a person with fast responsiveness, the stimuli sent were shorter. The length of the stimulus was also influenced by the slowness (or quickness) of the patient's movements. The stimuli sent to patient with parsimonious movements were longer that the stimuli sent to quick patients.

The hand raising could be characterized not only by the moment when the patient raises his hand, but also the height at which he does it. Figure 2.24(b) represents the signal for the hand raising detection, where the pulse grows equivalently to the height at which the patient raises his hand. By the combination of these two features that we store for each frame of the hand raising, we would able to determine different



(a) (b)

(c) (d)

Figure 2.23: Captions of the hand raising detection. (a) Hand down. (b) The patient starts to raise his hand. (c) Hand fully raised. (d) The patient begins to low his hand.

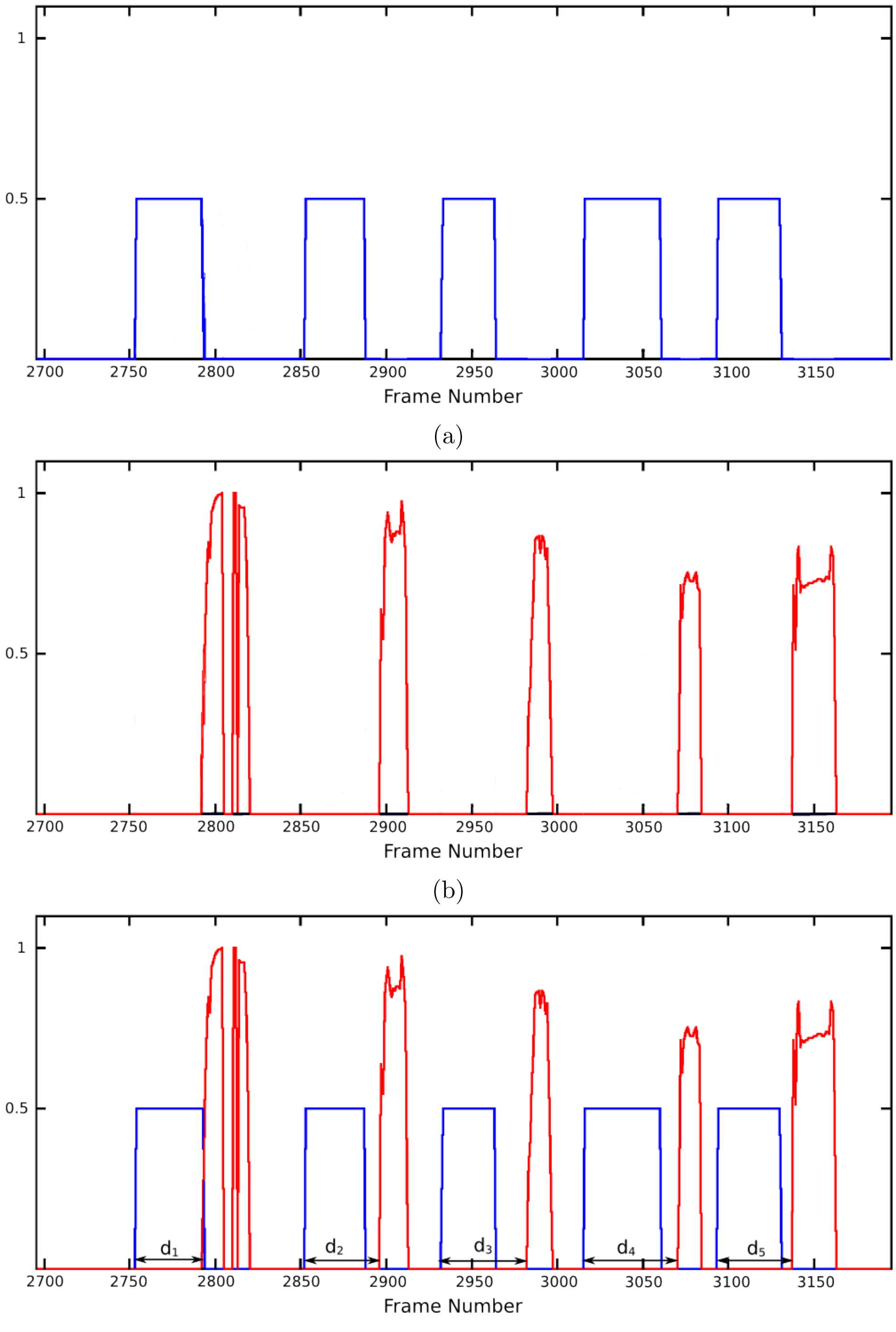
factors such as: the moment when the patient begins his response to the stimulus, the speed of his hand raising, how long he has his hand up, the maximum height reached by his hand, the speed with which he puts down his hand, and other possible features that would characterize his response.

Audiologists reported that one of the most relevant informations that can be derived from the analysis of this test is the evaluation of the patient's reaction times. Based on the impressions that the experts carry out during the performance of the audiometric evaluation, patients are classified according to their speed of response as "patients with a normal responsiveness" or "patients with a slow responsiveness". It is relevant for the audiologist to identify people with a behavior particularly slow compared to what is is considered as the average behavior since this could be a symptom of other pathologies such as inefficient central brain processing or also to lack of motivation.

Experts stated that the response times should be measured from the moment the auditory stimulus starts until the moment when the patient's reaction starts, or, said in a different manner, a reaction time is defined as the interval elapsing between the stimulus and the response. Considering this, we are going to combine the information from the auditory stimuli signal (Figure 2.24(a)) with the information from the hand raising signal (Figure 2.24(b)), and, this way, we compute the response times as the difference between the two initial moments as it is represented in Figure 2.24(c). Consequently, a patient will be characterized by a sequence of n distances d, where n is the number of reaction times measured for this particular patient, and each one of the d distances is a reaction time.

Figure 2.25 shows some examples of the correspondence between different moments during a sequence and the signal obtained by the system for that particular moment. In Figure 2.25(a) the auditory stimulus has already been sent, but the patient has not yet begun his reaction. A few frames later, in Figure 2.25(b), the patient he perceives the stimulus and he quickly raises his hand. In Figure 2.25(c) the patient maintains his response, and finally, in Figure 2.25(d) he starts to lower his hand.

With the proposed methodology we provide not only the detection of the patient's positive responses to the sound, but also a precise and objective measurement of his reaction times for each one of the auditory stimulus perceived. A set of experiments measuring the effectiveness of this approach and proposal of different metrics to classify patients according to their response times are addressed in the next Section.

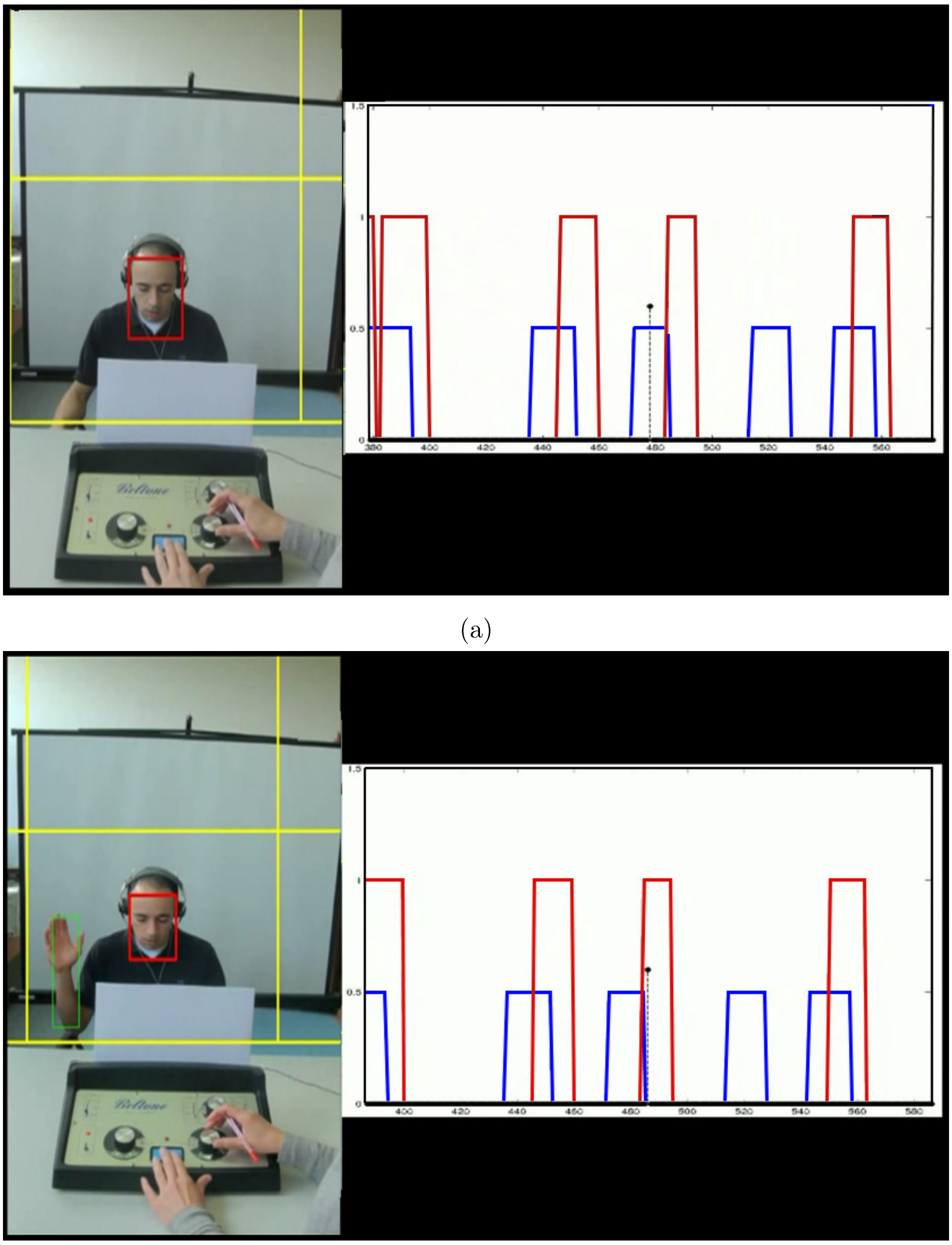


(a)

(b)

(c)

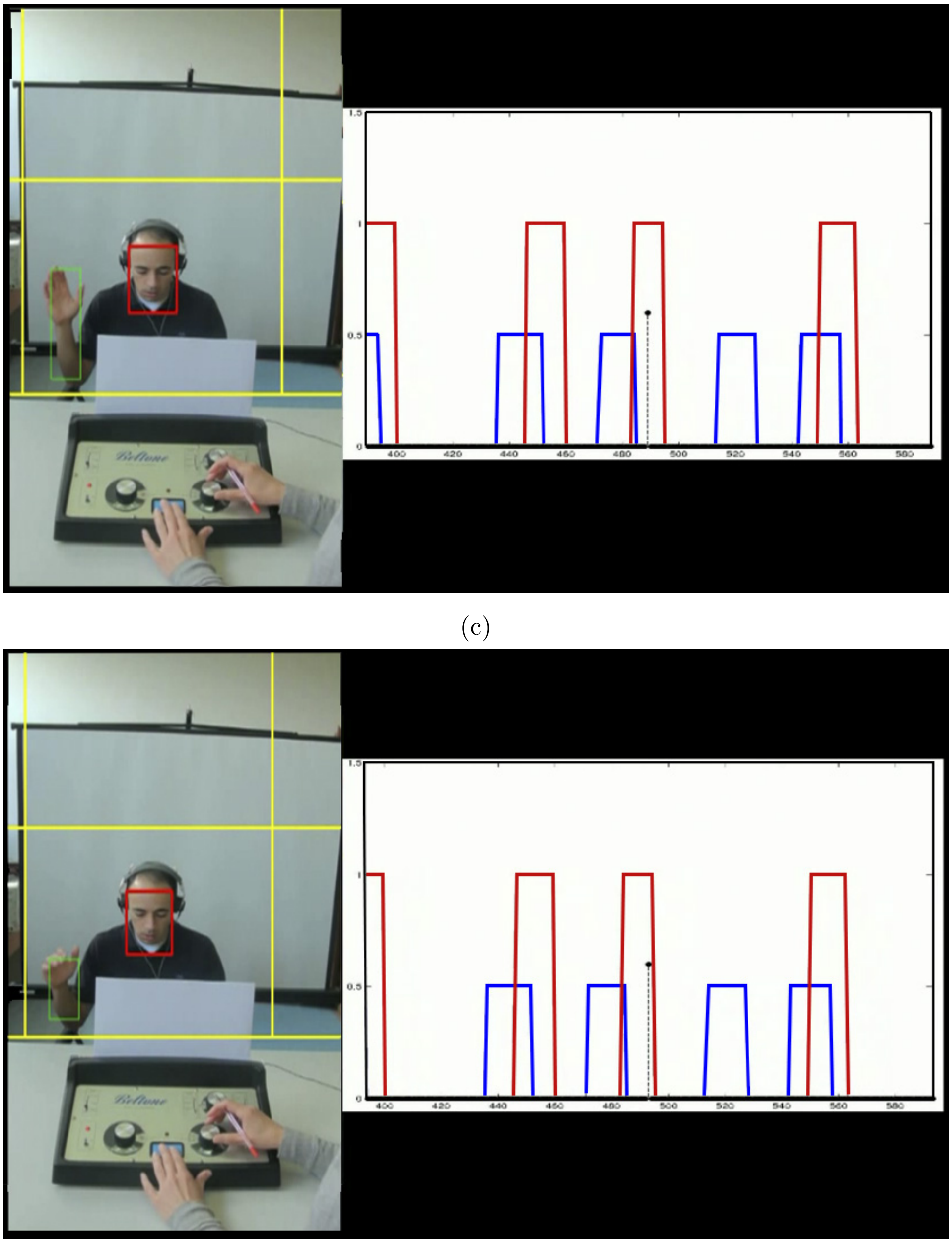
Figure 2.24: Measurement of the response times in a partial sequence. (a) Auditory stimuli signal. (b) Hand raising signal (where value 1 in the vertical axis is the maximum high at which the hand is raised by that patient). (c) Combination of (a) and (b) and measurement of the response times



(a)

(b)

Figure 2.25: Correspondence between video sequences and detection signals. (a) Auditory stimulus sent but not perceived. (b) The reaction starts. (c) The reaction continues. (d) The patient begins to low his hand.



(c)

(d)

Figure 2.25: Correspondence between video sequences and detection signals. (a) Auditory stimulus sent but not perceived. (b) The reaction starts. (c) The reaction continues. (d) The patient begins to low his hand.

2.6 Experimental results

For the evaluation of this methodology a dataset provided by the audiologists containing 77 video sequences recorded during the performance of audiometric evaluations has been considered. Patients undergoing this test were randomly selected, so they represent a sample from a normally distributed population. They are both male and female, with different ages. No significant hearing loss was known before the conduction of the test. Besides, for most of them, this was their first audiometric evaluation, so the were not conditioned by previous experiences. Since this was the first hearing assessment for most of them, the audiologist had not prior information about their hearing thresholds. According to the established clinical protocol each ear should be evaluated separately, thus, each video sequence corresponds to the evaluation of one ear. For this experiment, video sequences were recorded using a frame rate of 12 FPS (frames per second).

Each of these video sequences takes between two and four minutes, so, considering a frame rate of 12 FPS, it involves the evaluation of between 1500 and 3000 different frames per video sequence. At each of these video sequences between 30 and 60 auditory stimuli are delivered to the patient. Among these 30-60 auditory stimuli, the patient usually perceives only between 15 and 35, and he positevely responds to them by raising his hand. Since the purpose of the audiologist is to test the patient's hearing capacity, he systematically delivers different auditory stimuli looking for the lowest hearing level (in dB) that the patient is able to perceive. That is why not all the auditory stimuli will be answered, since some of them are too low for the patient to notice, but the expert should try them anyway in order to find the softest sounds the patient is able to perceive across a range of frequencies.

The provided video sequences are consistent in location but they are not consistent in illumination conditions. They had been recorded on different days and at different hours during the day, therefore, there is significant variability in the brightness conditions. There are video sequences with more light intensity than others (in most cases, due to the outside weather conditions), and also video sequences with natural light versus other video sequences with artificial light.

Experimental results presented throughout this section are divided into two parts: an initial validation of the detection techniques (the detection of the stimuli delivery and the detection of the hand raising), and a second one where we study the measurements obtained by the proposed methodology. But first of all, it is necessary to deal with the problem of the illumination changes.

2.6.1 Adaptation to illumination changes

It was observed that changes in illumination conditions between day scenes and night scenes with artificial light were very abrupt. This causes that the skin detection becomes imprecise when working with a fixed threshold. This situation suggest the use of an adaptive threshold dependent on the illumination of the scene for a better behavior of the methodology.

To compute threshold values we put a white screen as background behing the patient with the idea of using it as an indicator of the brightness of the scene. To that end, we take as reference the brightness component L of the white screen that appears as the background of the scene. This white screen is virtually divided into two halves, computing the average of the L component of the left side as *avg_white_left*, and equivalently to right side as *avg_white_right* (as represented in Figure 2.26). The *skinthresholds* defined taking into account these considerations are shown in Table 2.1.

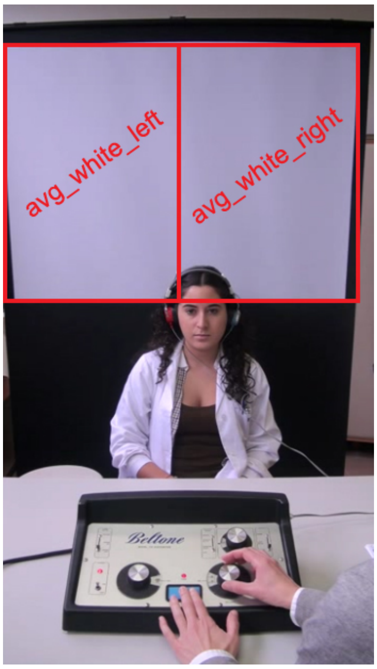


Figure 2.26: Estimation of image brightness

Table 2.1: *Skinthreshold* values

1) avg_white_left $\geq$ 185

    1.a) if (diff $>$ 15) $\longrightarrow$ *skinthreshold* = 0.0006;

    1.b) if (diff $\leq$ 15) $\longrightarrow$ *skinthreshold* = 0.00125;

2) (avg_white_left $\geq$ 170) and (avg_white_left $\leq$ 170)

    2.a) if (diff $\geq$ 28) $\longrightarrow$ *skinthreshold* = 0.0006;

    2.b) if (diff $\geq$ 20) and (diff $<$ 28) $\longrightarrow$ *skinthreshold* = 0.00125;

    2.c) if (diff $<$ 20) $\longrightarrow$ *skinthreshold* = 0.0005;

3) (avg_white_left $\geq$ 150) and (avg_white_left $\leq$ 165)

    $\longrightarrow$ *skinthreshold* = 0.00125;

4) avg_white_left $\leq$ 150

    $\longrightarrow$ *skinthreshold* = 0.004;

where diff = avg_white_left - avg_white_right

2.6.2 Validation of the detection techniques

The accuracy of the methodology has been tested for both the detection of the light stimuli indicator (presented Section 2.3) and the detection of the hand raising (presented in Section 2.4.2). To this end, a subset composed by 10 video sequences has been manually labeled and then evaluated. Results are presented in Table 2.2. The detection methodology offers a satisfactory rate of success to avoid the loss of patient's reactions.

Table 2.2: Accuracy of the methodology

	Stimuli	Hand raising
Number of events	479	288
Number of detected events	479	286
Accuracy	100%	99.31%

2.6.3 Measurements of response times

Once the accuracy of the detection methodology has been validated, the next and most important step is to apply the proposed method over all the video sequences at our disposal in order to test the quality of measurement of the patients' response times.

For each of the video sequences in our dataset we obtain a sequence of n response times, depending on the number of positive reactions manifested by the patient.

The speed of response of the patient must be evaluated taking into consideration all these response times. To this end, five different metrics were evaluated to summarize the responses of a patient into a single measure. It it important to clarify at this point that the non-reactions to the auditory stimuli are not considered since they are already manually registered by the audiologist, and at this stage we are only interested on measuring response times for positive reactions, which only occur when the patient correctly perceives the auditory stimuli.

To perform this experiment patients were selected completely randomly, which should imply that our test population corresponds to a sample from a normally distributed population. A series of auditory stimuli was delivered to these patients, and their response times were automatically measured with the proposed methodology. Independently, patients were classified by the experts as “patients with normal response times” and “patients with slow response times”. The target of this study is to determine if the labels assigned by the experts are consistent with the measurements obtained by the methodology.

The measurements obtained for each patient and each one of the five metrics are detailed in Table 2.3. The first metric (Mean in Table 2.3) is the mean and the second one (Median) represents the median, for the whole set of the patient’s measures. The third one (Mean(Q1-Q3)) is the average of all the elements between the first and the third quartile. The next two measures are defined considering that response times of less than half second (less than 6 frames) can not occur but for an error in measurement; and considering also that the highest times of the sequence may be due to moments of doubt or they may be occur because at the beginning the patient does not know what he has to perceive and takes longer to respond; this will cause that it may exist times higher than the average, but they are not representative because the real speed of response of the patient is the one that he shows when there is no doubt and the behavior is normal. As a result, in Gt6-15% measure the response times of less than 6 frames are removed, and to discard spurious high times we will remove the 15% of the largest times of the sequence; with the remaining elements we compute the mean. For Gt6-25% measure, we also removed times lower than 6 frames and the 25% of the largest times of the sequence, as in the previous case, with the remaining elements we compute the mean. The corresponding response times for all the video sequences are detailed in Table 2.3, it should be noted that all times are shown in number of frames (the video sequences have a frame rate of 12 FPS).

Table 2.3: Comparative table of the different measures for the response times

Video	Mean	Median	Mean(Q1-Q3)	Gt6-15%	Gt6-25%
001	33.6522	35	33.5333	31.7368	29.3750
002	24.8333	25.5	23.5000	20.9000	19.7778
003	14.8788	10	10.7727	11.2400	10.3182
004	10.5000	10	10.1176	9.5000	9.1053
005	12.9167	12	11.4000	11.3000	10.7778
006	17.2812	16	15.2000	14.8148	14.0833
007	14.8387	14	14.8947	13.3462	12.7391
008	17.0769	14.5	15.1176	15.0909	14.0526
009	21.3214	22	21.5000	20.8261	20.3000
010	23.4500	24	23.7273	21.5294	20.5333
011	11.5357	10.5	10.1111	10.1250	9.8095
012	12.2692	11	11.4000	11.1818	10.5789
013	11.8571	12	11.5000	11.3478	10.9000
014	13.0455	12	12.3846	11.9444	11.5000
015	16.5357	14.5	15.4286	14.7083	13.5714
016	11.7308	11	11.4000	10.7727	10.1053
017	15.9600	14	14.4375	14.3810	13.6842
018	15.5238	14	14.7857	13.7222	13.0625
019	10.0000	10	9.2105	9.5000	8.8235
020	11.5500	9.5	10.1429	9.7647	9.3333
021	11.3500	10	10.0667	10.0667	9.6154
022	12.0000	10	10.4667	10.7500	10.2143
023	10.1538	10	9.6316	9.3810	9.2105
024	9.8000	8	8.5000	8.6842	8.1875
025	10.8148	10	9.8889	9.6818	9.1053
026	11.7083	11	10.8125	10.6000	10.2222
027	12.2632	12	12.1429	11.1875	10.7143
028	14.6154	15	14.1000	13.7273	13.4000
029	10.4400	10	10.2143	9.8000	9.5000
030	12.8077	12	11.9412	12.0526	11.1875
031	11.3333	10	9.9500	10.3043	9.6500
032	11.5128	11	10.9310	10.2424	9.8966
033	11.0000	9	9.6842	9.8000	9.3182
034	15.5862	15	15.4211	14.8333	14.3333
035	14.0588	14	14.1429	13.6923	13.3478

Video	Mean	Median	Mean(Q1-Q3)	Gt6-15%	Gt6-25%
036	14.7143	14	13.7391	13.5833	12.9524
037	13.8621	12	12.1579	12.6000	11.8636
038	10.7436	10	9.8889	9.6875	9.1071
039	11.5556	11	10.9524	10.3913	10.0000
040	13.2593	13	12.9412	12.3478	11.9500
041	14.8387	14	14.4000	13.7600	13.1364
042	11.2800	11	11.1111	10.6190	10.3684
043	12.3704	12	12.1176	11.4348	11.0500
044	12.1905	12	12.1875	11.7222	11.4378
045	11.6579	10.5	10.7391	10.8333	10.0385
046	14.3611	12.5	14.2273	12.8333	12.0385
047	13.1765	13	12.5833	12.2143	11.7200
048	11.2750	11	11.3214	12.4400	11.7727
049	12.5128	11	11.1111	10.9032	10.3212
050	13.6786	13.5	14.0556	13.2273	12.7895
051	13.2667	13	12.9412	12.5600	12.0455
052	15.6296	12	13.2500	14.6842	13.2500
053	12.7308	12	12.1111	11.7273	11.3158
054	14.6552	14	14.0556	12.9200	12.2273
055	21.9375	23.5	22.8889	21.0714	20.3333
056	13.3333	12	12.1739	11.8065	11.0741
057	12.2647	11	10.6818	11.0000	10.3750
058	10.5625	10	9.6818	9.8077	9.3043
059	13.5717	13	13.0000	12.4000	12.0000
060	12.3750	11	10.6190	10.2692	10.0000
061	12.8684	11	11.7917	11.4839	10.4444
062	15.5172	15	14.8889	14.3200	13.7273
063	11.7895	12	11.6190	11.4483	10.5200
064	9.0000	9	9.0500	8.5417	8.2381
065	10.9615	9	9.8889	9.9048	9.3157
066	11.5200	11	10.7778	10.6190	10.3158
067	13.5600	13	12.9286	12.3810	11.8947
068	13.9615	13	13.2000	12.9091	12.3684
069	13.1852	12	12.3500	12.2609	11.8000
070	13.8889	13	13.8947	12.7826	12.1500
071	13.8000	12	12.6500	12.4286	11.8421

Video	Mean	Median	Mean(Q1-Q3)	Gt6-15%	Gt6-25%
072	14.7000	13.5	13.7895	13.1600	12.3182
073	13.1071	12	12.2500	11.2917	10.7143
074	11.3333	11	11.3077	10.2222	9.7500
075	12.6667	11	11.5000	11.1500	10.6111
076	12.6667	12	12.1579	11.5217	11.1500
077	13.0000	13	12.7059	11.5200	10.7727

In order to see these results in a more visual way, in Figure 2.27 we show the typical box and whisker plot for all the metrics. In these graphics, the bottom and top of the box are always the 25th and 75th percentile (the lower and upper quartiles, respectively), and the band near the middle of the box is always the 50th percentile (the median); the whiskers extend to the most extreme points the algorithm does not consider as outliers, and the outliers are plotted individually. According to this definition, our method spots five outliers for each one of the five measures. By consulting the experts, it is confirmed that the five detected outliers correspond to the "slow" patients in our test population. Thus, a 100% agreement with the audiologists is achieved, confirming that the automatic measurements correspond with the same rating that the experts establishes.

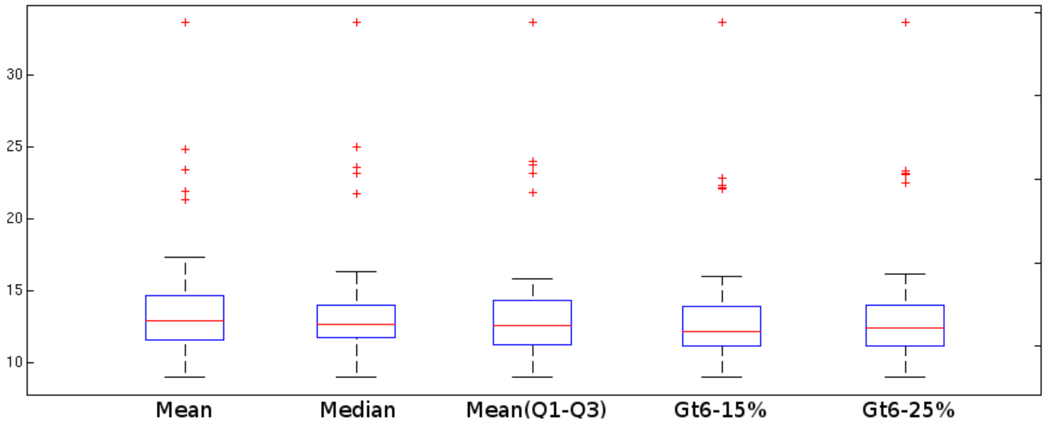


Figure 2.27: Box and whisker plot for all the metrics

Moreover, although we may not have enough samples, it would be interesting to carry out a preliminary study in order to determine if "normal" patients correspond to a normally distributed population. Towards determining if the obtained results for the normal patients can be approximated by a normally distributed population we are going to use the Lilliefors test. The Lilliefors test is a 2-sided goodness-of-fit test suitable when a fully-specified null distribution is unknown and its parameters must be estimated. The obtained results for this test allow us to accept the default null hypothesis at the 5% significance level for all measures but the Median measure. Except for the Median measure (whose normality has been rejected), the other measures obtained high p-values (where the lower of these p-values is of 0.1338), that allow us to accept the hypothesis of normality. Considering these results, it can be established that the response times of the healthy population come from a Gaussian distribution.

These results justify the experiment, in which the starting point consisted on taking individuals from a normal population (which is approximated by a Gaussian function). Since the patients were randomly taken from a normal population, the results approximate the same way to a normal population (with a Gaussian distribution).

In order to show the capacity of the methodology for distinguishing between "slow" and "normal" patients, probability density functions (pdf) were calculated. Results for all the measures are shown, except for the Median measure which showed not to be normally distributed. Calculating the probability density function for the slower measure from the "normal patients" versus the probability density function for faster measure from the "slow patients", it can be observed the capacity of each one of the measures for distinguishing between both types of patients. These values are presented in Table 2.4, showing for all cases that the distance between both types is wide enough. These results will allow the choice of any of these measures and the establishment of a threshold in future works, where we can have a wider selection of video sequences.

Table 2.4: Distance between extremes

	Mean	Mean(Q1-Q3)	Gt6-15%	Gt6-25%
Slower "normal"	1.2e-3	8.9e-3	8.6e-3	1.08e-2
Faster "slow"	5.4675e-10	2.2365e-12	1.7271e-12	6.4593e-12

2.7 Discussion

We have presented here a screening method to automatically measure the response times during the performance of an audiometric evaluation. One of the premises of this work consisted on trying to modify as minimum as possible the traditional protocol of the audiometric assessment, for this purpose, we adapted the proposed methodology to the way the audiologist proceeds during the assessment and also to the audiometer that he uses.

The auditory stimuli light indicator was properly detected (for two different models of audiometers), in order to synchronize stimulus and response. It is necessary to assume that we have a dependency on the device which implies that this method is not as general as it would be desirable; however, the adaptation to the specific device is quite simple and it can be easily included within the methodology. So that, either the method is combined and synchronized with a digital device or an analogical device it would be necessary and ad-hoc method for detecting the signal adapted to each particular device.

For the detection of the hand raising, skin color information was used. Since when working with color filtering the proper choice of the color space becomes very relevant, different color spaces were evaluated for this specific task. From this survey it was concluded that the TSL color space was the most suited one for our skin color detection. By the combination of both auditory stimuli and hand raising detection, the method finally provides a precise and objective measure of the patient's reaction times.

Throughout Section 2.6 the obtained results for the evaluation of the methodology are depicted. The obtained results for the detection and measurement of the response times are highly promising. Furthermore, we have studied several metrics in order to combine the patient's response times to the different auditory stimuli to which he responds during the session. It is noted that, although every measure has its peculiarities, all of them allow us to establish a gap between the "normal" patients and the "slow" ones.

The main strengths of this method begin by highlighting its capacity as screening method in order to objectively identify patients abnormally slow (those in which the experts are concerned because they could suffer other cognitive problems or pathologies), and it continues by offering to the experts the possibility of precisely quantify the patient's speed of response and to carry out more detailed studies using this information. It is important to note that the impact of the method is not only the detection of slow people by itself, but also the possibility of objectively and precisely measure their response times in order to establish comparisons or to conduct different clinical studies.

The limitation of this proposal arrives when the audiologist needs to evaluate patients with cognitive decline or severe communication disorders. With this group of patients it is not possible to maintain a conventional interaction and it is highly unlikely that they follow the instructions given by the expert, so a hand raising response should not be expected. Although the audiological evaluation of these patients becomes much more complex, it is still possible if the expert focus his attention on the detection of spontaneous gestural reactions to the sound. This problematic will be addressed across the next Part of this thesis.

We shall conclude this discussion by emphasizing that the proposed methodology has proven to be valid as screening method to objectively and precisely measure reaction times of a patient during the performance of an audiometric evaluation. This provides not only an objective measurement method for comparison and evaluation between experts, but also a system for the detection of those patients who have an abnormally slow response times.

Chapter 3

Web application

In addition to the proposed methodology a web application has been developed in order to provide an interface that facilitates the use of the automatic methodology by the audiologists.

The web application has restricted access through authentication, and it is simple and easy to use for the audiologists. It allows the audiologist to easily manage all the information related with his patients and their examinations. Furthermore, through this tool they will be able to easily manage and to process the video sequences associated to each patient. The application has been internationalized to Spanish, English and Galician.

The application is managed by an administrator which is responsible for singing up new audiologists and new clinical institutions. The administrator can manage the audiologist and institutions but for reasons of Data Protection he is not allowed to access to the patient's information (unless the audiologist voluntarily allows him in order to help him to solve an specific problem).

Once the audiologist has been signed up into the application by the administrator, when he needs to access to the web application he must enter his user name and password in the authentication screen (see Figure 3.1). There is also an option in case the audiologist forgets his password (option "I forgot my password"). By using this option, he will receive an email in the email address he has specified that will allow him to enter to the web application and to restore his password. The "Remember me' option could be selected by the audiologist in order to remain logged in as long as it has open the web navigator, otherwise, the session will expire after several minutes of inactivity.

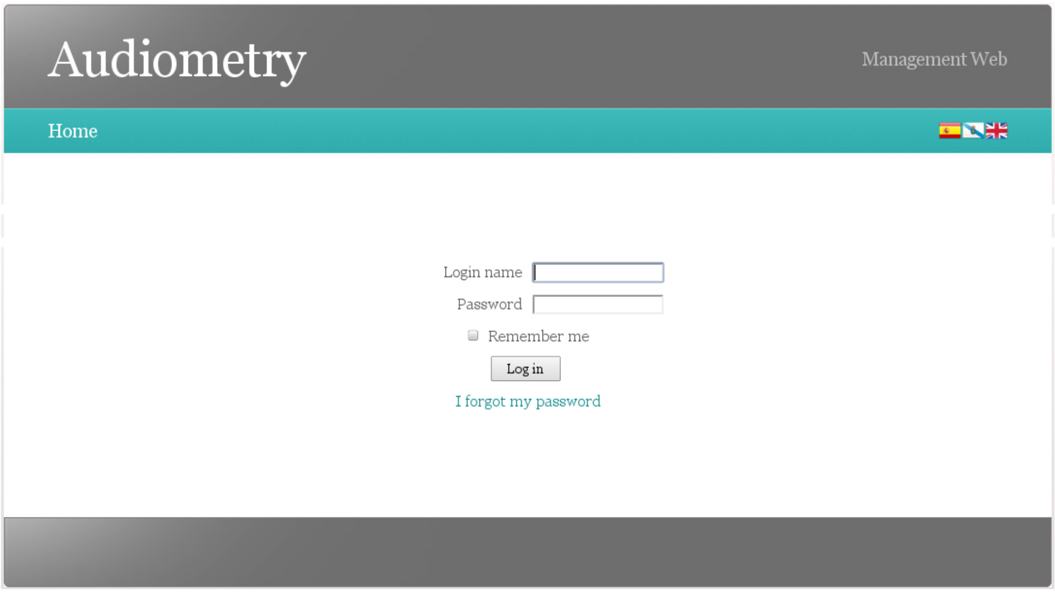
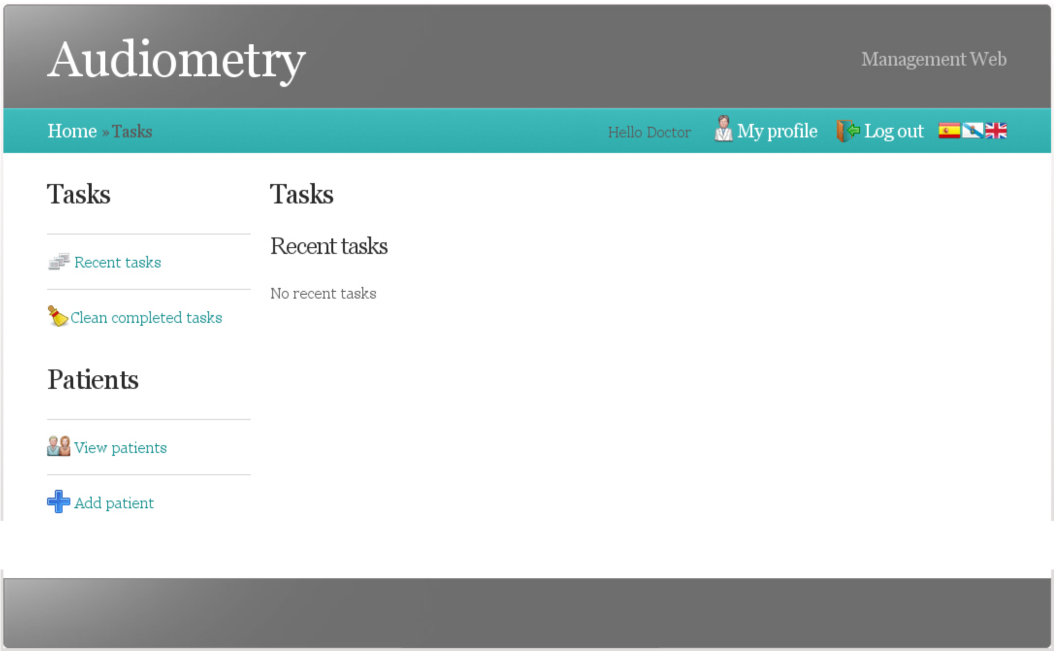


Figure 3.1: Authentication screen.

Once he has logged in into the system, the audiologist will be directed to the Recent Tasks screen (see Figure 3.2). At this screen it will be displayed the last tasks the system has processed for him, for example: a new video sequence has been recently uploaded, an uploaded video has been processed by the system, if an error has occurred while uploading or processing a video sequence, etc. This information will allow him to know which are the last video sequences that he has evaluated and had been already processed by the system or if any problem has recently occurred. There is an option on the left sidebar that allows the user to clean the completed tasks, by choosing this option, all the recent tasks will be automatically removed from this screen.

Through option My profile the audiologist can manage and modify his account information: name, surname, email address and phone number. It is important to maintain updated the email address since it is the email account used in the case the user forgets his password. Through My profile option the audiologist can also establish a new password. To that end, and for security reasons, it is necessary to introduce first the actual password which will be checked before allowing the password modification.

As mentioned before, each one of the audiologists must be initially registered by the administrator on the web application. When the administrator user reg-



Audiometry

Management Web

Home » Tasks

Hello Doctor My profile Log out

Tasks

Recent tasks

Clean completed tasks

Patients

View patients

Add patient

Tasks

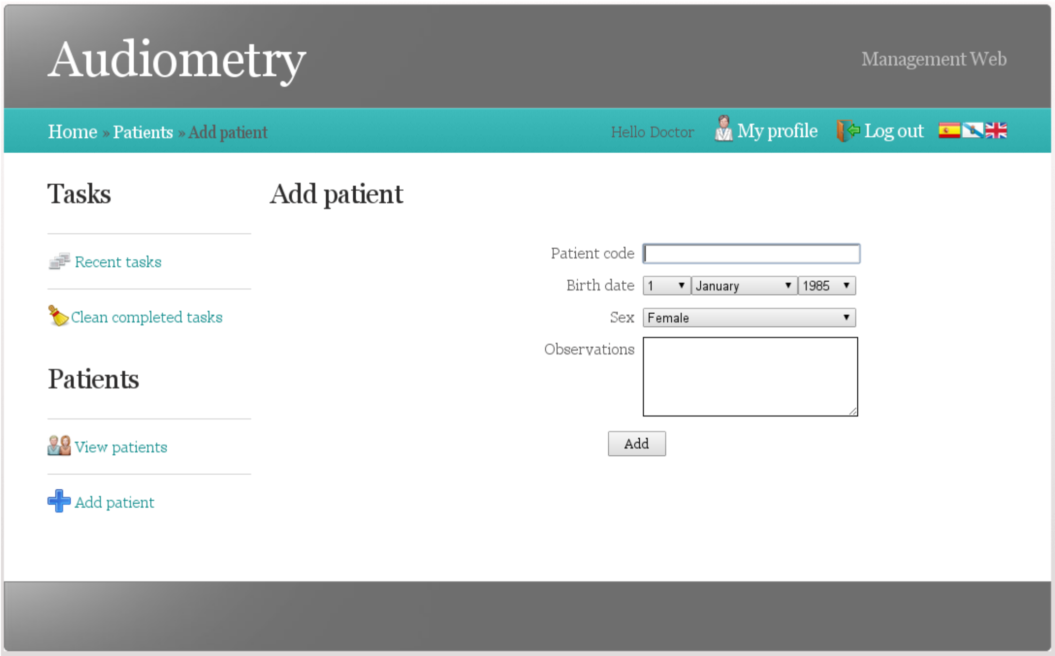
Recent tasks

No recent tasks

Figure 3.2: Recent tasks screen

isters the audiologist he associates him to an specific institution (institutions can only be created by the administrator too). In order to sign up an institution the administrator should introduce a name, a location and a contact phone number. The relation audiologist-institution is a 1-to-1 relation; it means, at a determined time, one audiologist can only belongs to one institution. However, it is possible that one audiologist has belonged to different institutions at different moments in time. The information of the institution and his relation with it can not be directly modified by the audiologist, only the user with administration priviliges can apply this modification.

In relation with patients, an audiologist can add a new patient through option Add patient. In this case, he will be redirected to the Add patient page (presented in Figure 3.3). To register a patient, the audiologist needs to enter a patient code, and also the birth date and the sex of the patient. Patients are identified by a patient code and not by name in order to keep the patient's information anonymous in the web application for Protection Data reasons. Sex (male or female) and birth date are introduced by using a selector. The audiologist can also include any type of observations about the patient using a text field. All the fields (except the observations field) are mandatory.



Add patient

Patient code:
Birth date: 1 January 1985
Sex: Female
Observations:

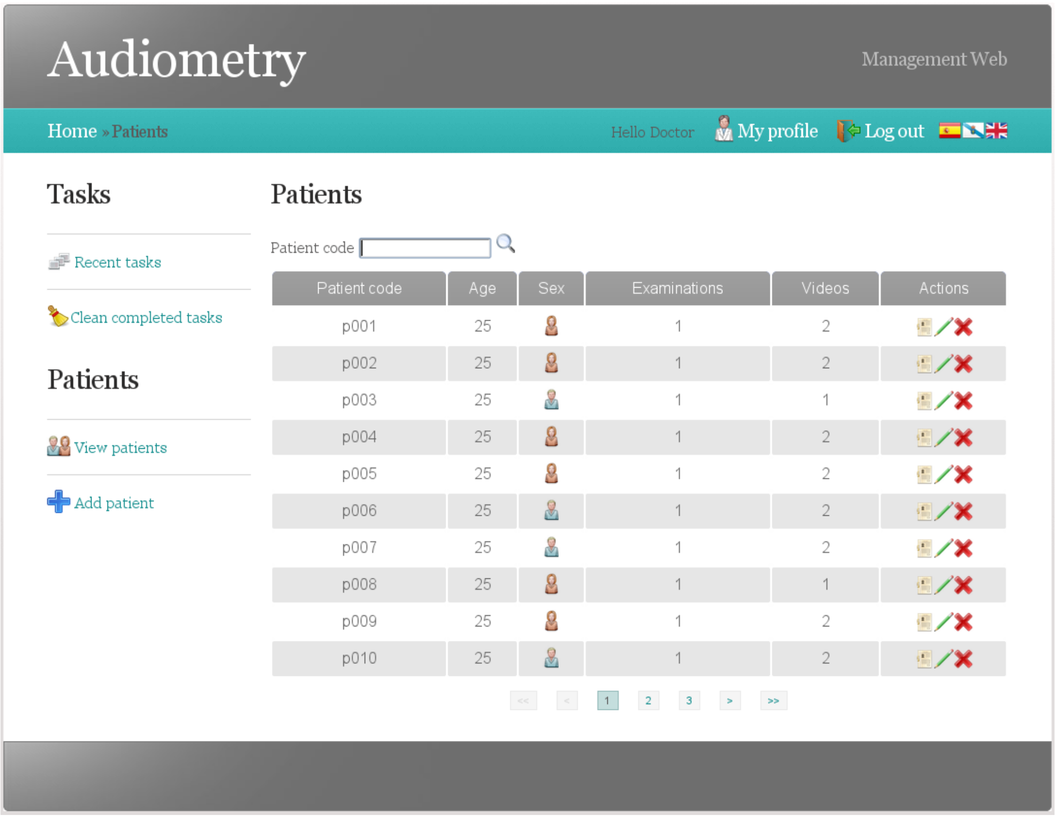
Figure 3.3: Add patient screen

Through option View patients the audiologist can access to the complete list of his patients ordered by patient code. The audiologist accesses to a paginated list where 10 patients are displayed at each page (see Figure 3.4), different buttons will allow the access to the next and previous pages (specific buttons for the two next/previous pages, a button for the next/previous page and another one for direct access to the last/initial) page). At this screen he also has search bar where he can search for a patient by using the patient code. For each patient he is going to see the most relevant information: patient code, age, sex, number of examinations registered, and number of videos for that patient. Moreover, he has three available options: see report, update and delete.

Through option Update the audiologist will be able to modify the patient's information. He will be redirected to a page similar to Figure 3.3 where he can modify the fields: birth date, sex and observations. The only restriction is that the patient's code can not be modified by the audiologist since it is used as key value in the database.

Action Delete will allows the audiologist to delete a patient from the web application. Once he pushes the delete button a confirmation message appears in order to confirm if he really wants to delete the patient or not. If he confirms the order, all the examinations and video sequences associated to that specific patient will be automatically deleted by the system.

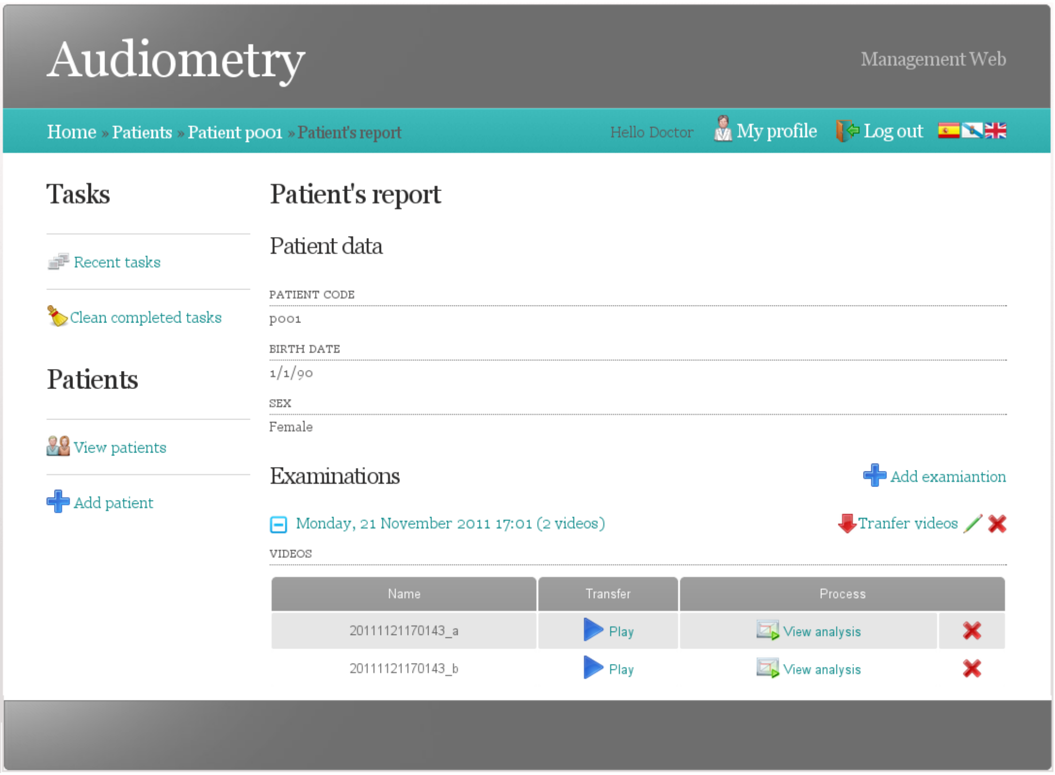
Finally, through option See report the audiologist will be able to access to patient's audiological report. The patient's report screen can be observed in Figure 3.5. At this screen he will see the patient's information: code, birth date, sex, and observations (if they were registered by the audiologist). Next, all the examinations registered for that particular patient will be displayed. The examinations are identified by date, and the number of video sequences associated to that examination is also detailed in the title. For a particular examination three actions are available: transfer videos (which allows to associate new video sequences to the examination), update (which allows to modify the examination's information), and delete (which displays a confirmation message before removing the examination and it has as consequence the removal of the associated video sequences). Once a video sequence has been transfered to the web application it will be automatically processed by the system.



Patient code	Age	Sex	Examinations	Videos	Actions
p001	25	♀	1	2	📄 ✎ ✗
p002	25	♀	1	2	📄 ✎ ✗
p003	25	♂	1	1	📄 ✎ ✗
p004	25	♀	1	2	📄 ✎ ✗
p005	25	♀	1	2	📄 ✎ ✗
p006	25	♂	1	2	📄 ✎ ✗
p007	25	♂	1	2	📄 ✎ ✗
p008	25	♀	1	1	📄 ✎ ✗
p009	25	♀	1	2	📄 ✎ ✗
p010	25	♂	1	2	📄 ✎ ✗

Figure 3.4: View patients screen

By displaying a particular examination the audiologist will be able to see the associated video sequences. For each of the videos three actions are available: play, view analysis and delete. Option Play reproduces the original video as it was uploaded by the expert. Though Delete option the video sequence will be removed (a confirmation message will be displayed before removing the item). And option View analysis shows the processed video sequence (the information displayed for this option will be discussed next).



Audiometry

Management Web

Home » Patients » Patient p001 » Patient's report

Hello Doctor · My profile · Log out

Tasks

Recent tasks

Clean completed tasks

Patients

View patients

Add patient

Patient's report

Patient data

PATIENT CODE
p001

BIRTH DATE
1/1/90

SEX
Female

Examinations · Add examiantion

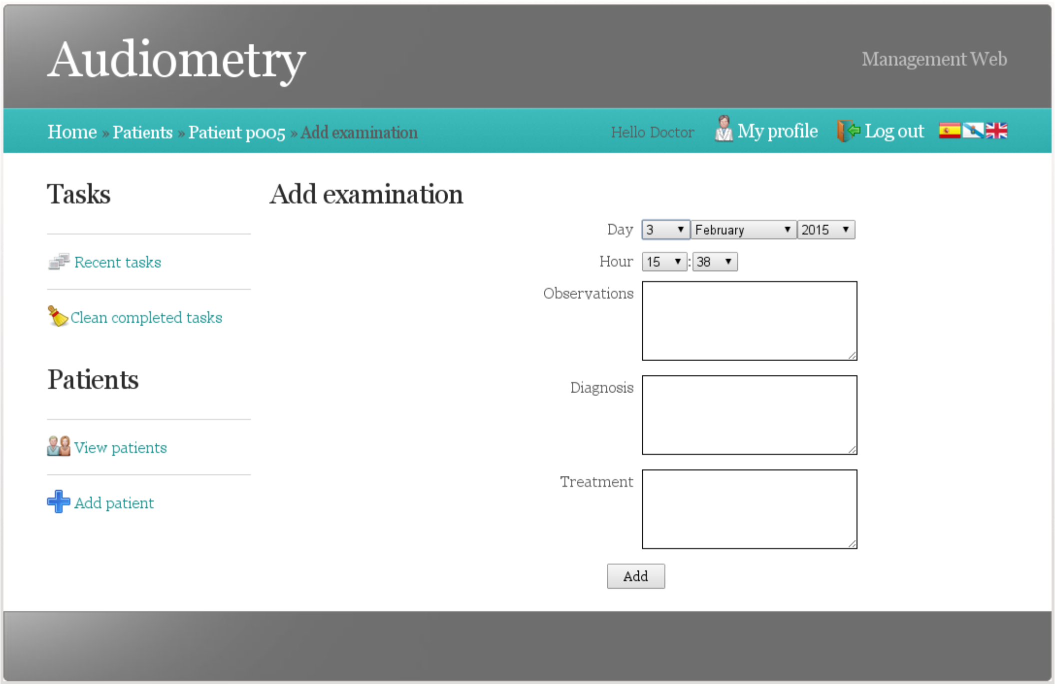
Monday, 21 November 2011 17:01 (2 videos) · Tranfer videos

VIDEOS

Name	Transfer	Process	
20111121170143_a	Play	View analysis	✗
20111121170143_b	Play	View analysis	✗

Figure 3.5: Patient's report screen

When the audiologist wants to upload a new examination he needs to access to the patient to which he wants to associate it. Once he is at the Patient's report screen (see Figure 3.5) by clicking in option Add examination he will be redirected to the Add examination screen (see Figure 3.6). A examination is identified by the date when it was conducted, for this reason, the audiologist must specify day and hour of the examination. The examination can also have associated information that the audiologist would want to register so hi can manage all the relevant information of a patient through this tool. Three text fields are displayed at this screen to that end: observation, diagnosis and treatment.



Day 3 February 2015

Hour 15 : 38

Observations

Diagnosis

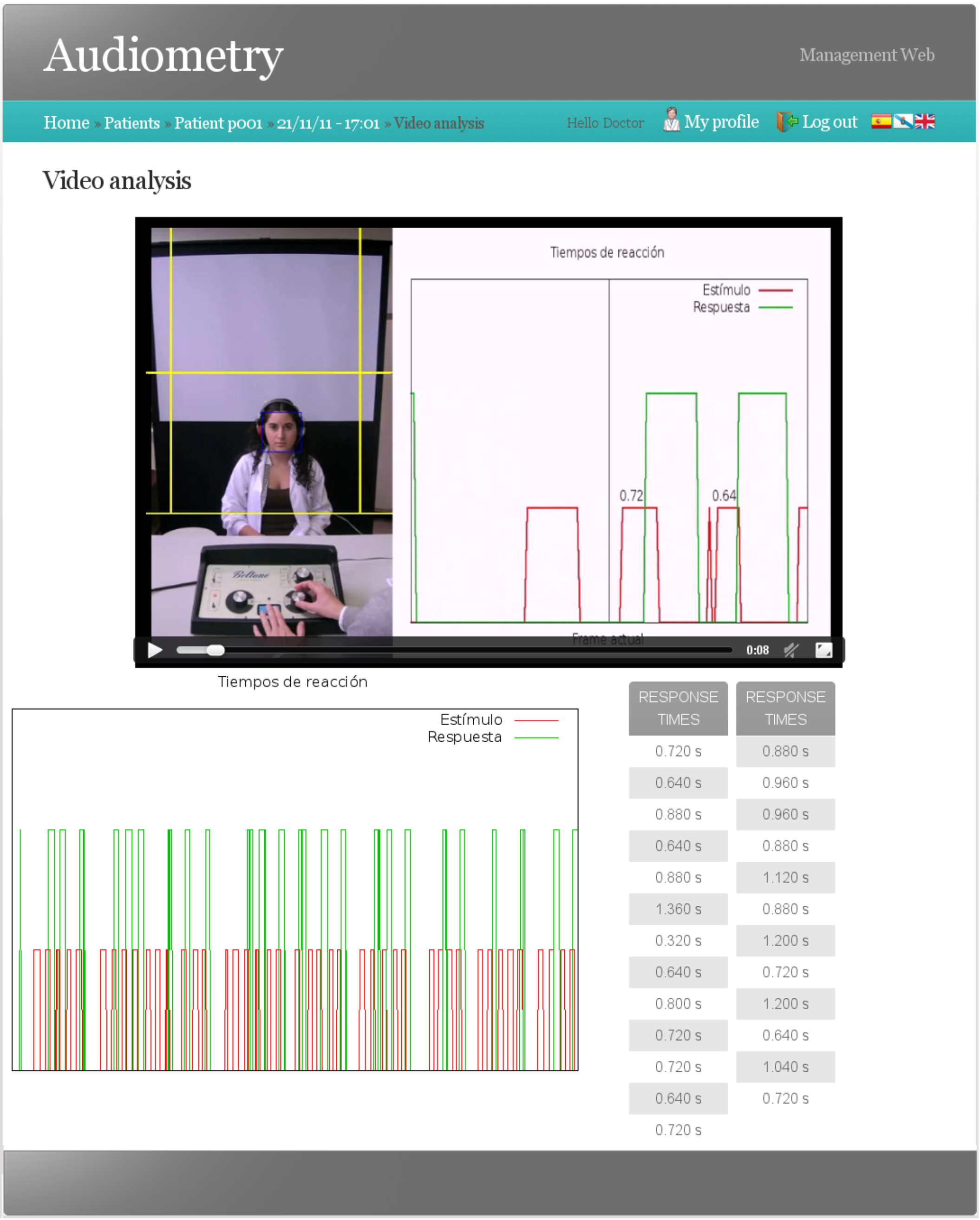
Treatment

Figure 3.6: Add examination screen

Finally, the most relevant screen is the one that shows the processed video sequences. As mentioned before, this option is available for each of the video sequences associated to an examination. By choosing the option View analysis the audiologist will be redirected to the Video analysis screen (see Figure 3.7).

In the first place, the processed video sequences will be displayed and it can be directly reproduced. On the left side of the image the original video sequence is displayed highlighting the positive responses detected by the methodology. On the right side of the image a graph showing the pure tone delivery, the positive responses and the measured response times is represented. This graph moves forward according to the video. Red pulses of the graph represent the pure tone sound delivered by the audiologist, green pulses represent the positive responses provided by the patient, and response times are measured from the moment when the stimulus begins until the moment when the reaction begins (these times are also displayed at the graph).

In second term, a graph representing all the pure tones delivered by the audiologist (red pulses) and all the positive response provided by the patient (green pulses) is displayed (zoom option is available). Each one of the response times measured during the evaluation are represented on a table on the right side.



Audiometry

Management Web

Home » Patients » Patient p001 » 21/11/11 - 17:01 » Video analysis

Hello Doctor My profile Log out

Video analysis

RESPONSE TIMES	RESPONSE TIMES
0.720 s	0.880 s
0.640 s	0.960 s
0.880 s	0.960 s
0.640 s	0.880 s
0.880 s	1.120 s
1.360 s	0.880 s
0.320 s	1.200 s
0.640 s	0.720 s
0.800 s	1.200 s
0.720 s	0.640 s
0.720 s	1.040 s
0.640 s	0.720 s
0.720 s	

Figure 3.7: Video analysis screen

Part II

Analysis of gestural reactions in hearing assessment

Over the course of Part I a methodology for the automatic measurement of response times during typical audiometry was presented. However, as mentioned in the Introduction, not all the patients are able to maintain the level of interaction that the traditional development of this test requires. In the case of patients with cognitive decline or other severe communication problems, the standard protocol of hearing assessment becomes unenforceable since no active interaction audiologist-patient is possible. With this specific group of patients there are some limitations when it comes to maintaining a normal interaction, limitations that are aggravated as the cognitive decline worsens.

Although the evaluation of these patients becomes much more complex, it is still possible if the audiologist is experienced enough. Whereas "normal patients" react by raising a hand (or with voice), patients with cognitive decline typically react unconsciously with subtle facial reactions. These facial reactions occur mainly on the eye region, so the audiologist needs to focus his attention within this region in order to detect changes in the gaze direction, eye opening or closing, or another specific expression change that could indicate some kind of perception to the sound by the patient.

It is important to emphasize that the gestural reactions are particular for each patient, even the same patient may react in different ways during the same session. This variability requires from the audiologist broad experience so he can be able of properly detecting and interpreting the gestural reactions. The subjectivity involved in the gesture interpretation makes this task an imprecise problem, prone to errors, and it greatly limits the reproducibility and robustness of the measurements performed in different sessions or by different experts, leading to inaccuracies in the assessment.

All these considerations make clear the improvements that an automated solution could offer, helping the audiologists in the detection and interpretation of these unconscious gestural reactions. During the next sections, we are going to propose a novel method for the analysis of the eye movements specifically designed for this field. This proposal makes use of computer vision techniques in order to analyze video sequences recorded during the performance of the audiometry. The methodology needs to detect the patient, locate the eye region, and be able to detect movements produces within the eye region.

It is necessary to clarify at this point that other techniques aimed to the interpretation of facial expressions are not directly applicable in this domain. Most of these techniques (such as (Happy, George, & Routray, 2012) or (Chew, Rana, Lucey,

Lucey, & Sridharan, 2012)) are focused on the classification of the facial expressions into one of the typical expressions (anger, surprise, happiness, disgust, sadness, fear, etc.). The facial expressions of this particular group of patients do not directly correspond to any of those categories. They are specific to each patient, without following a fixed pattern, and, as commented before, they can even vary within the same patient.

The main challenge of this methodology is the identification of gestures associated with reactions to the auditory stimuli, which are totally dependent on the patient. In most cases, these reactions are associated with changes on the gaze direction (some samples can be observed in Figure 3.8), namely, when the patient perceives a sound through one of his ears, he unconsciously changes his gaze direction to that specific side.

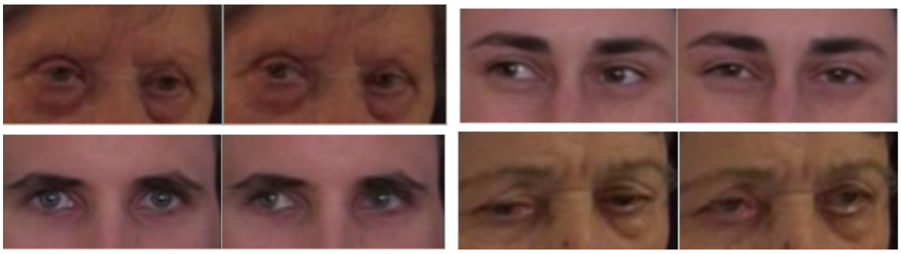


Figure 3.8: Sample of the different eye movements target of detection

From these images, it can be inferred that reactions can be more subtle or marked depending on the patient; in addition, sometimes, the presence of wrinkles or the absence of eyebrow modify the appearance and the features of the area; besides, also changes in the illumination or other lighting conditions may affect the process. To the best of the authors' knowledge, this problem has never been attempted to address through a computational solution, which may be very helpful for the audiologists when evaluating this particular group of patients. In order to provide an automated solution to this specific problem, different approaches are proposed throughout this Part of the thesis.

The development of an automated methodology for the detection and interpretation of the gestural reactions will be of great relevance for improving the objectivity and repeatability in the evaluation of specific group of these patients. By the combination of different computer vision solutions over the video sequences recorded during the audiometric evaluations, we have developed a method aimed to support the audiologists in the detection of eye-based gestural reactions as a response to the auditory stimuli.

A general schema of the global methodology is presented in Figure 3.9. This methodology receives as input a video sequence recorded during the performance of an audiometric evaluation and it gives as a result the detection of the gestural reactions to the sound. This final methodology was obtained through the development of different approaches and improvements which are going to be presented across the next Chapters.

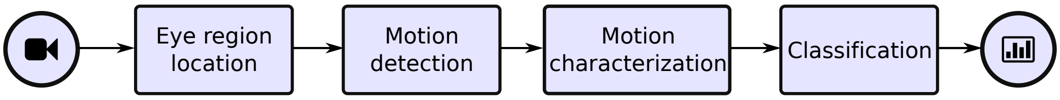


Figure 3.9: Main steps of the gestural reactions analysis.

Chapter 4

1st approach: Optical flow approximation

An initial approach based on the use of optical flow for the detection of movements within the interest region was proposed. The schematic representation of the main steps of this methodology can be seen in Figure 3.9. The method will receive as input a video sequence recorded during the performance of the audiometric evaluations, and this video sequence is going to be processed frame by frame. This proposal has been initially presented in (Fernández, Ortega, Penedo, Vázquez, & Gigirey, 2014), and it is going to be addressed in detail next. The first step of the method is the location of the eye region, which is the region where we wan to detect the movement. After that, the motion is detected using the optical flow, then characterized, and finally classified.

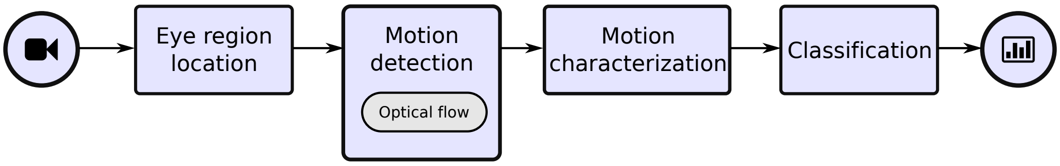


Figure 4.1: Main steps of the optical flow information part.

Since this is the first proposed approach, it will provide the basis for the rest of the process.

4.1 Eye region location

Proper location of the eye region is the first step since the recorded scene is larger and it contains many information that is not relevant at this step. So, for this method we need to establish the eye region as our region of interest (hereinafter ROI). The recorded video sequences are processed frame by frame. In order to detect the ROI, we first locate the face, and after that, within this region we locate the eye region. The initial location of the face allow us to narrow the search area, reduces the computational cost of the next step and makes it less error prone, as already mentioned. Face location is addressed on the Appendix A.

Once the face area has been delimited, the next step is the location of the eye region. We could have considered the detection of gestural reactions all over the face, but the extensive experience of our audiologists with this type of patients allow them to claim that the gestural reactions that really correspond with responses to the auditory stimuli most prominently occur within the eye region. This statement allows us to limit the movement analysis to this particular area and to work without considering other movements that may occur in the rest of the face leading to confusion or inaccuracies.

Eye detection can be broadly divided into three types: template-based, feature-based, and appearance-based methods. For example, in (Jorge, Carvalho, Manuel, & Tavares, 2007) deformable templates are used to extract the eye boundaries. A sample of the second group of methods can be found in (Kawato & Tetsutani, 2004), where blinks are detected based on differences between successive images. The appearance-based ones can be integrated with machine-learning techniques and have been widely developed by the research community during recent years. Some representative algorithms can be found in (Murthy & Natarajan, 2011), where a neural-network-based approach is proposed, and (Zhu & Ji, 2005), where a SVM is applied.

For the location of the eye region the Viola and Jones object detection framework (Viola & Jones, 2001) was considered. A cascade was specifically trained for this study using more than 1000 images of the eye area. Each one of these 1000 images was manually selected in order to delimit our ROI. The training images were cropped from different face images from different face databases. An accuracy of the 98% was obtained during the evaluation of this eye detector. It is capable of reliably detecting the eye region regardless of the expression and even when the eyes are closed, which is a relevant feature given the unconstrained and unpredictable gestures and expressions of target patients. Samples of eye region detections can be observed in Figure 4.2.

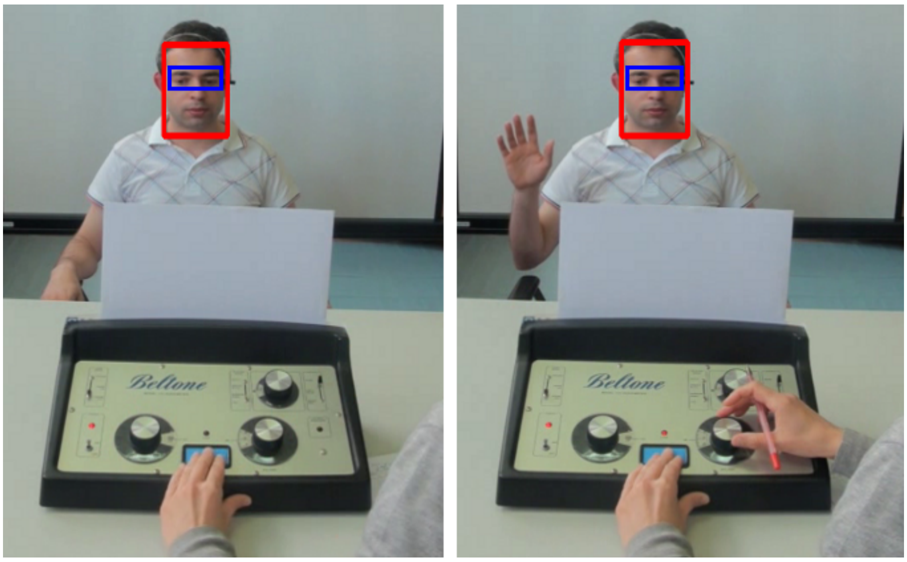


Figure 4.2: Eye detection at different times during the test.

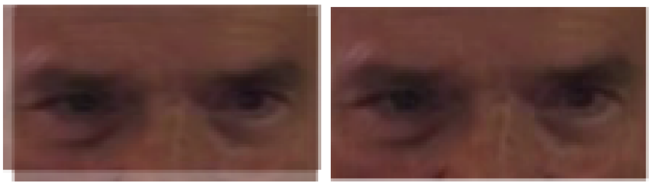
With the aim of facilitating the subsequent steps, it was established that the eye regions captures during an audiometric evaluation must have the same size. Since the Viola and Jones object detector does not fulfill this condition, a later correction is required. To that end, a fixed size is established based on the measurements of the first location. The subsequent eye locations are going to be scaled to this fixed size.

Although the locations provided by the Viola-Jones detector are fairly stable, there might be a small displacement of a few pixels between locations of consecutive frames. Even though this displacement is almost non-significant for the human eye, since the aim of this methodology is the analysis of movements within these regions, it may introduce noise to the results. To solve this, cross correlation between images is calculated.

The cross correlation between images calculates the greater similarity R of a template T inside an image I, according to the classical equation (4.1). In this case, the template corresponds with the eye region located in the previous frame, and the image used for the correlation is based on the current location of the eye region slightly enlarged.

$$R(x,y) = \frac{\sum_{x',y'}(T(x',y')I(x+x',y+y')}{\sqrt{\sum_{x',y'} T(x',y')^2 \sum_{x',y'} I(x+x',y+y')^2}} \tag{4.1}$$

Figure 4.3 illustrates the impact of applying cross correlation normalization. In Figure 4.3(a) the cross correlation was not applied, so, by the overlap of the images it can be observed that there exists a displacement of several pixels. However, when the cross correlation is applied, the displacement is almost non-existent as it can be seen in Figure 4.3(b). Figure 4.4 shows examples of eye region detection for a number of sequences.



(a) (b)

Figure 4.3: Eye region detection: (a) without considering cross correlation, (b) applying cross correlation.

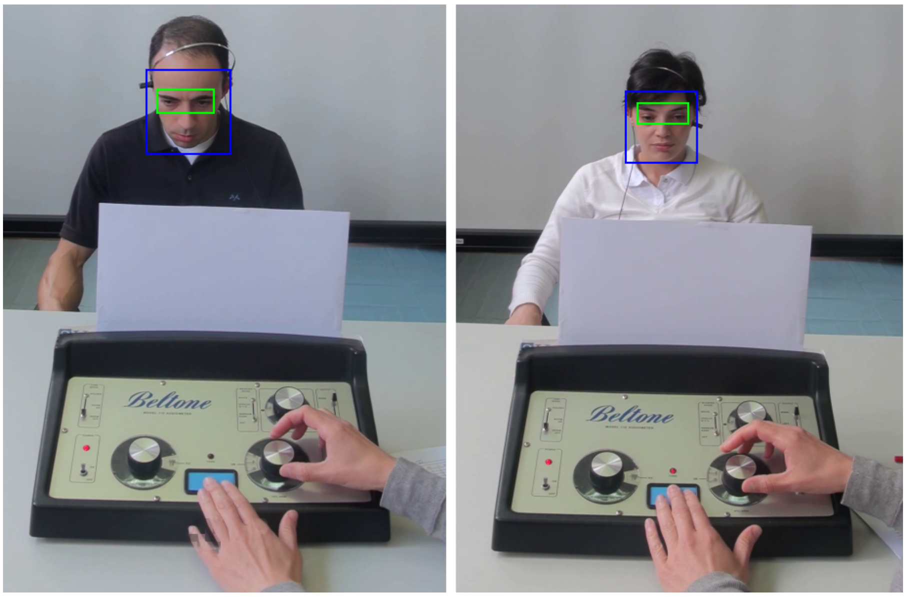


Figure 4.4: Eye region detection samples.

After the eye region location, the proposed methodology makes use of the optical flow information in order to detect the movements occurred within this particular region.

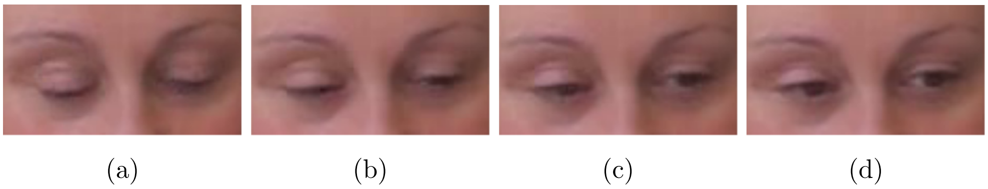
4.2 Motion detection

After the eye region location, this step is aimed to start the detection and analysis of movements or expression changes that occur within this particular area as a reaction to the sound. Due to the nature of the problem, movements are analyzed in a global sense, so a classical point to point feature registration (such as in (Geetha, Ramalingam, Palanivel, & Palaniappan, 2009)) is less effective when the expected set of movements cannot be initially expressed as a function of a particular point in the ROI. Besides, each individual may show different gestures as a reaction and even the same patient may act erratically performing different movements along the audiometric test. Therefore, a template analysis (e.g. (Kumano, Otsuka, Yamato, Maeda, & Sato, 2009; Akakin & Sankur, 2011)) is not possible either, since the reaction gestures of these patients are erratic and they do not correspond to any typical gestural expression.

In order to address this problem, a novel approach specifically aimed to this domain and based on global movement analysis for description was proposed. By the evaluation of the domain and the features of the images to be treated, it was decided to analyze the optical flow between eye region images. The motion is estimated by the use of the iterative Lucas and Kanade (Lucas & Kanade, 1981) optical flow method with pyramids (Bouguet, 2000). Optical flow has shown optimal results in the identification of general and unconstrained movements produced by expression changes.

The recorded video sequences have associated a particular frame rate. If the frame rate is high, comparisons between a frame and the next one may not show changes notable enough, because expression changes cannot occur so quickly. With the purpose of allowing expression changes notable enough, we consider a time window (t) between considered frames, i.e. optical flow is computed between frame i and frame i+t. The t parameter must be chosen as a trade-off between ignoring irrelevant movements and not losing relevant movement. For our particular video sequences with a frame rate of 25 FPS (frames per second), t was empirically established in 3 (see Figure 4.5).

The optical flow operation is based on the detection of interest points. An interest operator (typically, Good Features to Track (Shi & Tomasi, 1994)) is applied over the first reference image and their corresponding points are then localized over the frame i+t. Usually, this interest operator is applied over the first reference image and the obtained points are used for all the images in the video sequence. In this

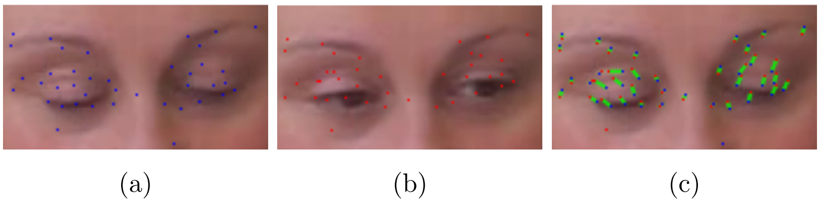


(a) (b) (c) (d)

Figure 4.5: Sequence of consecutive eye region locations. For *t*=3 the optical flow would be performed between (a) and (d).

case, this behavior was modified, so that the interest operator is applied over the first frame of each comparison. This modification is done because changes in the eye expression (e.g. open eyes versus closed eyes) may highly affect the amount of detected interest points and their features. As mentioned, Good Features to Track is the interest operator commonly associated to the optical flow; anyway, a study about the influence on the interest operator for this domain was conducted, this study is presented in Appendix C.

A sample of the application of the interest operator and the optical flow can be observed on Figure 4.6, where Figure 4.6(a) represents the interest points detected over the reference frame *i*, Figure 4.6(b) shows the correspondence of the interest points located over the second frame *i*+*t*, and, Figure 4.6 shows the motion vectors with origin at the interest point in frame i (represented in blue) and end at their corresponding point in frame *i*+*t* (represented in red), for a *t=3* (in 25 FPS video).



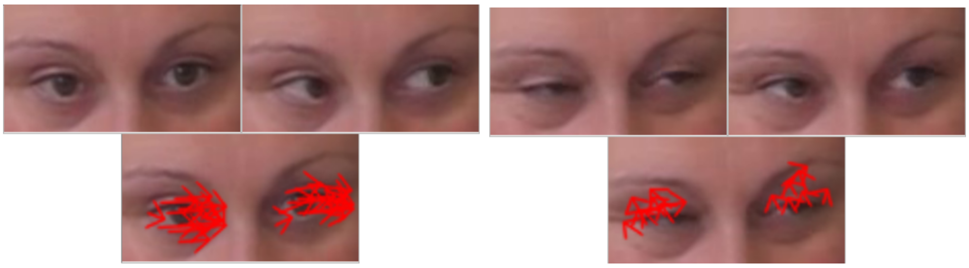
(a) (b) (c)

Figure 4.6: Motion estimation with optical flow. (a) Detected reference points in frame *i*. (b) Optical flow results with the location of the reference points over the frame *i*+*3*. (c) Motion vectors (from blue to red).

Since vectors in Figure 4.6(c) represent direction and amount of movement, this representation can be modified in order to show arrows instead of vectors, where the arrow for a particular point represents its movement from the initial moment to the final one. The length of each arrow represents the magnitude of the movement and arrowhead indicates the direction of the movement.

Figure 4.7 shows a couple of samples with this type of representation. In Figure

4.7(a) the direction of the gaze changes, the optical flow is able to detect this movement and it provides as a results vectors pointing to the side. In the case of Figure 4.7(b) the eye opening increases, so vectors are pointing up, properly representing the movement produced.



(a) (b)

Figure 4.7: Sample of movement vectors represented as arrows for different eye movements: (a) gaze shift and (b) eye opening.

In order to adapt the obtained results to this specific domain and to consider only the significant vectors for the subsequent steps, some considerations are applied over the obtained vectors.

4.2.1 Non-significant vectors removal

Since every movement is detected regardless of its strength, it can be considered that small movements should not be considered in our domain since they do not represent significant movements. This approach removes the non-significant vectors in order to only consider the vectors that really correspond with a significant movement and, thus, facilitate the movement classification.

Movement vectors are ranked according to their magnitude into three different categories. This clustering was established empirically after evaluating the movement vectors of this domain. Since the eye region has been fixed at the beginning of the procedure, the thresholds can be normalized according to these proportions. Equation (4.2) shows the established thresholds for an eye region size of 115 x 62pixel, and Figure 4.8 shows a sample of this classification.

$$\text{vector classification} \begin{cases} 0px \leq v_{short} \leq 1.5px \\ 1.5px \leq v_{interm} \leq 2.5px \\ 2.5px \leq v_{long} \leq 13px \end{cases} \tag{4.2}$$

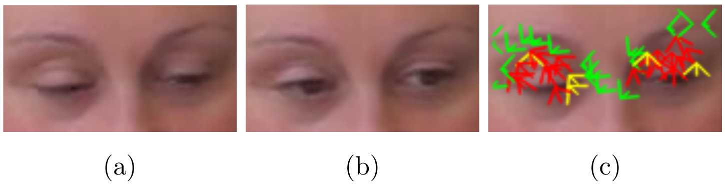


Figure 4.8: Movement vectors ranked by magnitude. (a) reference frame (b) frame to be compared and (c) shows the ranked movement vectors: green for v_{short}, yellow for v_{interm} and red for v_{long}.

Vectors labeled as v_{short} are considered too small to be significant and will be removed, in Figure 4.9 it can be observed this situation. It can be visually concluded that between Figure 4.9(a) and 4.9(b) there are not significant differences. Consistently, the optical flow only detects slight movements corresponding with short vectors (Figure4.9(c))). Since these slight movements are not relevant for our domain, it is correct to discard them.

The second category, v_{interm}, contains those vectors with an intermediate length that does not always correspond with relevant movements, therefore, in principle, they are not considered. Vectors in v_{long} have a length significant enough to always correspond with significant movements, so they are the vectors considered for the next stages of the methodology.

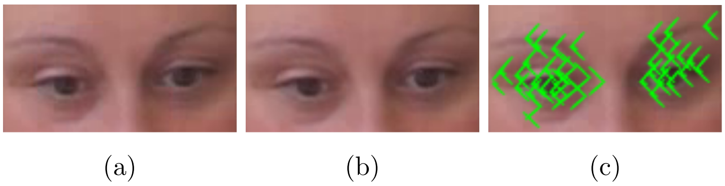


Figure 4.9: Movement vectors ranked by magnitude. (a) reference frame, (b) frame to be compared and (c) shows the ranked movement vectors.

It can be observed that there is also an upper limit for vectors in v_{long}; as occurs with too small vectors, very long vectors must be removed. These vectors are usually related to inaccurate associations or interest points that do not appear in the second image (these behavior can be observer on Figure 4.10).

4.2.2 Discarding the displacement component

It can occur sometimes that the detected motion is due to global movements between the two images instead of movements within the eye region. This global displacement

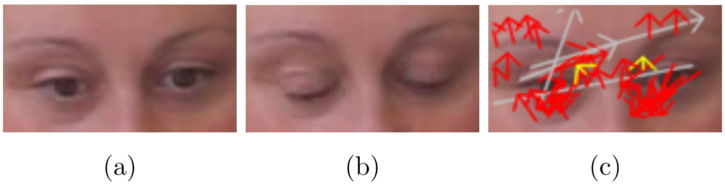


Figure 4.10: Correction for too long movement vectors. (a) and (b) are the frames to be compared, (c) shows the movement vectors. Vectors in grey are removed due to their excessive size.

implies a significant number of vectors with the same strength and direction. In order to correct the displacement component of the image, the number of equal vectors (both in angle and magnitude) is considered. This value is defined by (4.3).

$$C_{\theta,m} = \{v \in C \mid \theta_v \simeq \theta \ \wedge \ |v| \simeq m\} \tag{4.3}$$

where $C_{\theta,m}$ will contain the set of vectors with a similar angle (θ) and magnitude (m). v represents the vector, C the entire set of vectors, θ_v is the angle of the vector and $|v|$ the magnitude of the vector.

Between all the $C_{\theta,m}$ the one with a higher number of elements is chosen (following (4.4)), since, if there is a global displacement this occurs only in one direction.

$$C_{mode} = C_{\theta,m} \mid \forall \theta', m', \ \theta' \neq \theta \vee m' \neq m, \ |C_{\theta,m}| > |C_{\theta',m'}| \tag{4.4}$$

where θ' and θ are the angles and m' and m the magnitudes.

To consider a global displacement, a high number of vectors with the same angle and magnitude is required. This is established with (4.5).

$$C_{mode} \geq |C| \cdot \lambda \tag{4.5}$$

where λ is a parameter that sets the limit for discarding vectors depending on the number of vectors in C_{mode}.

When a global displacement is detected the removal of the displacement vectors is not enough as it is also necessary to correct the remaining vectors. To that end, a subtraction of the vectors is computed, where the displacement component is subtracted to the remaining vectors. The consequences of this optimization can be observed in Figure 4.11, where it can be noted that after the correction of the displacement vectors, no significant movement is detected.

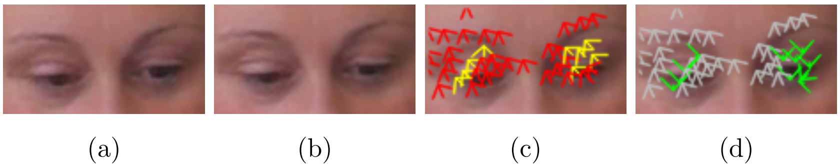


Figure 4.11: Correction for displacement vectors. (a) and (b) are the frames to be compared, (c) shows the movement vectors, which point up when there is no real movement and (d) show the vectors after the correction (vectors in gray are the discarded vectors).

4.3 Motion characterization

The information provided by the optical flow serves as basis to characterize the produced movements. Movement detection allows the identification of those instants during the process where the patient shows a sign of perception to the sound. Since each patient is going to react differently to the auditory stimuli, a classical solution for the global characterization of the facial expression changes is not applicable in this case. Considering this, this proposal relies on the detection of basic gestures within the detected ROI. So that, the aim is to process the complete video sequence by the analysis of the optical flow in order to detect where the significant movements occur.

In order of being able of reliably distinguishing the patient's movements it is needed to characterize the movement when it occurs using as base a group of properties associated to movement. It is necessary to find a set of features that describe the movement that manifests the patient as a sign of perception in such a way that all these movements are equally described. This is an important contribution of this work, since it enables the possibility of modeling any spontaneous movement from the patient in a compact and homogeneous feature space which allows the subsequent analysis of these in a formal and repeatable way. When no significant movement occurs the classification is not applicable, so it is not needed to characterize the optical flow in these time intervals. Instead, when a significant movement occurs, it needs to be characterized. With the aim of capturing all those features that are relevant for the motion characterization we propose a descriptor based on some features that are going to be detailed next.

A set of relevant features is considered in order to try to group all vectors obtained after the previous step. The considered features are: orientation, magnitude and dispersion. A general idea of this feature extraction can be observed from Algorithm 4.1.

Algorithm 4.1: Optical flow feature extraction

```
Data: $iSamples$ which contains the optical flow vectors
Result: $histOrient$, $histMagnit$ and $histDist$
for $i = 0$ to $iSamples.size()$ do
    $p \leftarrow iSamples[i].p$;   ▷ p is the origin $q \leftarrow iSamples[i].q$;   ▷ q is the end
    $angle \leftarrow iSamples[i].angle$;   ▷ angle in degrees
    $histOrient[ceil(angle/45) - 1] + +$;
    $histMagnit[ceil(angle/45) - 1] + = euclideanDistance(p, q)$;
    $histPos[ceil(angle/45) - 1].pushback(q)$;
for $i = 0$ to $histOrient.size()$ do
    if $histOrient[i] \neq 0$ then
        $histMagnit[i] = histMagnit[i]/histOrient[i]$;
for $i = 0$ to $histPost.size()$ do
    $centroides[i] \leftarrow$ compute centroid as the average;
for $i = 0$ to $histPost.size()$ do
    for $j = 0$ to $histPost[i].size()$ do
        $distances[i] \leftarrow$
        $distances[i] + euclideanDistance(centroid[i], histPos[i][j])$;
    $histDist[i] = distances[i]/histPost[i].size()$;
```

First, it must be clarified that each eye is considered separately at this stage, so that each movement generates two movement descriptors, one for the right eye and the other one for the left eye. When generating a movement descriptor, one of the relevant features is the orientation of that movement. The orientation vector provides information about the direction of the movement produced within the eye region. This orientation is different for a change in the gaze direction than for a movement of eye closure or even for a movement of eye opening. For the definition of these descriptors, vectors are divided into eight different equally distributed ranges according to their angle. This classification can be represented mathematically as in (4.6).

$$R_i^* = \left\{ v \in C_f^* \mid \theta_v \in [45 \cdot i, 45 \cdot (i+1)] \right\} \tag{4.6}$$

where $* \in \{L, R\}$, indicating the differentiation between left (L) and right (R) eye, and i takes values from 0 to 7. This way, vectors are grouped according to their angle and the 8 first values of the descriptor correspond to the number of vectors in each range (see (4.7)).

$$n_i^* = |R_i^*| \tag{4.7}$$

It is also important to know the vector's magnitude, because this feature provides information about the intensity of the movement. Considering this, the next eight values of the descriptor are associated with the vector's magnitude. With vectors grouped by ranges (the angle ranges previously defined), the average of the module of the vector is calculated according to (4.8). This feature provides information about the intensity of the movement, allowing to distinguish between strong and soft movements.

$$m_i^* = \frac{1}{n_i^*} \cdot \sum_{v \in R_i^*} |v| \tag{4.8}$$

Finally, the dispersion of the optical flow vector contributes with other eight values to the descriptor. The dispersion of the optical flow allows us to discriminate between localized and global movements. The computation is considered by range, it means, according to the angle of the vectors. From each one of the vectors ($v = \overrightarrow{AB}$) the destination point is taken $B = (B_x, B_y)$ and the center of them is calculated according to (4.9) locating the centroid.

$$c_i^* = \left(\frac{1}{n_i^*} \cdot \sum_{v=\overrightarrow{AB}, v \in R_i^*} B_x, \frac{1}{n_i^*} \cdot \sum_{v=\overrightarrow{AB}, v \in R_i^*} B_y \right) \tag{4.9}$$

Once the centroid is calculated, the dispersion is computed through the calculation of the average distance to that center, according to (4.10).

$$d_i^* = \frac{1}{n_i^*} \cdot \sum_{v=\overrightarrow{AB}, v \in R_i^*} d(B, c_i^*) \tag{4.10}$$

where $d(p, q)$ is the euclidean distance between p and q.

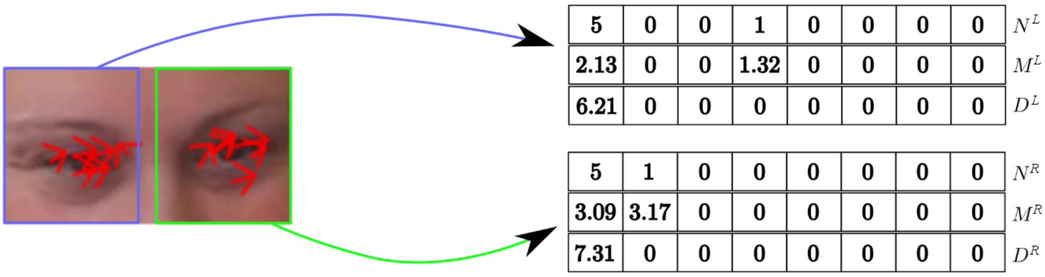
With all this, the descriptor is comprised by a vector of 24 values where the 8 first correspond to orientation, N^*, the 8 next are related to magnitude, M^*, and the last 8 provide information about dispersion, D^*, as in (4.11), (4.12) and (4.13).

$$N^* = \{n_i^* | i \in \{0...7\}\} \tag{4.11}$$

$$M^* = \{m_i^* | i \in \{0...7\}\} \tag{4.12}$$

$$D^* = \{d_i^* | i \in \{0...7\}\} \tag{4.13}$$

A sample of these descriptors can be observed in Figure 4.12.



5	0	0	1	0	0	0	0	N^L
2.13	0	0	1.32	0	0	0	0	M^L
6.21	0	0	0	0	0	0	0	D^L

5	1	0	0	0	0	0	0	N^R
3.09	3.17	0	0	0	0	0	0	M^R
7.31	0	0	0	0	0	0	0	D^R

Figure 4.12: Descriptors sample. Left image shows the movement vectors for each eye and right tables the corresponding descriptors for each eye. First row represents orientation, the second one magnitude and the last one dispersion.

Once the vector descriptors are computed for each movement, the next step is their classification, according to to movement classes determined for this domain.

4.4 Classification

The last step of the methodology is the classification of the descriptors. It must be noted that this classification is not strictly required, different alternatives could be attempted. This methodology allows to cover many more problematic, but considering that the experts tend to classify the type of movements that the patient shows as a reaction, in this initial approach, we are going to handle the problem as similar as possible to the experts' procedure.

At this point, the different movement descriptors are associated with the different movement categories and established by the audiologists for this domain. After reaching an initial consensus with the experts, five typical movement were identified as the most relevant: eye opening (EO), eye closure (EC), gaze shift to the right (GR), gaze shift to the left (GL) and global movement (GM). Also, an extra category was included to categorize those descriptors corresponding to small or insignificant movements (class NM).

At this stage, classification is conducted independently for each eye, meaning that the descriptor for one of the eyes can be classified as eye closure (EC), whereas the descriptor of the other eye can be classified into a different category, i.e. gaze shift to the left (GR). It must be also noted that if the obtained movement vector is totally composed by 0 values, the classification is not conducted since no movement exists.

A previous training of the classifier is needed to accomplish this step. To that end, a supervised training is conducted for several classifiers. The different classifiers considered for this evaluation are: Naive Bayes, Random Forest, Random Committee, Logistic Model Tree (LMT), Random Tree, Logistic, Multilayer Perceptron and Support Vector Machines (SVM). The obtained results are shown in the next Section, obtaining as result the most suited classifier.

Despite of the high resolution of the video sequences, the obtained eye regions do not have the same quality. This is motivated by the need of a general scene including elements other than the patient and also by changes in the lighting conditions during the audiometric evaluation. These considerations increase motion detection difficulties.

4.5 Experimental results

To perform these experiments different video sequences were analyzed and the eye movements produced during them were manually labeled.

The audiologists were equipped with an audiometer Madsen Xeta from Otometrics (represented in Figure 2.8(b)). The auditory stimuli used were pure-tone and the frequency range was between 125-8000 Hz for air conduction, and 250-8000 Hz for bone conduction. The stimulus levels can be set from -10 to 120 dB(HL) with 5 dB(HL) steps for air conduction, and from -10 to 70 dB(HL) also with 5 dB(HL) steps for bone conduction.

The video sequences considered for the following experiments had Full HD resolution (1080 x 1920 pixels) and 25 frames per second (FPS). The device used for recording them is a conventional video camera with full HD resolution and no particular hardware requirements. The only requirement is to try to maintain favorable and constant lighting conditions in order to improve the recorded images and to avoid shadows or occlusions. As mentioned before, the video sequences are focused on the patient who is seated in front of the camera. The image shows the patient waist up and also the surrounding scenario: the audiometer, the hand of the audiologist handling the audiometer, the background, etc, exactly as in the previous part of this work.

Despite of the high resolution of the video sequences, the obtained eye regions do not have the same quality. This is motivated by the need of a general scene including elements other than the patient and also by changes in the lighting conditions during the audiometric evaluation. These considerations increase motion detection difficulties.

Although we have recorded more than 150 audiometric evaluations, only 8 of them are suited for this initial evaluation. As mentioned in the Introduction, these gestural reactions are very specific and they only occur when the patient suffers from cognitive decline or in the case of patients very expressive facially, otherwise, they will raise their hands or respond vocally. This justifies the low number of video sequences available for this evaluation, since it is a ratio that corresponds to the percentage of people with these characteristics within a normal population. It is also important to note the difficulties for recording this specific group of patients; most of people with severe cognitive decline are entered in special centers and special permits and authorizations are needed to record them.

There are 8 considered video sequences, all of them are from adult patients, both male and female and with ages ranging from 45 to 85 years. Each of these video sequences takes between 4 and 8 minutes, so considering a frame rate of 25 FPS, it involves the evaluation of between 6000 and 12000 different frames per video sequence.

The experiments that are going to be addressed next show: a preliminary analysis about the quality of the movement detection; subsequently, over the classification of the detected movement a high number of different classifiers are evaluated in order to select the one that offers better results using the proposed motion descriptor; after that, applying that classifier, a more detailed analysis is conducted over specific video sequences of patients with cognitive decline. And finally, the movements are associated to the auditory stimuli in order to establish validity of characterized reactions respect to the stimuli, setting the suitability of our methodology in the target domain and purpose.

4.5.1 Quality of the movement detection

In order to evaluate the quality of the movement detection, a number of 1500 labeled frames were analyzed. Each one of them were processed to check if any significant movement took place. The results of the classification are detailed in Table 4.1 using a contingency table. It can be noted from Table 4.1 that most frames are classified as non-significant movements; this is an expected consequence of the fact that, generally, the patients are static.

Table 4.1: Contingency table for detection

		Labels		
		No movement	Movement	
Classifications	No movement	1176	12	
	Movement	42	270	
	Total:	1218	282	1500
	Sensitivity:	95.74%		
	Specifity:	96.55%		
	Accuracy:	96.4%		
	F-score:	90.9%		

An important measurement in these cases is the sensitivity, i.e. the ability to detect significant movements when these occur. In our case, the sensitivity rate has a value of 95.74%. Combining sensitivity with specificity (ability to detect non-significant movements) we obtain an accuracy of 96.4%. The F-score of our method is 90.9%.

4.5.2 Classifier assessment in the domain

Several classifiers are trained and evaluated for all the available video sequences in this Section. Only those frames where the optical flow detected a significant movement are considered for the classification. It it important to note that most of the frames do not show a significant movement, since, by default, the patient does not show any reaction. When a significant movement was detected, it was manually classified into one of the possible categories (i.e., eye opening (class EO), eye closure (class EC), gaze shift to the left (class GL), gaze shift to the right (class GR), global movement (class GM) and no movement (class NM)). No additional categories were needed in this dataset.

A total number of 820 descriptors were detected as significant movements and they were classified into one of the possible categories obtaining the distribution showed in Table 4.2.

Table 4.2: Distribution of the significant movements between the considered categories.

	Eye open (EO)	Eye close (EC)	Gaze left (GL)	Gaze right (GR)	Global mov. (GM)	No mov. (NM)
Number of samples	241	339	64	34	108	34

As it can be observed in Table 4.2, the number of samples is not well balanced because changes in the gaze direction are not very common. In order to improve the training dataset, it is necessary to balance the number of samples of the different classes. To that end, for those classes with a high number of samples (class EO, class EC and class GM) a limit number of 75 samples was established. Thus, the final dataset will be composed by 357 frames. Since in order to balance the training datasets, 75 samples were randomly for three of the six classes, it is necessary to conduct several trainings in order to obtain reliable results. For these reasons, 10 experiments were conducted, where each one of them corresponds to a ten-fold cross validation. Furthermore, each one of the 10 training datasets was trained for each one of the eight different classifiers considered for this experiment: Naive Bayes, Random Tree, Logistic, Logistic Model Tree (LMT), Perceptron, Random Forest, Random Committee and Support Vector Machines (SVM). Results of this experiment are summarized in Table 4.3. In this Table, each column corresponds with one of the eight considered classifiers and each row corresponds with one of the ten experiments. Each cell shows the accuracy for the combination of training dataset and classifier. Finally, the last two rows show the average and the variance of the ten experiments for each classifier.

Although for reasons of space and simplicity, only a summary of the results is shown here, all the experiments were studied and discussed in detail. In Table 4.3 only the global accuracy of the experiment is shown, however, the accuracy was also analyzed considering the different classes in order to discuss the behavior of the classifier related to each class. It was observed that, whereas almost all the classifiers offer balanced accuracy for all the classes, the Naive Bayes classifier provides high accuracy in the classification of class GL but very low for classes EO or GM. This behavior can be observed in Table 4.4, where the accuracy by classes is detailed for several experiments. From the global results in Table 4.3 it can be concluded that Naive Bayes is the worst classifier in terms of accuracy, but even if this did not happen, it would not be a valid classifier due to the imbalance of the different

Table 4.3: Accuracy of the classifiers for 10 different trainings. Last two rows show the average and the variance for each classifier.

	Naive Bayes	Random Tree	Logistic	LMT	Perceptron	Random Forest	Random Comm.	SVM
T. 1	55.6%	73.3%	73.3%	73.6%	71.4%	77.5%	76.1%	75.8%
T. 2	55.0%	72.8%	69.7%	72.5%	72.8%	76.7%	77.2%	77.2%
T. 3	51.9%	68.9%	71.7%	73.3%	69.2%	74.2%	73.1%	75.6%
T. 4	58.6%	72.2%	73.6%	74.7%	72.5%	76.4%	76.9%	75.8%
T. 5	53.6%	70.6%	67.8%	70.6%	70.3%	74.2%	75.6%	75.8%
T. 6	55.3%	71.4%	69.7%	73.6%	72.2%	75.8%	78.6%	77.5%
T. 7	56.6%	73.6%	71.4%	72.8%	76.9%	78.1%	79.2%	79.2%
T. 8	55.3%	73.6%	67.5%	71.7%	71.7%	77.2%	76.9%	77.5%
T. 9	55.2%	71.3%	69.1%	69.6%	74.1%	77.2%	75.2%	76.3%
T. 10	59.2%	68.1%	69.2%	70.0%	73.1%	74.2%	75.6%	81.4%
Avg	55.6%	71.6%	70.3%	72.2%	72.4%	76.1%	76.4%	77.2%
Var	4.55	3.78	4.54	2.94	4.52	2.20	3.10	3.40

classes. The problem of imbalance does not happen to the other classifiers, so they can be considered.

Table 4.4: Accuracy by classes of Naive Bayes for different experiments.

	Test 3	Test 5	Test 7	Test 9
Class NM	0.529	0.529	0.529	0.471
Class EO	0.303	0.329	0.408	0.276
Class EC	0.632	0.697	0.776	0.789
Class GL	0.859	0.859	0.859	0.891
Class GR	0.647	0.647	0.618	0.618
Class GM	0.276	0.263	0.263	0.307
Global	51.9444%	53.6111%	56.6667%	55.1532%

Analyzing the global results from Table 4.3 it can be observed that the best results are obtained with the Random Committee and the SVM classifiers. The average accuracy is better for the SVM classifier, whereas the variance is better for Random Committee. Although the variance for SVM is not the minimum, it is one of the lowest values, and thus, it is acceptable. These results only show the global accuracy of the classification, so in order to evaluate the classification by classes and check for the imbalance, in Table 4.5 the accuracy by classes is detailed for the

Table 4.5: Accuracy of the classifiers by classes for the training dataset number 10. Last row shows the average for each classifier.

	Naive Bayes	Random Tree	Logistic	LMT	Perceptron	Random Forest	Random Comm.	SVM
Class NM	47.1%	67.6%	55.9%	55.9%	61.8%	64.7%	58.8%	69.1%
Class EO	35.5%	67.1%	65.8%	67.1%	61.8%	73.7%	73.7%	73.6%
Class EC	73.7%	57.9%	76.3%	76.3%	77.6%	69.7%	75.0%	92.5%
Class GL	90.6%	73.4%	76.6%	76.6%	79.7%	78.1%	82.8%	76.9%
Class GR	64.7%	55.9%	38.2%	38.2%	50.0%	58.8%	61.8%	85.0%
Class GM	44.7%	80.3%	78.9%	81.6%	89.5%	86.8%	85.5%	89.9%
Avg	59.2%	68.1%	69.2%	70.0%	73.1%	74.2%	75.6%	81.4%

experiment number 10.

As it can be observed from Table 4.5, no major imbalances occur neither for Random Committee neither for SVM classifiers. So, going back to the main table (Table 4.3), if we analyze the obtained accuracies, it can be observed that the best accuracy is obtained for the combination of experiment number 10 and the SVM classifier (accuracy of 81.4%). So, this combination (experiment 10 and SVM classifier) is selected as the most suited for the classification task and it will be the one applied for the following experiments.

4.5.3 Classifier evaluation

In order to evaluate the performance of the trained classifier, it was applied to five different sequences from three different patients, who reacted with some kind of eye movement. These three patients were elderly, but they did not have any cognitive impairment, this is why their spontaneous reactions expressed like eye movements were few and far between.

A total number of 1950 frames were analyzed in this experiment, within these 1950 frames, a total number of 545 were classified as significant movements, for the remaining it was considered that no significant movements occur. Result of this classification are detailed in Table 4.6. The columns of this table correspond to the number of frames evaluated, the number of frames where a significant movement was detected, the number of frames correctly classified and, finally, the accuracy of the classification in terms of percentage.

Table 4.6: Initial classification results.

	Number of frames	Significant frames	Correctly classified	% Accuracy
Video 1, seq. 1	350	124	87	70.16129%
Video 1, seq. 2	400	134	80	59.70149%
Video 2, seq. 1	400	90	68	75.55556%
Video 2, seq. 2	400	122	80	65.57377%
Video 3, seq. 1	400	75	59	78.66667%
Global	1950	545	374	68.62385%

By the evaluation of the classification results, it was observed that a couple of optimizations could be applied. First, having into consideration the domain knowledge, it can be established that it must exist continuity along the movement, i.e., if a movement of eye closure (EC) is detected for three consecutive frames in one eye, and in the other eye two frames are classified as eye closure (EC) too, whereas an intermediate frame is classified as global movement (GM), it is very likely that a miss-classification has occurred and that particular frame should be classified as eye closure too. By the application of a voting system, based on the requirement of this continuity, some miss-classifications may be corrected and, this, the accuracy may be improved.

Furthermore, considering the domain and according to the experts opinion, isolated movements of only one frame of length are discarded, because a movement without continuity does not represent a significant movement. Moreover, this system attempts to automate the expert behavior, and the expert does not consider movements of one frame of length since he is not able to detect them in real time, and thus, they are irrelevant for the characterization of patients as they can induce error.

Taking into account these two considerations, the results were optimized and the new obtained results are detailed in Table 4.7. For an easier comparison the accuracies before and after the optimizations are compared in Table 4.8. This last table shows the improvement in accuracy due to the optimizations applied. It can be noted that for Video 3, seq. 1, the accuracy suffers a greater increase. By the analysis of this sequence, it was observed that the lighting conditions were significantly better; thus, under optimal recording conditions, the influence of the proposed optimizations is even greater. Nevertheless, under normal recording conditions, the optimizations also provide an improvement.

Table 4.7: Classification results after the optimizations

	Number of frames	Significant frames	Correctly classified	% Accuracy
Video 1, seq. 1	350	124	93	75.0%
Video 1, seq. 2	400	134	82	61.19403%
Video 2, seq. 1	400	90	73	81.11111%
Video 2, seq. 2	400	122	86	70.49180%
Video 3, seq. 1	400	75	69	92.0%
Global	1950	545	403	73.94495%

Table 4.8: Comparative of the accuracy before and after the optimizations

	% Accuracy before optimization	% Accuracy after optimization
Video 1, seq. 1	70.16129%	75.0%
Video 1, seq. 2	59.70149%	61.19403%
Video 2, seq. 1	75.55556%	81.11111%
Video 2, seq. 2	65.57377%	70.49180%
Video 3, seq. 1	78.66667%	92.0%
Global	68.62385%	73.94495%

4.5.4 Association of movements and stimuli reactions

Finally, the most relevant results are related to the correct detection of eye gestural reactions to the stimuli. Previously, it has been seen that the movements are detected and they are correctly classified. Now, it must be demonstrated that by the correlation of the detected movements and the auditory stimuli, the system is able to detect the reactions to the stimuli.

For this last analysis, the amount of data is not high, but even so, it is interesting to conduct a preliminary analysis to determine if the reactions can be correctly associated with the detected movements. With five different video sequences from three different patients, the aim is to corroborate if the detected reactions correspond to the ones labeled by the experts.

For this experiment, it is considered that a gestural reaction exists when a significant movement occurs during two consecutive frames or more. For this particular group of patients, besides all the rest of eye movements, they expressed their unconscious reactions by gaze movements, so, in this case, only movements of classes gaze shift to the left (GL) and gaze shift to the right (GR) are interpreted as positive reactions to the stimuli.

Besides, it is expected that the eye gestural reactions occur after an auditory stimulus has been delivered. In order to correlate the information from the eye movements and the auditory stimuli it is necessary to know when the auditory stimuli are sent, the methodology that handles the stimuli detection is addressed in Appendix B. A sample of the correlation between stimuli and reactions can be observed in Figure 4.13.

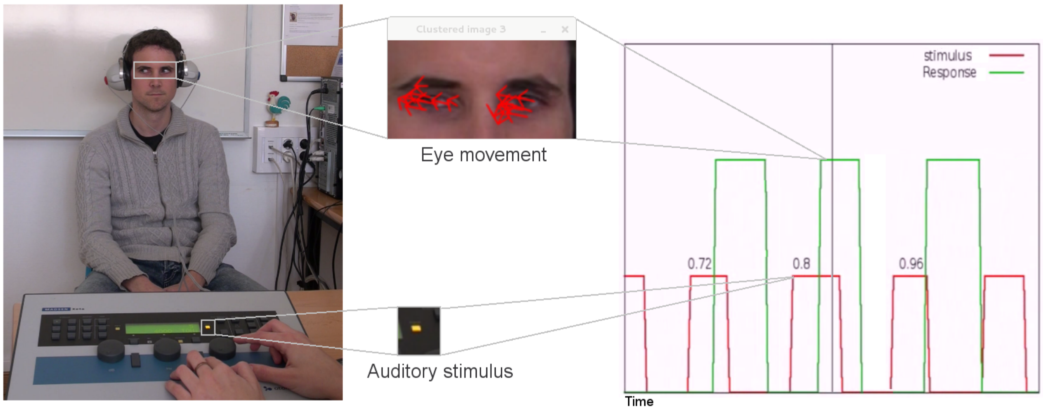


Figure 4.13: Correlation between the stimulus and the reaction. Red signal for the stimuli (signal up when stimulus delivered) and green for reaction (signal up when reaction occurs).

Classification results are processed in this final experiment, and it is established that an eye gestural reaction exists when two or more consecutive frames are classified as gaze sift to the left (GL) or gaze shift to the right (GR). According to this, the number of detected reactions for each video sequence is detailed in Table 4.9. As it can be derived from the results of this Table, all the eye gestural reactions detected by the experts are correctly detected with this methodology too, achieving a 100% of accuracy in the detection of these reactions, which is the main goal of this proposal.

Table 4.9: Number of existing and detected eye gestural reactions for each video sequence. The correlation between natural reactions and our methodology is maximum

	Classification accuracy	Number of reactions	Detected reactions
Video 1, seq. 1	75.0%	2	2
Video 1, seq. 2	61.19403%	1	1
Video 2, seq. 1	81.11111%	3	3
Video 2, seq. 2	70.49180%	1	1
Video 3, seq. 1	92.0%	1	1
Global	73.94495%	8	8

Finally, associated to the auditory stimuli we have shown here the reactions labeled by the experts against those obtained by the system. Although the data are insufficient for a definitive validation, it has been shown that the methodology here proposed offers useful characteristics for this domain.

4.6 Discussion

The methodology presented until this point is capable of characterizing the eye movements of patients with communication difficulties in the audiometric domain, something that had not been addressed so far.

One of the premises of this work is to modify as minimum as possible the traditional protocol of the audiometric assessment. When working with this particular group of patients it is very important to avoid distractions, for this reason, the only requirement of our approach is to place a video camera behind the audiologist performing the assessment.

The auditory stimuli light indicator is properly detected for two different models of audiometers in order to synchronize stimulus and response. Besides this step is dependent on the device, it could be easily adapted to different models.

The detection of the eye gestural reactions is addressed using as base optical flow information. The obtained results showed promising results in the detection and classification of these unconscious eye movements. Besides, the proposed methodology is not only able of classifying the eye movements with reasonable classification rates, but also these rates seem to indicate that the methodology is appropriate for the detection or gestural reactions to the stimuli, paving the way for the development of an automatic tool for this particular domain. The methodology has shown encouraging and positive results, especially considering that it is the first fully automated approximation proposed for this domain.

Furthermore, one of the highlights of this proposal is that its behavior suggests that it may be a useful tool in different domain where the eye gestural reactions could provide relevant information. The proposed steps could be adapted to different domains or to different classification requirements.

Anyway, it can be observed that classification accuracy could still be improved. There are some situations where the a global analysis as the one provided by the optical flow is not enough, i.e., when face movements introduce noise to the movement detection. This situation leads us to consider the proposal of a complementary approach more focused on local features.

Chapter 5

2nd approach: Color information from sclera

As commented before, the proper hearing assessment of patients with cognitive decline or other communication problems becomes a challenge for the audiologists. In Chapter 4 an automated solution based on the use of optical flow was proposed in order to detect the unconscious eye gestural reactions to the sound of this particular group of patients. In this case, we propose a different approach in order to provide an alternative solution based on an alternative feature: the color information from the sclera. The sclera (see Figure 5.1), also known as the white of the eye, is the opaque, fibrous, protective, outer layer of the eye containing collagen and elastic fiber.

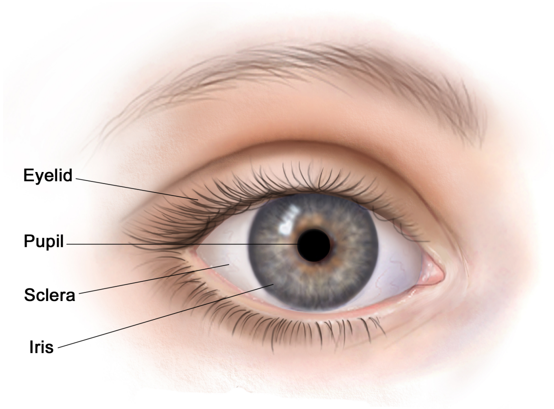


Figure 5.1: Structure of the eye.

In humans the whole sclera is white, contrasting with the colored iris. The human eye is relatively rare for having an iris that is small enough for its position to be plainly visible against the sclera. This makes it easier to infer where an individual is looking at. The relation between the location of the iris and the white distribution of the sclera allows to determine the gaze direction. This relation must be identified and characterized in order to determine if, in our specific domain, a movement has occurred as a reaction to the auditory stimuli. This way, the methodology will enable the proper assessment of patients when no typical interaction is possible, but unconscious gestural reactions to the sound occur.

The development of an automatic solution capable of analyzing the eye movements and detecting gestural reactions to the stimuli would be very helpful for the hearing assessment of patients with severe cognitive decline or other communication difficulties. The proposed solution will receive as input a video sequence recorded during the development of the hearing assessment, and it is going to by analyzed frame by frame.

Using as starting point the general schema presented in the previous Chapter, we propose here an alternative approach for the movement analysis. The proposed methodology is divided into five main stages represented in Figure 5.2. In the first one, we locate the eye region, which is our region of interest. After that, we obtain the location of the pupils' centers. Then we delimit the eyes by the location the eyes' corners, and finally, we characterize and classify the eye position using color information about the sclera. Each one of these steps is going to be discussed next, unless the Eye region location step which is addressed as introduced in Chapter 4 Section 4.1.

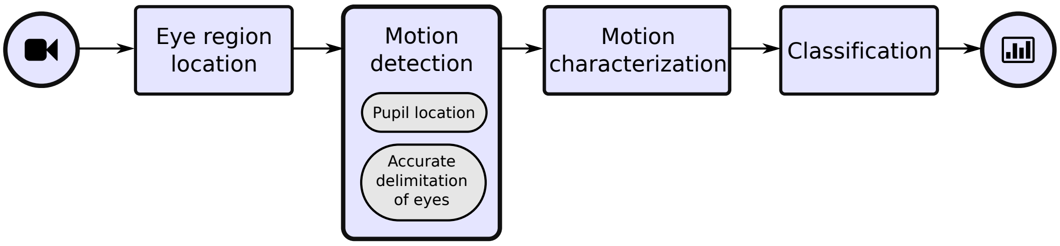


Figure 5.2: Main steps of the color information from the sclera.

5.1 Motion detection

Once the eye region is located, at this step the motion will be detection through the detection of the pupil's center and the accurate delimitation of the eye's boundaries, as depicted in Figure 5.3.

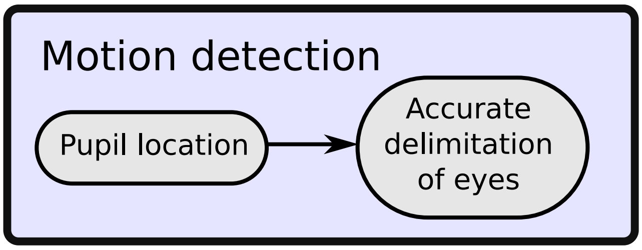


Figure 5.3: Motion detection substeps.

5.1.1 Pupil location

After the location of the eye region, this step is aimed to the location of the center of both pupils, so we can use this information as a reference point in the subsequent steps of the methodology. To that end, a method based in gradients (Timm & Barth, 2011) is applied. The yellow points in Figure 5.4 correspond to the pupil's locations provided as a result by the proposed method for some eye region samples from different patients.

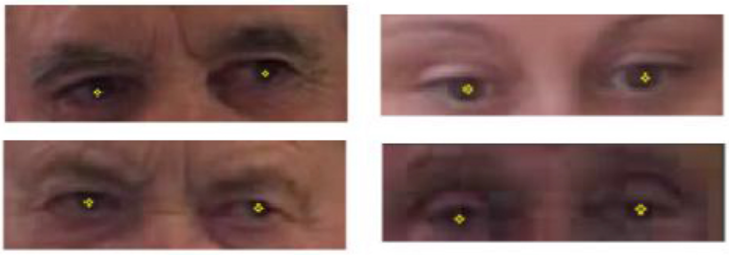


Figure 5.4: Yellow points represent the center of pupil obtained at this step

Two more different methods aimed to the location of the pupil's center were considered too: a method also based in gradients (Kothari & Mitchell, 1996) and the Starburst method (Li, Winfield, & Parkhurst, 2005). However, the Timm & Barth, 2011 approach was the one that show the most accurate results for our domain. The experiment conducted for the choice of the pupil location method is detailed in next. This experiment consisted on measuring the distance between the locations provided by each one of the three methods and the expected pupil's locations manually labeled by the experts. Since the Timm & Barth, 2011 approach was the one the provided the minimum distances, it was the selected solution.

Choice of the pupil location method

In this experiment three different alternatives were compared for the pupil location task. The three methods for the location of the pupil's center analyzed in this study are: method 1 (Starbust (Li et al., 2005)) and methods 2 and 3 (both based on the gradient (Kothari & Mitchell, 1996) and (Timm & Barth, 2011)).

The test set was established used in this experiment was built from 10 different video sequences recorded during hearing assessment. From each one of these 10 video sequences, 20 frames were selected. From is each of these frames, the eye region of each one of the eyes was labeled, obtaining this way a total number of 40 samples for each video sequence. Considering that we have 10 video sequences, a total number of 400 samples will be evaluated, 200 for the right eye and 200 more of the left eye.

In order to conduct this experiment the expected pupils' centers were previously labeled so a comparison with the obtained results from the three different methods could be established. Once we have both the expected center P_e and the center provided as a result by the method P_c, we compute the error in the location as indicated in (5.1). Since both P_e and P_c are pixel locations, the obtained error is also measured in pixels.

$$error = |P_e - P_c| \tag{5.1}$$

Table 5.1 contains the average and the standard deviation (expressed in number of pixels) of the obtained errors for the complete dataset. Each row corresponds with one of the ten video sequences.

For better understanding, based on the data presented in the above table, Figure 5.5 shows a graph representing the average global error (measured in number of pixels) for each one of the ten considered video sequences. It can be observed from this graph that the highest errors are obtained with video sequences 3 and 5, which have an average error considerably higher than the errors obtained for the remaining video sequences.

Table 5.1: Individual results for the pupil location methods.

		Error method 1	Error method 2	Error method 3
Video 1	Average	3.1810	1.0028	2.0688
	Std. dev.	0.8841	0.3156	0.9450
Video 2	Average	2.0700	1.7400	1.6200
	Std. dev	1.0483	1.1902	0.8084
Video 3	Average	2.8263	7.1919	7.0540
	Std. dev	1.2967	7.3473	6.9875
Video 4	Average	4.1156	3.5661	1.1719
	Std. dev	0.9843	2.3030	0.3917
Video 5	Average	6.9201	6.0677	9.8813
	Std. dev	8.5483	7.2433	14.419
Video 6	Average	2.6337	1.6156	1.6697
	Std. dev	0.9093	0.8837	0.6646
Video 7	Average	2.9304	1.2061	1.4733
	Std. dev	1.2719	0.5860	0.9176
Video 8	Average	2.0825	0.9825	0.6255
	Std. dev	0.7741	0.6550	0.5528
Video 9	Average	4.8747	1.6059	1.2394
	Std.dev	5.5548	0.7163	0.5876
Video 10	Average	2.8151	3.0882	0.7506
	Std.dev	0.7367	4.1131	0.7339

After a detailed study of each one of these two videos, it was detected the existence of shades and abrupt illumination changes that noticeably modified the appearance of the eye region. Thus, it can be concluded that the existence of these particular circumstances directly influenced on the results obtained for these specific video sequences. In the remaining video sequences there also exist illumination changes, but they are not so abrupt; so they do not modify as much the appearance of the eye region.

Finally, Table 5.2 presents the global results for this experiment in terms of average and standard deviation. Generally speaking, the results provided by the three different alternatives are quite similar in terms of error. Since the results provided by the Timm and Barth (2011)) approach are slightly better than the others, this is the method chosen in order to be included in our methodology.

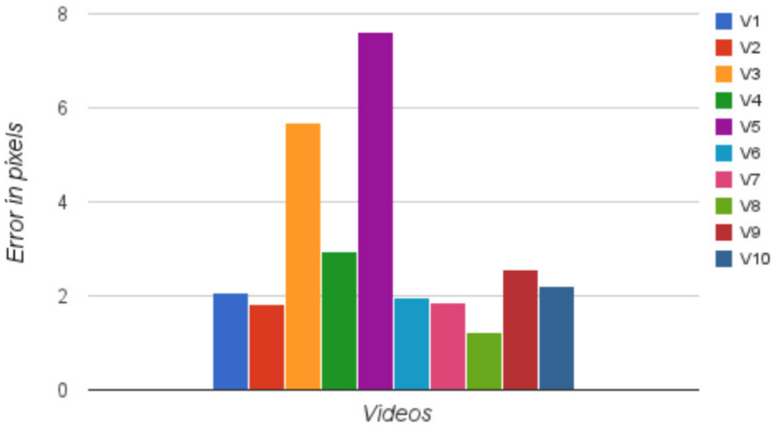


Figure 5.5: Average global error for each video sequence.

Table 5.2: Global error results for pupil location.

	Error method 1	Error method 2	Error method 3
Global average	3.4479	2.7711	2.7380
Global std. dev	2.0645	2.7467	2.7154

5.1.2 Accurate delimitation of eyes

This step aims to locate eye corners. The detection of eye corners is quite more complex than the detection of the iris. This occurs because the eye's corner is a corner located within a skin region that does not have specific features that makes it unique and it is not as well defined as the pupil. Furthermore, the presence of wrinkles or puffiness around the eyes will disturb the appearance of the corner in the image (Lam & Yan, 1996). These circumstances have as consequence that the eye's corner is a area difficult to characterize using edge descriptors, corners, textures or other low level methods (Santos & Proenca, 2011); so, a more complex solution is required here.

Using as information the eye region location and the pupil location provided by the previous steps, we have designed a method that locates the eyes' corners in three steps (see Figure 5.6): selection of the candidate points, selection of the reference points and choosing the best candidates. Each one of these steps is going to be addressed next.

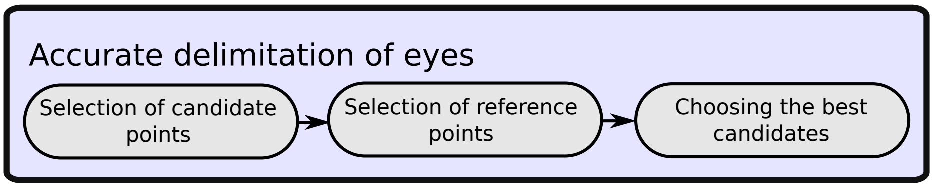

Accurate delimitation of eyes
Selection of candidate points
Selection of reference points
Choosing the best candidates

Figure 5.6: Phases of the delimitation of the eyes stage.

Phase 1: selection of the candidate points

First, we are going to detect points that can be considered as candidates to correspond with the eyes' corners. In order to facilitate this detection, four areas of interest are established using as reference the pupils' centers. For each eye, two ROIs are defined: one on the right side of the eye and the other on the left side (see Figure 5.7). These four areas correspond with those regions where the eyes' corners are expected to be. The main target of these ROIs is to reduce the search area and also to avoid the iris area, which would produce false negatives when applying an interest operator.

So, within these four regions, we are going to apply an interest operator. At this step, three different interest operators were evaluated for this task: Harris (Harris &

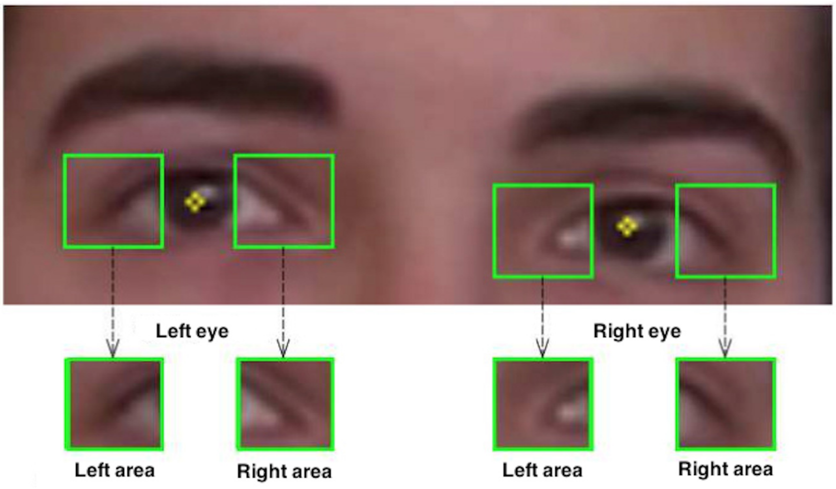


Figure 5.7: ROIs for the detection on the eyes' corners.

Stephens, 1988), Shi-Tomasi (Shi & Tomasi, 1994) and FAST (Rosten & Drummond, 2005) (Rosten & Drummond, 2006). Particularly, we have applied the Shi-Tomasi method. The choice of this method is justified in the study presented in Appendix D, where we compute the distance between the candidate points and the expected eye corner. A sample of the results obtained at this point can be observed in Figure 5.8.

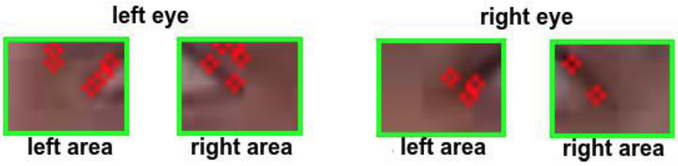


Figure 5.8: Interest operator applied over the four search areas.

As a result of this step, we obtain a set of points that are candidates to be the eyes' corners. Between them, we need to choose those that better represent the eyes' corners.

Phase 2: selection of the reference points

The aim of this phase is to find a set of reference points that allow the removal of false positives from the list of candidate points obtained in the phase previously presented.

Edge information is used at this step in order to obtain edges associated with the limits of the eyelids, so they can be used as a reference of the eyes' limits. The main idea is to apply a different technique in order to complement the information obtained in the previous step. The joint use of both methods provides greater robustness.

We use as input two areas of interest (one for each eye) containing the eye. In order to facilitate edge detection, we increase the enhancement of the eyelid by increasing the contrast of the image. First, we convert the images from the RGB color space to HSV color space, in order to use the saturation channel S. It mus be considered that, regardless of the skin color, pixels from the sclera have always low intensity on the saturation channel due to their white color. Next, an erosion filter considering the radius of the iris is applied. The image obtained as a result from the application of the erosion filter $S_f(x, y)$ is subtracted from the saturation images $S(x, y)$, obtaining this way the subtraction images $R(x, y)$ (as indicated in (5.2)). This process is showed in Figure 5.9, where it can be observed how the eyelid has now more contrast.

$$R(x, y) = S(x, y) - S_f(x, y) \tag{5.2}$$

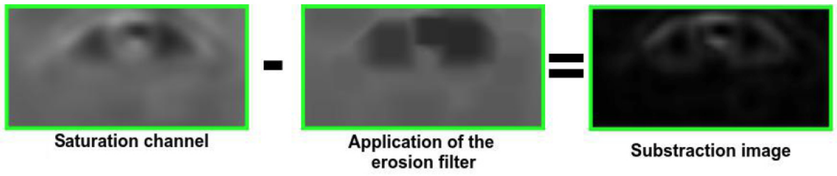


Figure 5.9: Process of the enhancement of the eyelids' contrast.

Next, a threshold for the binarization of the image is computed using as reference some features of the image according to (5.3), where μ is the average value of the pixels from the difference image I_{diff} and σ is the standard deviation.

$$th_s = \mu(I_{diff}) + 0.75 * \sigma(I_{diff}) \tag{5.3}$$

The binarization is computed according to (5.4) where I_{ths} is the thresholded image (see Figure 5.10).

$$I_{ths}(x,y) = \begin{cases} 1, & \text{if } I_{diff}(x,y) > th_s; \\ 0, & \text{otherwise.} \end{cases} \tag{5.4}$$

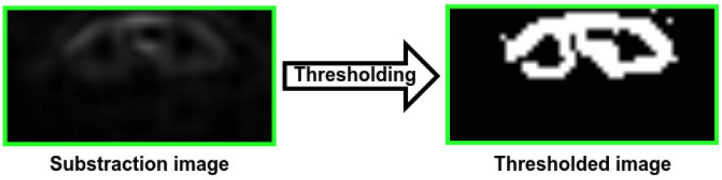


Figure 5.10: Thresholding the subtraction images $R(x,y)$.

Although the eyelids are now easily segmentable, there also exist other tiny elements that need to be removed in order to avoid errors. These elements are small clusters of pixels obtained from the thresholding. In order to remove them, we are going to group the connected pixels as blobs. Once all the pixels are grouped as blobs, we take the biggest one and remove the remaining blob. This step is represented in Figure 5.11, where the bigger blob stays and the three tiny blobs are removed.

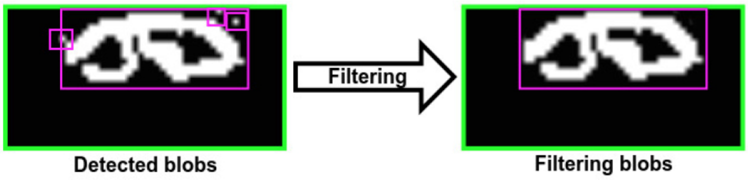


Figure 5.11: Blob filtering for removing noise.

Next, considering the anthropometric constraints that involve the human eye, we can define the eyes' corners as the intersections between the ellipses that represent the eyelid, which correspond with the lower and upper limits over the x coordinate of the blob previously obtained. In the case of obtaining two point with the same value for the x coordinate, we choose the one that has a lower value for the y coordinate. This way, the reference points obtained at the end of this phase can be observed in Figure 5.12.

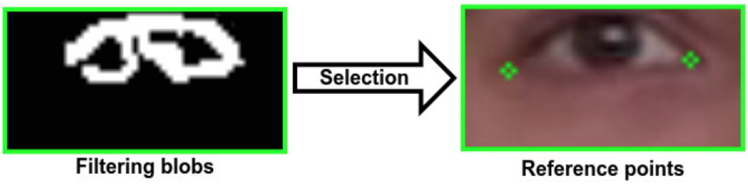


Figure 5.12: Reference points for a sample image.

Phase 3: Choosing the best candidates

At this point, we are going to consider the candidate points and the reference points obtained in the previous phases with the aim of choosing the best candidates to correspond with the eyes' corners.

In order to ensure the robustness of the points representing the eyes' corners, it is necessary to ensure that the reference point associated to those points are good enough, i.e, that they close to the eye corners. To that end, we evaluate the quality of those points based on the anthropometric references of the human eye. Since in our domain the patient remains seated throughout the audiometric evaluation with a stable position, we can ensure that they will not occur significant changes regarding the eye's features during the performance of the assessment.

The reference points obtained from the previous phase are labeled as Pr_1, Pr_2, Pr_3 and Pr_4, where Pr_2 and Pr_3 represent the internal reference points (see Figure 5.13). We also consider the average size of the eye tc_{eye} and the inner distance dc_{in} (where dc_{in} is the distance between Pr_2 and Pr_3) computed for the complete video sequence. According to (5.5) we accept the reference point whenever the distance between the ends of the eye is similar to the average size of the eye tc_{eye} computed for the complete video sequence. d_{right} represents the euclidean distance between Pr_1 and Pr_2, d_{left} is the euclidean distance between Pr_3 and Pr_4, and α is the allowed deviation. Otherwise, we reject those reference points. The same occurs for the inner reference points, where d_{int} is the euclidean distance between Pr_2 and Pr_3, dc_{int} is the inner distance computed for the video sequence and α is the allowed deviation (see (5.6)).

$$f(d_{left}, d_{right}, tc_{eye}, \alpha) = \begin{cases} |d_{left} - tc_{eye}| \leq \alpha \\ |d_{right} - tc_{eye}| \leq \alpha \end{cases} \tag{5.5}$$

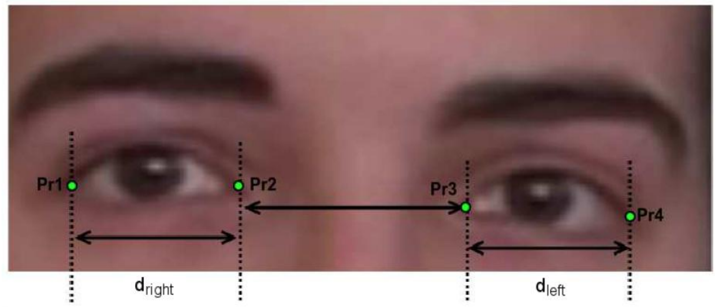


Figure 5.13: Location of the reference points.

$$f(d_{int}, dc_{int}, \alpha) = \begin{cases} |d_{int} - dc_{int}| \leq \alpha, & \text{Accept;} \\ |d_{int} - dc_{int}| > \alpha, & \text{Reject.} \end{cases} \tag{5.6}$$

Once the validity of the reference points has been checked, this information is used to we compute the distances between the candidate points and the associated reference points, finally choosing the candidate point nearest to the reference point P_e. In the case of two or more candidate points with the same euclidean distance to the reference point, we compute the average of those points according to (5.7), where P_c represents each one of the n candidate points with the same euclidean distance to the reference point.

$$P_e(x,y) = P_c\left(\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n}\right) \tag{5.7}$$

Finally, the quality of the selected points Pe_i is analyzed. If Pe_i is far from the nearest reference point Pr_i, considering β as the maximum distance allowed, that Pe_i is going to be discarded and replaced by the reference point Pr_i, as indicated in (5.8).

$$Pe_i = \begin{cases} Pe_i \quad \text{si} \quad |Pe_i - Pr_i| < \beta & \\ & \forall i \in \{1 \ldots 4\} \\ Pr_i \qquad \text{otherwise} & \end{cases} \tag{5.8}$$

The results of this step of the methodology can be observed in Figure 5.14, where the yellow points represent the candidate points, red points correspond with the reference points and green points represent the final points selected as eyes' corners.

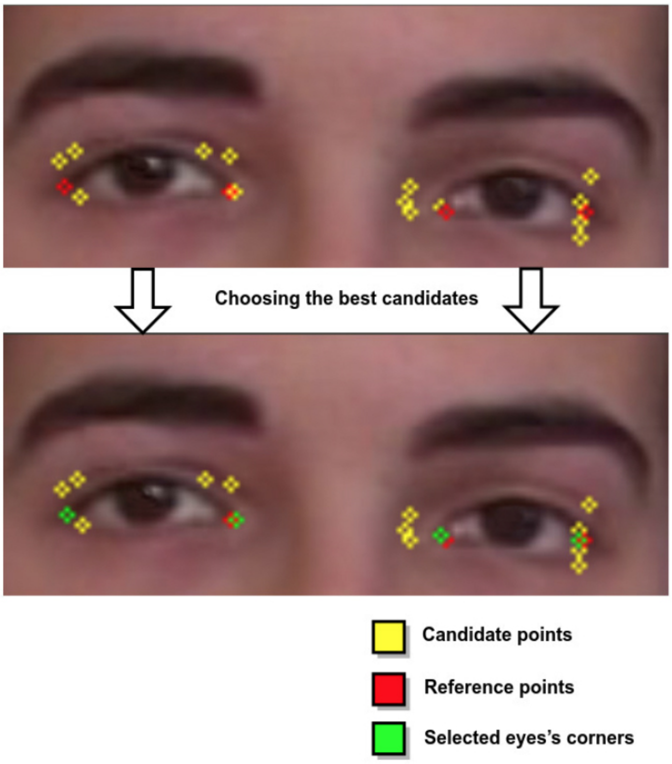


Figure 5.14: Choosing the best candidates: yellow for candidate points, red for reference points and green for the selected eyes' corners.

5.2 Motion characterization and classification

The last step is the characterization of the eye movement. This is going to be accomplished using color information from the sclera, the white area of the eye. To that end, we need to estimate the amount of white in the eye, using as reference the characteristic points previously obtained.

First, the input image is converted to gray-scale color space and a histogram equalization is applied over it. For the characterization of the movement, we are going to compute a gray level distribution representing the gray level for each one of the pixels located in the line connecting both eye corners. A visual representation of this step can be observed in the sample presented in Figure 5.15, where the sclera (since it is the white area in the eye) corresponds to high values in the gray level distribution, while the iris (the colored area of the eye) corresponds with low values in the distribution.

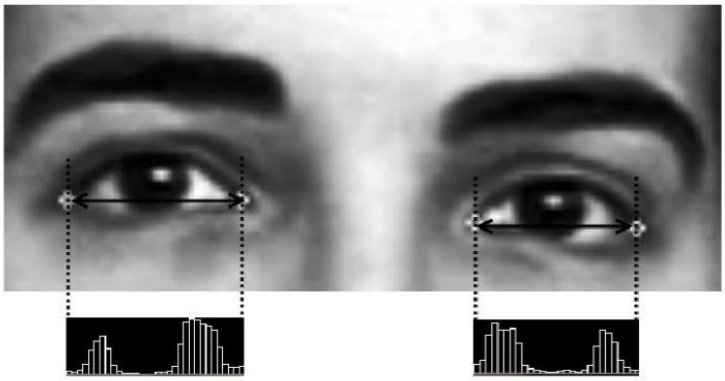


Figure 5.15: Sample of the gray level distribution.

Once the gray level distribution is computed it is possible to divide it into three areas of interest: iris, left side of the sclera and right side of the sclera. To that end, we make use of the information provided by the pupil's center and the estimation of the radius of the iris.

The proposed method is based on the analysis of significant color changes along the gray level distribution. This way, starting from the pupil center we go through the gray level distribution, both to the right and to the left, until we detect the first white pixel that indicates the boundary between the iris and the sclera. This value is accepted as boundary whenever it does not exceed the estimation of the radius of the iris. As a result of this step, we obtain the delimitation of the three areas of interest: iris, left side of the sclera and right side of the sclera. The distribution of the delimitation of these three interest areas can be observed in the sample presented in Figure 5.16.

The distribution of the gray levels in the sclera provides a useful representation of the location of the iris, allowing to deduce the direction of the gaze. According to the audiologists' criteria, four eye movements are considered as relevant in this domain: eye open, eye closed, gaze shift to the left and gaze shift to the right.

As it can be observed from Figure 5.15 where the direction of the gaze is centered, the gray level distribution grows in the sides (where the sclera is located) and it decreases in the center (where the iris is located). In this case, since the direction of the gaze is centered, the gray level distribution grows at both sides of the iris. In the case the direction of the gaze is focused to one side, the gray level distribution will grow on the opposite size.

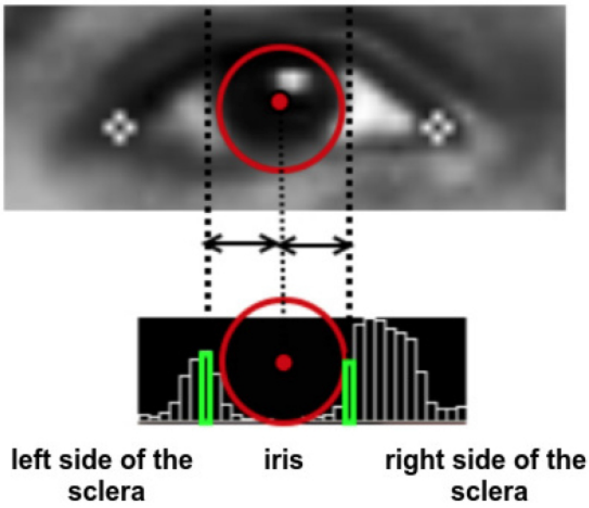


Figure 5.16: Delimitation of the three areas of interest over the gray level distribution.

This behavior is going to be clarified with the following rules:

5.2.1 Eye closed

Due to the absence of the sclera when the eye is closed, it is expected to have low intensity of white values over the gray level distribution. Considering this, we compute the summation of all the gray values G_i for all the points in the distribution. If the summation has a low value we can consider that the eye is closed. Mathematically, this rule can be expressed as (5.9), where θ is a threshold empirically established in order to distinguish between open and closed states as $\theta = 1700/n$, where n is the average size of the eye.

$$\sum_{i=1}^{n} G_i < \theta \tag{5.9}$$

A sample of this category is represented in Figure 5.17 where it can be observed that there is no white information along the distribution, which represents that the eye is closed.

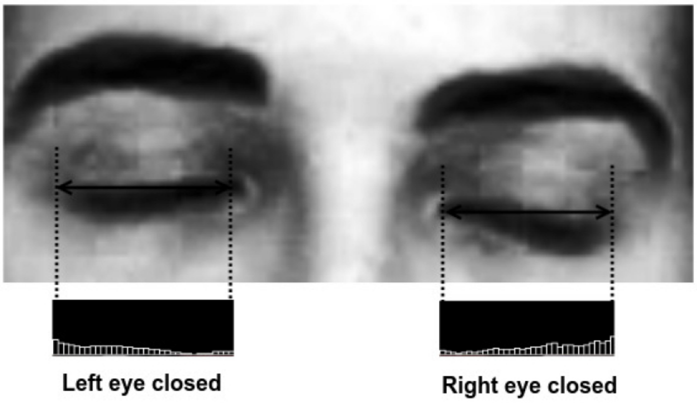


Figure 5.17: Gray level distribution for closed eye.

5.2.2 Eye open

The eye open status is the opposite case of the previous state eye closed, so, in this case, unlike previous case, white information associated to the sclera is present.

The mathematical expression that characterizes this status is just the opposite of the mathematical expression of the previous status (see (5.10)), where G_i represents the gray value of each one of the n points between the boundaries, and θ is the same threshold.

$$\sum_{i=1}^{n} G_i \geq \theta \tag{5.10}$$

Figure 5.18 shows an image sample where it can be observed from the gray level distribution that there is white information from the sclera. In this case, since the gaze is directed to a central fixed point, the white area is distributed at both sides of the iris.

This classification allows a subsequent classification between gaze shift to the left and gaze shift to the right. Only in the case the eye is classified as open the two next rules will be applied in order to determine the direction of the gaze at a specific moment.

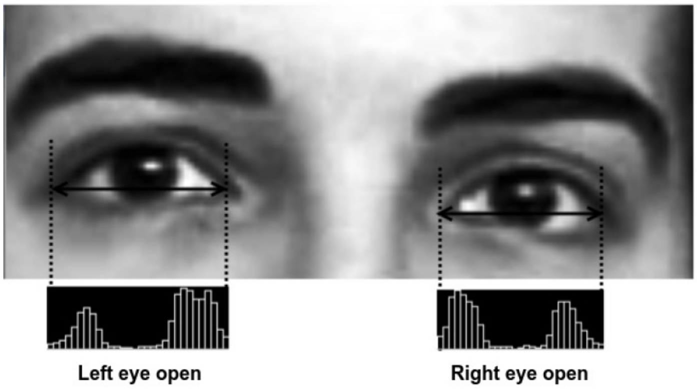


Figure 5.18: Gray level distribution for open eye.

5.2.3 Gaze shift to the right

This status is only possible when the eye has been previously classified open. Otherwise, this rule will not be applied.

In order to distinguish it, we are going to use the information previously obtained about the delimitation of the areas of interest, in this case: left side of the sclera and right side of the sclera. When the eye is classified as open, we compute the summation of the gray level values for each one of the sides of the sclera. Next, it is checked whether the summation of the right side of the sclera represents a small part of the total summation of both sides. This can be expressed mathematically as in (5.11) where Ed represents the n_r points located in the right side of the sclera, T_j represents the n point at both sides of the sclera and β is a threshold empirically established with value 0.20.

$$\sum_{i=1}^{n_r} Ed_i \leq \beta * \sum_{j=1}^{n} T_j \tag{5.11}$$

A sample of this status is presented in Figure 5.19. As it can be observed, since the gaze is oriented to the right, the iris is located near to the right boundary, so the white area is mainly distributed in the left side of the eye, while in the right side of the eye there is almost no white.

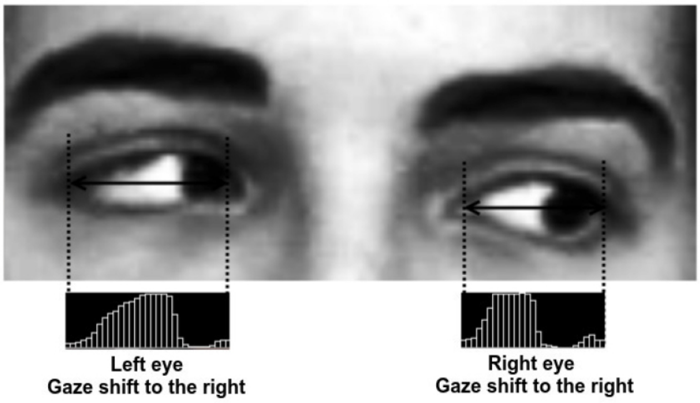


Figure 5.19: Gray level distribution for gaze shift to the right.

5.2.4 Gaze shift to the left

This status is defined analogously to the previous state, and equally, it is only possible when the eye is classified as open. In this case, it is checked whether the summation of the left side of the sclera represents a small part of the total summation of both sides. In (5.12), E_l represents the n_l points located in the left side of the sclera, T_j the n points at both sides of the sclera and β is the threshold. Figure 5.20 shows a sample of this status.

$$\sum_{i=1}^{n_l} El_i \leq \beta * \sum_{j=1}^{n} T_j \tag{5.12}$$

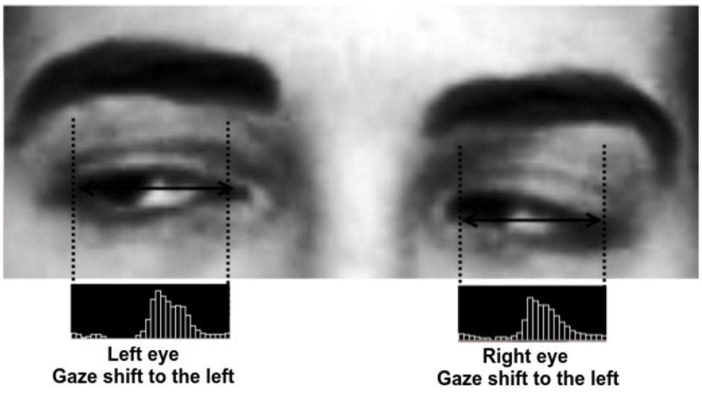


Figure 5.20: Gray level distribution for gaze shift to the leftt.

5.3 Experimental results

Given the preliminary nature of this study, the aim was to test the viability of the methodology over a small dataset.

The proposed methodology is applied frame by frame over three video sequences recorded during the performance of the audiometric evaluations. These video sequences were recorded in the same environment than the images used in the first approach. Video sequences are high resolution images (1080x1920 pixels) with a frame rate of 25 FPS (frames per second). Each video sequence corresponds with a different patient and they have an average duration of 6 minutes. So, with a frame rate of 25 FPS and an average duration of 6 minutes, we analyze an average of 9000 frames for each video sequence.

This experiment is divided into two studies: the analysis of accuracy in the classification of the eye movements and the analysis about the detection of eye gestural reactions to the sound.

5.3.1 Movement classification accuracy

The aim here is to study the suitability of the method in the classification of the eye movements. Three video sequences from three different hearing assessments were analyzed and classified frame by frame. Table 5.3 shows the accuracy for each one of the four eye movement categories considered in this domain: class Eye Closure (EC), class Eye Opening (EO), class Gaze to the Left (GL) and class Gaze to the right (GR). It must be noted that the category Eye open not only corresponds to the situation in which the eyes are open with the gaze fixed to a central point, but it also contains the categories Gaze shift to the left and Gaze shift to the right.

Table 5.3: Accuracy for each one of the eye movement categories

	Class EC	Class EO	Class GL	Class GR
% Accuracy	84.31%	98.2%	85.89%	82.84%

These results are quite acceptable since they are above 82.84%. The high accuracy obtained for the category Eye open is justified because the empirical threshold applied here is optimized for this class, since it contains the gaze movements that are the most relevant categories in this domain. It is important to emphasize that the main goal of this work is not the classification of the eye movements, but the detection of eye gestural reactions to the stimuli, which is the analysis that we are going to conduct next.

For reasons of comparison Table 5.4 present a brief comparative between the accuracy obtained with this approach and the accuracy previously obtained for the optical flow method with the most accurate classifier (SVM). It can be observed how in some cases the accuracies are complementary.

Table 5.4: Accuracy for each one of the eye movement categories: optical flow vs. color from sclera

	Class EC	Class EO	Class GL	Class GR
Optical flow accuracy	92.5%	73.6%	76.9%	85.0%
Color sclera accuracy	84.31%	98.2%	85.89%	82.84%

5.3.2 Detection of reactions to the sound

As commented before, the great contribution that we can provide to the audiologists is the proper detection of the eye movements associated with reactions to the auditory stimuli. Since the video sequences have a frame rate of 25 FPS we know for certain that a reaction will last more than one frame, this is why we are not concerned about obtaining a high success rate in classification, because a typical reaction lasts between 5 and 15 frames, so, the miss-classification of one frame will not affect the proper detection of a reaction.

For this experiment, we consider that a state can be established when three or more consecutive frames receive the same category in classification. Results are detailed in Table 5.5, where we evaluate the agreement between the methodology and the audiologists based on the number of reaction to the stimuli detected by each one of them. The agreement between the methodology and the audiologists is complete (100% of agreement) for the video sequences evaluated in this test. It must be considered here that the detections provided by the experts were obtained by the visualization of the recorded video sequences, in order to maximize their success and also to avoid the inaccuracies that may occur in real time.

Therefore, despite of the not optimum accuracy in the classification of eye movements (around the 87.8%), the detection of the eye gestural reactions is optimal

Table 5.5: Evaluation in the detection of reactions to the sound. Results are expressed in number of reactions.

	Gaze shift to left		Gaze shift to right	
	Expected	Detected	Expected	Detected
Video 1	17	17	15	15
Video 2	17	17	21	21
Video 3	20	20	18	18
Agreement	100%		100%	

(100% of agreement with the experts). This is because of the high frame rate of our video sequences; since a typical gestural reaction lasts more than 5 frames, the misclassification of one frame does not affect to the proper detection of the reaction. It must be also considered here that the detections provided by the experts were obtained by the visualization of the recorded video sequences, in order to maximize their success and also to avoid the inaccuracies that may occur in real time.

5.4 Discussion

This work proposed a novel methodology for the detection and identification of eye gestural reactions as a positive response to auditory stimuli with the aim of supporting the audiologists in the hearing assessment of patients when no cooperation exists. This task is accomplished using information about the color distribution of the sclera.

As in the approach previously presented in Chapter 4, one of the main premises of this work is to modify as minimum as possible the traditional protocol of the audiometric assessment, in order to not influence on the spontaneous behavior of the patient. The conducted experiments have proven the validity of the methodology placing the video camera behind the audiologist (which is seated in front of the patient), an unobtrusive location that minimally perturbs the patient's behavior and provides enough image resolution to apply the proposed methodology obtaining proper results.

As mentioned before, the detection of the eye gestural reactions is accomplished using information about the color distribution of the sclera. The results obtained in this first approach point out the suitability of the method for the detection of these specific kind of reactions. Although in this proposal we only consider four movement categories (the ones that are relevant for the experts in this domain), the methodology could be extended to more or alternative categories. This flexibility would allow the use or the adaptation of this approach to different domains where the eye movements provide relevant information.

The final contribution of this work might be very interesting for the audiologist community since it is a novel method for the detection of eye bases gestural reactions. This methodology will facilitate the hearing assessment of patients with severe cognitive decline or other communication difficulties, patients that can not be evaluated following a standard procedure. A proper hearing assessment of these patients is more difficult to conduct, but it is very important to solve this issue since a proper evaluation may help to treat the hearing loss and improve the quality of life of these patients.

Until this point we have proposed two different alternatives, one comprehensive and generic (the optical flow approach) and the other one more focused on local features (the white color distribution of the sclera). Considering this, we propose now to merge the two proposal in order to evaluate the impact. The combination of these two techniques could provide more robust results by complementing the strengths of each one of the proposals.

Chapter 6

3rd approach: Combining optical flow & color from sclera

Two different approaches were proposed until this point with the aim of providing an automatic support tool for the proper hearing assessment of patients with cognitive decline or other severe communication problems. The optical flow approach was firstly established as the reference method since it was designed in order to represent eye movements, while the color information from the sclera approach is more oriented to the static direction of the gaze, so they can be considered as complementary sources.

Since both approaches are of different natures and provide promising results it was considered that by the combination of both foundations, there obtained results could be supplemented, and therefore they could be improved. From this theory, a methodology combining optical flow information and color information from the sclera was proposed. It is expected that from the combination of these two techniques, the detection accuracy gets improved.

A schematic representation of the main steps of this method can be seen in Figure 6.1. As it can be observed, after the location of the region of interest (i.e. the eye region), the proposed methodology combines the optical flow information about changes in the region of interest and color information from the sclera in order to analyze the eye gestural reactions. Next sections discuss the main one stages presented in the schema. The eye region detection has been presented in Chapter 4 Section 4.1, Optical flow information corresponds with Chapter 4 and Color information from the sclera is addressed throughout Chapter 5.

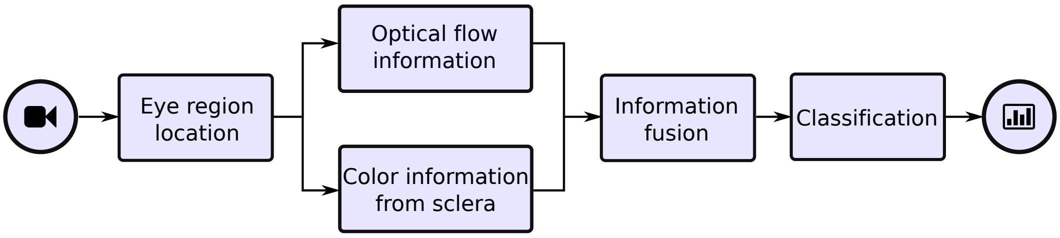


Figure 6.1: Main steps of the proposed methodology.

6.1 Information fusion

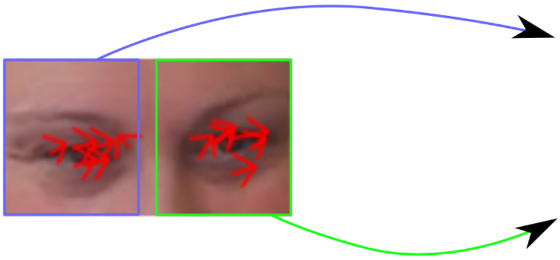
The optical flow provides global information about the movements occurred within the eye region, while the information provided by the color information from the sclera is more focused on the local level. In order to provide a more robust and precise characterization, both approaches are combined. To that end, it is necessary to join the features obtained with both approaches into a single descriptor. Next subsections remind us how the movement descriptors should be obtained for both proposals.

6.1.1 Optical flow information

The optical flow branch was presented throughout Chapter 4. With the aim of reliably distinguishing the patient's movements, it is necessary to characterize the movements detected by the optical flow.

The descriptor provided by this branch is going to be combined with a descriptor representing the color information from the sclera. The characterization of the optical flow vectors was addressed in Chapter 4 Section 4.3, where all the procedure and basis were explained in detail. Vectors were described in terms of: orientation, magnitude and dispersion; composing a descriptor of 24 values for each one of the eyes.

A sample of these descriptors can be observed in Figure 6.2. The formulas that define the each one of the rows in the descriptor can be consulted in Chapter 4 Section 4.3.



5	**0**	**0**	**1**	**0**	**0**	**0**	**0**	N^L
2.13	**0**	**0**	**1.32**	**0**	**0**	**0**	**0**	M^L
6.21	**0**	**0**	**0**	**0**	**0**	**0**	**0**	D^L

5	**1**	**0**	**0**	**0**	**0**	**0**	**0**	N^R
3.09	**3.17**	**0**	**0**	**0**	**0**	**0**	**0**	M^R
7.31	**0**	**0**	**0**	**0**	**0**	**0**	**0**	D^R

Figure 6.2: Optical flow descriptors sample. Left image shows the movement vectors for each eye and right tables the corresponding descriptors for each eye. First row represents orientation, the second one magnitude and the last one dispersion.

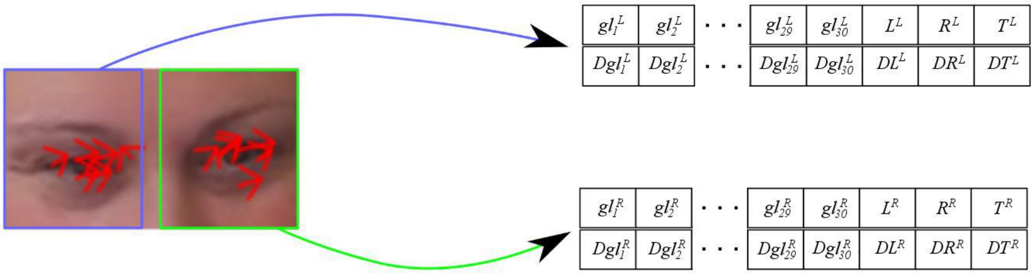
6.1.2 Color information from the sclera

This second branch of the methodology is based on the proposal presented throughout Chapter 5. In the original proposal, the eye movements were classified into the considered categories using a set of rules based on the distribution of the gray values. For this new approach, we want to characterize the color distribution as a movement descriptor equivalent to the one defined for the optical flow branch. It would be interesting to define these descriptors in a way they also represent changes in the gaze direction, equivalently as how the optical flow branch does it. The set of rules established in the original proposal classified the direction of the gaze throughout the actual position, but they did not consider previous information. It is important to consider the previous information since we are interested on changes within the eye region, not only on the actual state.

For this proposal, a movement descriptor composed by 66 values is defined, 33 values for the actual state and 33 more values considering previous information. First, the gray level distribution is normalized to 30 values, so each one of this new 30 values in the gray level distribution is a descriptor in the final movement descriptor. Moreover, considering the delimited sides of the sclera, the summation of the gray level for each one these sides is computed. These two values, along with their sum, are added as 3 additional values of to the movement descriptor.

Besides, since we are interested not only in the actual current moment but also on the movement with respect to previous frames, we are going to compare the actual descriptor with a previous descriptor. As in the optical flow case (Chapter 4), with the purpose of allowing expression changes notable enough, we consider a time window (t) between considered frames, i.e. we are going to compare the descriptors between frame i and frame i+t. The t parameter has been established in

3 for our video sequences (which have a 25 FPS frame rate). Considering this, the subtraction of the descriptors of frames i and i+t is computed adding this way 33 new values to the final descriptor. This way, the movement descriptor is composed by 66 values according to Figure 6.3.



gl_1^L	gl_2^L	$\cdots$	gl_{29}^L	gl_{30}^L	L^L	R^L	T^L
Dgl_1^L	Dgl_2^L	$\cdots$	Dgl_{29}^L	Dgl_{30}^L	DL^L	DR^L	DT^L

gl_1^R	gl_2^R	$\cdots$	gl_{29}^R	gl_{30}^R	L^R	R^R	T^R
Dgl_1^R	Dgl_2^R	$\cdots$	Dgl_{29}^R	Dgl_{30}^R	DL^R	DR^R	DT^R

Figure 6.3: Color descriptors sample. In the first row, the first 30 values correspond with the normalized gray level distribution, L is the summation of the gray levels in the left side of the sclera, R is with the summation of the gray levels in the right side of the sclera and T is the summation of both sides. In the second row, the first 30 values correspond with the subtraction of the normalized gray level distribution of frame i and frame i+t, DL is the subtraction of the summation of gray levels in the left side of the sclera, DR is with the subtraction of the summation of gray levels in the right side of the sclera, and DR is the subtraction of the summations at both sides.

In the first row of Figure 6.3, the first 30 values correspond with the normalized gray level distribution, L represents the summation of the gray levels in the left side of the sclera, R corresponds with the summation of the gray levels in the right side of the sclera and T is the summation of both sides. In the second row, the first 30 values correspond with the subtraction of the normalized gray level distribution of frame i and frame i+t (6.1), DL represents the subtraction of the summation of gray levels in the left side of the sclera (6.2), DR corresponds with the subtraction of the summation of gray levels in the right side of the sclera (6.3), and DR is the subtraction of the summations at both sides (6.4).

$$Dgl_n = gl_{i+t,n} - gl_{i,n}, \mathit{where}\ n \in [1, 30] \tag{6.1}$$

$$DL = L_{i+t} - Li \tag{6.2}$$

$$DR = R_{i+t} - Ri \tag{6.3}$$

$$DT = T_{i+t} - Ti \tag{6.4}$$

Once the vector descriptors are computed for each movement, the next step is their combination with the optical flow information and their classification according to the movement categories determined for this domain.

6.1.3 Descriptor fusion

The final descriptors are composed by 90 elements: 24 descriptors from the optical flow and 66 descriptors from the color information from the sclera. This situation is represented on Figure 6.4, where the first three rows of each eye correspond with the optical flow descriptors, while the fourth and fifth rows correspond with the color information from the sclera descriptors.

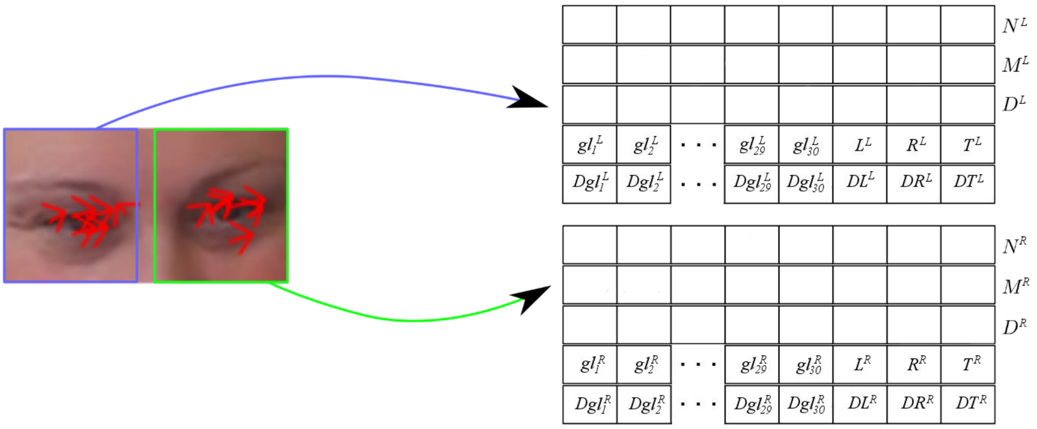

N^L
M^L
D^L
gl_1^L gl_2^L . . . gl_{29}^L gl_{30}^L L^L R^L T^L
Dgl_1^L Dgl_2^L . . . Dgl_{29}^L Dgl_{30}^L DL^L DR^L DT^L
N^R
M^R
D^R
gl_1^R gl_2^R . . . gl_{29}^R gl_{30}^R L^R R^R T^R
Dgl_1^R Dgl_2^R . . . Dgl_{29}^R Dgl_{30}^R DL^R DR^R DT^R

Figure 6.4: Descriptor sample. The first three rows of each eye correspond with the optical flow descriptors, while the fourth and fifth row correspond with the color information from the sclera descriptors.

6.2 Classification

Using the combined descriptors obtained after the previous step, the last step of the methodology is their classification into the considered movement categories. A final experiment on classification was conducted in order to find the most suited classifier. A supervised training is conducted with the following classifiers: Naive Bayes, C4.5, Random Forest, Random Committee, Logistic Model Tree (LMT), Random Tree, Logistic, Multilayer Perceptron and Support Vector Machines (SVM). The complete experiment and the final results are detailed in the next Sections.

6.3 Experimental results

For the development of this evaluation different video sequences recorded during the performance of audiometric evaluation were analyzed and the eye movements occurred during them were manually labeled.

The considered video sequences had Full HD resolution (1080 x 1920 pixels) and 25 frames per second (FPS). The device used for recording the experiments is a conventional video camera with full HD resolution and no particular hardware requirements. The only requirement is to try to maintain favorable and constant lighting conditions in order to improve quality of the recorded images and to avoid shadows or occlusions. Another important consideration is to place the video camera behind the audiologist (which is seated in front of the patient), an unobtrusive location that minimally perturbs the patient's behavior and provides enough image resolution to apply the proposed methodology obtaining proper results.

As mentioned before, the video sequences are focused on the patient who is seated in front of the video camera. The image shows the patient waist up and also the surrounding scenario: the audiometer, the hand of the audiologist handling the audiometer, the background, etc. The reason for such a general scene is the need of recording the audiometer in order to detect the moments when the auditory stimuli are being sent, which allows to correlate this information with the detected eye movements.

As it occurred in the previous approaches, despite of the high resolution of the video sequences, the obtained eye regions do not have the same quality. This is motivated by the need of a general scene including elements other than the patient and also by changes in the lighting conditions during the audiometric evaluation. This limitation of resolution increases motion detection difficulties.

Another major barrier is the difficulty on obtaining video sequences of this specific group of patients. Unconscious gestural reactions tend to occur as a reaction when the patient suffers from cognitive decline or other severe communication disorders. This situation implies two problematics: first, not all the population will exhibit these gestural reactions to the sound, which implies the necessity of specifically searching for patients in this situation. Second, most of people with severe cognitive decline are entered in specific institutions and special permits and authorizations are needed to record them.

To the eight video sequences considered for the experiments in Chapter **??**, three new videos with several eye gestural reactions to the sound were achieved for this final experiment. This 11 video sequences correspond to male and female patients with different ages. Each one of the video sequences takes between 4 and 8 minutes, and each one of them is analyzed frame by frame (remember that the frame rate is 25 FPS).

The considered movement categories for this experiment are: eye opening (EO), eye closure (EC), gaze shift to the right (GR) and gaze shift to the left (GL); since they are the four movement categories indicated as relevant for this domain by the audiologists. The test video sequences were analyzed frame by frame, and each one of the relevant movements was labeled into the corresponding category. A total number of 1180 descriptors were classified as significant movements and they were classified into the corresponding category, obtaining the descriptors distribution showed in Table 6.1.

Table 6.1: Distribution of the significant movements between the considered categories.

	Eye open (EO)	Eye close (EC)	Gaze left (GL)	Gaze right (GR)
Number of samples	408	310	230	232

As it can be observed in Table 6.1 we still have a small imbalance for the categories gaze shift to the right (GR) and gaze shift to the left (GL). However, this imbalance is less pronounced than the one we had in the experiments of Chapter **??** (see Table 4.2), since the new video sequences obtained for this experiment outweigh the situation.

With all the 1180 labeled descriptors, a supervised training is conducted for several classifiers. The considered classifiers are: Naive Bayes, C4.5, Random Forest, Random Committee, Logistic Model Tree (LMT), Random Tree, Logistic, Multilayer Perceptron and Support Vector Machines (SVM). A 10-fold cross validation was applied for the experiment. Results of this experiment are detailed in Table 6.2, where a comparison between applying only the optical flow descriptors vs. the optical flow and color from the sclera descriptors is presented.

As it can be observed from these results, the classification accuracy is improved in all cases with the introduction of the complementary information from the color distribution of the sclera. Worst results are obtained for the C4.5 and Random Tree classifiers, both cases with an accuracy under the 90%. The rest of the classifiers achieve classification accuracies over the 90%. Best results are obtained for the Support Vector Machines, with an accuracy of the 97.46%, which is highly accurate result for this domain.

Table 6.2: Classification accuracy comparative: left column for optical flow descriptors and right column for optical flow & color from the sclera descriptors.

	% Accuracy	
	Optical flow	Optical flow &
Method		color sclera
Naive Bayes	84.5059%	93.2203%
C4.5	87.2697%	89.8305%
Random Tree	86.8509%	88.3051%
Logistic	88.7772%	90.3390%
LMT	89.7822%	97.0339%
Perceptron	88.9447%	97.2034%
Random Forest	90.1173%	94.5763%
Random Committee	90.3685%	95.6780%
SVM	**91.4573%**	**97.4576%**

For a deeper study, Table 6.3 presents the true positive rate by classes for the combined approach.

Table 6.3: Classification results by classes for the combined approach using a 10-fold cross validation for different algorithms

		True positive rate			
Method	% Accuracy	Class EC	Class EO	Class GL	Class GR
Naive Bayes	93.2203%	0.944	0.955	0.939	0.875
C4.5	89.8305%	0.924	0.903	0.896	0.849
Random Tree	88.3051%	0.895	0.887	0.896	0.862
Logistic	90.3390%	0.951	0.877	0.896	0.862
LMT	97.0339%	0.983	0.965	0.978	0.961
Perceptron	97.2034%	0.983	0.961	0.978	0.961
Random Forest	94.5763%	0.973	0.945	0.952	0.892
Random Committee	95.6780%	0.973	0.961	0.943	0.935
SVM	97.4576%	0.980	0.977	0.978	0.957

6.4 Discussion

The previous approach presents a methodology for the detection and classification of eye gestural movements as a response to auditory stimuli during the performance of audiometric evaluations. This methodology will allow the automatic detection of positive unconscious gestural reactions which can be interpreted as positive reactions in the hearing assessments of patients with cognitive decline or other severe communication difficulties. When no cooperation exists, the audiologist focus his attention of the detection of these particular reactions, but the development of an automated tool will facilitate his work, avoid subjectivities and make it less error prone.

In this novel proposal, the detection of eye-based gestural movements is accomplished by the combination of the optical flow information and the analysis of the color distribution of the sclera. These two approaches were presented and justified along the previous chapters (Chapter 4 and Chapter 5). The results obtained with this combined proposal allow us to conclude that the inclusion of the information provided by the color distribution of the sclera improves the obtained results. In the presented results it can be observed how the combined descriptors obtain in all cases an accuracy higher that the one obtained when considering only the optical flow descriptors. The most accurate results were obtained for the SVM classifier, with a classification accuracy of the 97.46%, which is highly accurate result for this domain. However, if the true positive rate by classess is considered, it can be noted that class GR (gaze shift to the right) is the one with the lower value. In most cases, classes GL and GR are going to be considered as the relevant ones, so it would be interesting to improve somehow their classification accuracy.

As in the individual approaches presented before, one of the main premises is to not alter the traditional protocol of the audiometric assessment. As in the previous proposals, the only requirement of the methodology is to place the video camera behind the audiologist at an unobtrusive location that does not pertub the patient's behavior.

In clinical terms, the manual analysis conducted by the audiologists can be automated with the main benefit of being unaffected by subjective factors when evaluating patients with cognitive decline or severe communication difficulties. Besides the fact that the automatic proposal produces unbiased results, it also saves times for the experts and provides a detailed identification of the eye-based gestural reactions. In this sense, the audiologist can have a more objective detection of the unconscious positive reactions to the sound, which means a great help in the proper evaluation of the hearing hearing assessment. The proper diagnosis of hearing loss will allow the prescription of appropriate hearing aids, thus improving the quality of life of these particular group of patients.

As it was also commented for the individual approaches, the proposed methodology could be easily adapted to other domains where the eye movements provide relevant information.

6.5 Machine learning techniques for improving relevant categories

In the final proposal presented until this point the movement categories considered as relevant by the experts were: eye closure, eye opening, gaze shift to the left and gaze shift to the right. However, two are the categories typically associated with gestural reactions to the sound: gaze shift to the left and gaze shift to the right. This is justified because patients with cognitive decline or severe communication difficulties tend to stay still and passive during the hearing evaluation, but, when they perceive an auditory stimulus they usually tend to direct their gaze to the side on which they perceive the sound. This is a unconscious reaction to the sound that has been considered as consistent by the experts.

After extracting the final features from the video sequences, machine learning techniques are applied aiming at automatically classifying the data into one of these four classes. Specifically, we have applied several classifiers and at this stage we are going to try to improve the true positive rate of the most important classes by using oversampling techniques. It is expected that the proposed methodology will enable the proper assessment of patients when no interaction is possible with high classification accuracy rates.

There are several aspects that might influence the performance achieved by classification algorithms. It has been reported that one of these aspects is related to class imbalance. A dataset is imbalanced if the classification categories are not approximately equally represented. In this situation the learning system may have difficulties to learn the concept related to the minority class. Most of real-world datasets are predominately composed by one class with an abundant number of instances (known as the majority or negative class) and the other class with few number of samples (known as the minority or positive class). Moreover, the classes which are underrepresented are usually the cases under consideration on the study; therefore, its correct identification becomes even more important. This problem has gained much importance in the last years because of its presence in lots of real-world applications such as medical diagnosis, software defects detection, finances, or bioinformatics.

The imbalance in class distribution poses a major challenge to standard machine learning algorithms because the search process that is embedded in most of these is guided by a global search measure that does not consider the differences in the number of instances that belong to each class. Machine learning algorithms is typically evaluated using predictive accuracy, something that might not be appropriate when the data is imbalanced and/or the costs of different errors vary markedly. In this manner, the instances of the minority class are usually neglected during the model construction as its identification is performed using specific learning rules and its representation inside the dataset is not strong enough. These specific rules are usually ignored in favor of more general rules, which are precisely the rules that cover the majority class.

Learning with imbalanced data is one of the emergent challenges in machine learning. The machine learning community has addressed the issue of imbalanced datasets in two different ways: by assigning distinct costs to training examples (Domingos, 1999) or by re-sampling the original dataset. The re-sampling can be achieved by over-sampling the minority class or/and under-sampling the majority class (Kubat & Matwin, 1997; Japkowicz, 2000). The re-sampling is applied until the dataset is nearly balanced, before feeding it into any classifiers. Over-sampling methods (Bunkhumpornpat, Sinapiromsaran, & Lursinsap, 2009; Han, Wang, & Mao, 2005) address the imbalance by adding to the new dataset instances from the minority class in order to gain importance, under-sampling methods (Laurikkala, 2001; García & Herrera, 2009) aim of equalizing the number of examples of each class by deleting instances from the majority class; and hybrid methods (Batista, Prati, & Monard, 2004) combine the two previous approaches, usually starting with an oversampling step that creates new samples for the minority class and then applying undersampling in order to delete samples from the majority class. All these techniques are sampling the dataset until the classes are approximately equally represented.

6.5.1 Experimental results

The methodology is going to been tested on a normalized dataset which is composed of 1194 samples and 66 features. Four different classes are included: eye closure (class EC) (408 samples), eye opening (class EO) (314 samples), gaze shift to the left (class GL) (233 samples) and gaze shift to the right (class GR)(239 samples). Notice that the most important classes to detect are gaze shift to the left and gaze shift to the right, since they are associated with gestural reactions to the sound. However, not every patient shows the same reaction, so the detection of the other eye movements could be relevant in particular cases. Table 6.4 shows the classification results using different classification algorithms, trying to determine which one is more appropriate for the problem at hand.

Table 6.4: Classification results using a 10-fold cross validation for different algorithms

			True positive rate			
Method	Accuracy	ROC Area	Class EC	Class EO	Class GL	Class GR
C4.5	0.9000	0.9343	0.9152	0.9083	0.8948	0.8666
Naive Bayes	0.9331	0.9848	0.9487	0.9559	0.9447	0.8713
k-NN	0.9593	0.9721	0.9747	0.9532	0.9619	0.9368
SVM	**0.9712**	0.9899	**0.9838**	**0.9623**	**0.9787**	**0.9579**
RandomForest	0.9585	**0.9972**	0.9714	0.9567	0.9666	0.9357

In light of the results depicted in Table 6.4, the highest accuracy and TP rates were obtained with SVM. This fact is not surprising at all, since Support Vector Machines have demonstrated to be successful modeling and prediction tools for a variety of applications. Regarding the true positive rates, it is easy to note that these algorithms are often biased towards learning the majority class (López, Fernández, García, Palade, & Herrera, 2013), leading to higher misclassification rates for the minority class instances (gaze left and gaze right). Furthermore, it is worth pointing out that the minority classes are the ones that have the highest interest in the problem at hand, so we will try to improve their classification.

Numerous techniques are used to deal with imbalanced datasets in classification, among all of them we would like to highlight the use of over-sampling methods, which aim to balance the class distribution by adding to the new dataset instances from the minority class. In particular, the SMOTE algorithm (Synthetic Minority Oversampling TEchnique) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) is an over-sampling method that adds synthetic minority class examples to the original dataset until the class distribution becomes balanced. In order to do so, the SMOTE algorithm generates the synthetic minority class examples using the original minority class examples in the following way: the SMOTE algorithm searches the k nearest neighbors of the minority class sample that is going to be used as base for the new synthetic sample. Then, in the segment that unites the minority class sample with one or all of its neighbors, a synthetic sample is randomly taken and is added to the new oversampled dataset.

As stated before, the usual approach is to replicate the minority classes until the class distribution becomes balanced, so we applied an over-sampling rate of 100% on the classes gaze left and gaze right. In the second row of Table 6.5 we can see the results of this experiment, in which the true positive rates for the two classes that are the most important in this problem (gaze left and gaze right) have increased, while maintaining the global accuracy. Based on some ideas raised in (del Río, López, Benítez, & Herrera, 2014), and considering that what was of real interest for us was to improve the identification of these two minority classes, we performed some experiments increasing the oversampling rate to 200%, 300% and 400%, as can bee seen in the remaining rows of Table 6.5. For the sake of comparison, the first row shows the classification results when no oversampling technique is applied.

Table 6.5: SVM classification results using a 10-fold cross validation and applying different levels of oversampling

			True positive rate			
% Oversampling	Accuracy	ROC Area	Closure	Opening	Left	Right
0	**0.9712**	0.9899	**0.9838**	0.9623	0.9787	0.9579
100	0.9703	0.9907	0.9793	0.9589	0.9829	0.9607
200	0.9695	0.9898	0.9773	**0.9625**	0.9787	0.9615
300	**0.9712**	0.9901	0.9769	0.9589	0.9860	0.9667
400	0.9686	**0.9909**	0.9723	0.9528	**0.9882**	**0.9707**

The best results for the gaze shifts were obtained when the level of oversampling is 400%. In this case, we achieve maximum true positive rates for the detection of the most important classes (gaze left and gaze right) whilst the global accuracy has been deteriorated in less than 0.3%. However, with 300% oversampling, we can maintain the global accuracy with an important increase in the detection of gaze shifts. Notice that, with these configurations, the TP rates for the detection of closure and opening decrease, but they are still in acceptable values considering their lower incidence in the detection of gestural reactions.

6.5.2 Discussion

In previous Chapters we have proposed a methodology for the correct classification of eye gestural reactions to the auditory stimuli in order to facilitate the hearing assessment of patients when no cooperation exists. There are four eye movement categories: eye closure, eye opening, gaze shift to the left and gaze shift to the right. The two gaze shifts are typically associated with gestural reactions to the sound, so its correct detection is of utmost importance. However, these two categories are the ones with the smallest number of classes, so we proposed to include oversampling techniques in our methodology.

After extracting the features from the existing videos, we applied several state-of-the-art classification algorithms to determine which one was more appropriate. Since SVM clearly outperformed the other techniques, we decided to use it for our case study. Then, and trying to increase the true positive rates for the two classes of interest (gaze left and gaze right), we opted for applying oversampling techniques, in particular, SMOTE. The novelty of this work lies in the fact that, instead of replicating the minority classes until they are balanced with respect to the majority classes, we decided to apply higher oversampling rates so that the minority classes have now much more instances than the previous majority classes. In this way, our hypothesis was that by forcing the classifier to learn from a data where the two classes of interest were the majority classes, their true positive rates would increase. Indeed, by choosing an oversampling rate of 300%, we were able to obtain true positive rates over 0.986 and 0.966 with a global classification accuracy over 97%.

In conclusion, the proposed methodology is able to classify the different movements with reasonable detection rates, especially for the two classes that deserve more importance. This methodology has shown encouraging and positive results, paving the way to its inclusion in an fully automated tool. As future work, there is a need of more video sequences of patients with this particular conditions so a more comprehensive analysis can be conducted. Also, we plan to try feature selection methods to check if accuracy can be improved.

Chapter 7

Conclusions

Hearing loss is the most common sensory deficit in the elderly, and it is becoming a severe social and health problem. Impaired hearing results in distorted or incomplete communication, thus significantly impacting everyday life, causing loneliness, isolation, withdrawal, dependence, and frustration, as well as communication disorders. The proper assessment of hearing and the use of hearing aids or hearing rehabilitation process will improve the quality of life of people with hearing disabilities.

The evaluation of the hearing capacity has sometimes some associated assessments or particular development limitations which hinder the expert's work. Therefore, the development of a computer-based analysis is highly desirable for assisting the audiologists and providing more objectivity and accuracy to their traditional assessments.

Several automated assessments related with the evaluation of the hearing capacity have been proposed and developed throughout this PhD thesis. The two main aims are the measurement of the response times and the detection of eye-based gestural reactions as a response to the sound. These automated assessments are not intended to override the judgment of an expert, but they should prove helpful in the conduct of clinical routine and research.

In Part I, a screening method to automatically measure the response times during the performance of an audiometric evaluation has been presented. In this case, the positive response most commonly requested by the audiologist was the one that we have considered as reference to measure the response times: the patient's hand raising. The process is carried out using color analysis techniques for the detection of the hand, obtaining an accuracy over 99% for the detection of this reaction. Besides, the auditory stimuli light indicator was properly detected (for two different models of audiometers), in order to synchronize stimulus and response. By the combination of both auditory stimuli and hand raising detection, the method finally provides a measure of the patient's reaction times. Several metrics were evaluated in order to combine the patient's response times; although every measure has its peculiarities, all of them allow us to establish a gap between the "normal" patients and the "slow" ones. The proposed method not only precisely quantifies the patient's speed of response but it also allows to objectively identify patients abnormally slow (those in which the experts are concerned because they could have other cognitive problems or pathologies). Furthermore, it paves the way to the development of more complex clinical studies.

Once the measurement of the response times was successfully resolved, the remaining work of this thesis was focused on the detection of the unconscious gestural reactions that patients with cognitive decline or severe communication difficulties show as a response to the auditory stimuli. Throughout Part II, several approaches were proposed in order to finally obtain a methodology which solves this second main aim.

Firstly, in Chapter 4 the detection of the eye gestural reactions is addressed using as base optical flow information and machine learning algorithms. The obtained results showed promising results in the detection and classification of these unconscious eye movements. Although accuracy rates are not optimum, the proposed method it is still able to detect all the reactions of the test dataset. Thus, considering that it is the first fully automated approximation proposed for this domain, the results are encouraging.

In Chapter 5 an alternative method for the detection and identification of eye gestural reactions was proposed. This task is accomplished using information about the color distribution of the sclera. This solution shows interesting classification accuracies (over the 82.8%) and it is able to detect all the reactions of the test dataset.

Finally, Chapter 6 the detection of eye-based gestural movements is accomplished by the combination of the optical flow information and the analysis of the color distribution of the sclera. The optical flow is more focused on global features, while the color distribution of the sclera works in a local level. They both provide complementary results, so its combination is highly recommended. The results obtained with this combined proposal allow us to conclude that the inclusion of the information provided by the color distribution of the sclera improves the results obtained with the use of the optical flow.

For all the approaches presented throughout this thesis one of the main premises was to modify as minimum as possible the traditional protocol of the audiometric assessment, in order to not influence on the spontaneous behavior of the patient. The conducted experiments have proven the validity of the methodology placing the video camera behind the audiologist (which is seated in front of the patient), an unobtrusive location that minimally perturbs the patient's behavior and provides enough image resolution to apply the proposed methodology obtaining proper results. The most accurate results were obtained for the SVM classifier, with a classification accuracy of the 97.46%, which is highly accurate result for this domain and it improves all the previous approaches.

Furthermore, a final improvement was applied in order to improve the accuracy of the most relevant categories (gaze shift to the left and gaze shift to the right) by using machine learning techniques. Trying to increase the true positive rates for the two classes of interest, we opted for applying oversampling techniques, in particular, SMOTE. The hypothesis was to increase their true positive rates by forcing the classifier to learn from a data where the two classes of interest were the majority classes. Indeed, by choosing an oversampling rate of 300%, we were able to obtain true positive rates over 0.986 and 0.966 with a global classification accuracy over 97%.

This final approach presents a methodology for the detection and classification of eye gestural movements as a response to auditory stimuli during the performance of audiometric evaluations. This methodology will allow the automatic detection of positive unconscious gestural reactions which can be interpreted as positive reactions in the hearing assessments of patients with cognitive decline or other severe communication difficulties. When no cooperation exists, the audiologist focus his attention of the detection of these particular reactions, but the development of an automated tool will facilitate his work, avoid subjectivities and make it less error prone.

In clinical terms, the manual analysis conducted by the audiologists can be automated with the main benefit of being unaffected by subjective factors when evaluating non-cooperative patients. Besides the fact that the automatic proposal produces unbiased results, it also saves times for the experts and provides a detailed identification of the eye-based gestural reactions. In this sense, the audiologist can have a more objective detection, which means a great help in the proper evaluation of the hearing hearing assessment. The proper diagnosis of hearing loss will allow the prescription of appropriate hearing aids, thus improving the quality of life of these particular group of patients.

Besides, the fact that the recorded video sequences can be visualized after they have been analyzed by the proposed methodology will facilitate the training of new experts in the detection and interpretation of unconscious reactions. As mentioned in the Introduction, since these reactions are quite inconsistent and subtle their proper interpretation requires broad experience from the audiologist. If we provide to the inexperienced audiologist a tool where they can observe previous assessments where the eye-based gestural reactions are properly labeled we will facilitate their training.

Although in this thesis we are focused on the audiometric domain, the proposed methodologies for the detection of eye movements could be easily adapted to other domains where these movements could provide relevant information.

7.1 Further research

In the case of the analysis of the eye-based gestural reactions detection there is a need of more video sequences of patients with cognitive decline or other severe communication disorders and which show gestural reactions as a response to the sound so a more comprehensive analysis can be conducted. The problem here is the difficulty in obtaining the required permissions for recording this particular group of patients. Most of these patients are entered in special centers and special permits and authorizations are needed to record them.

Although the final methodology only considers four movement categories (the ones that were highlighted by the experts as the most relevant) it could be easily extended for new categories. In this case, the classification training should be rebuild by using the SVM classifier.

The behavior of the method for the detection of eye movements as a reaction to the sound suggest that it may be a useful for different domains. With this idea in mind, it might be proposed the adaptation of our methodology to different domains where eye movements could provide relevant information.

Appendix A

Face location

The location of the face is an initial step in different stages of this thesis. Proper face location will allow us to narrow the subsequent search areas. This way, this initial location reduces the computational cost of the next steps and makes them less error prone.

Although the location of the face is a natural process for a human being, it becomes a challenging task in computer vision. In addition to the inherent complexity of defining a face for a computer, the variations in scale, orientation, pose, facial expression, lighting conditions, and background, increase the complexity of the problem.

Different methods have been developed in order to detect faces in a scene. First, earliest and easiest proposals worked in monocolor backgrounds or with a predefined static background, in these cases, the face was obtained by removing the background. Other approaches (e.g. (Sandeep & Rajagopalan, 2002)) use color information as base, seeking for skin color regions. The use of color as the main base of the method may involve limitations with some skin colors or with varying lighting conditions. If we work with video sequences instead of static images, we can use motion information to find the face if we can consider that face is almost always moving (Graf, Cosatto, Gibbon, Kocheisen, & Petajan, 1996). The problem with motion information arrives when we have a non static background. There also exist many other different alternatives: using feature analysis, active shape models, neural networks, and so on.

In our case, the domain is very stable in terms of location: the audiologist is always seated in front of the patient and the video camera is located behind the audiologist to ensure that the patient's face will always be recorded in frontal position. This particular setup can be observed in Figure A.1. The certainty of having a frontal position of the patient's face allow us to apply the Viola and Jones approach (Viola & Jones, 2001).

The Viola-Jones detector is a general object detection framework which provides competitive object detection rates in real time. It can be trained to detect a variety of different objects; however, its initial motivation was to provide a solution for the face detection task. As consequence of this, an optimized classifier for the face detection was obtained. Particularly, a classifier for the detection of frontal faces is available in the OpenCV library. The fact that faces must be in frontal position in order to be correctly detected may slightly compromise the requirement for being unconstrained, but considering that the detection algorithm most often will be succeeded these demands seem quite reasonable. This classifier is not as flexible as other approaches, but it is low computational and very robust for the detection of frontal faces, so it is a good solution for this domain.

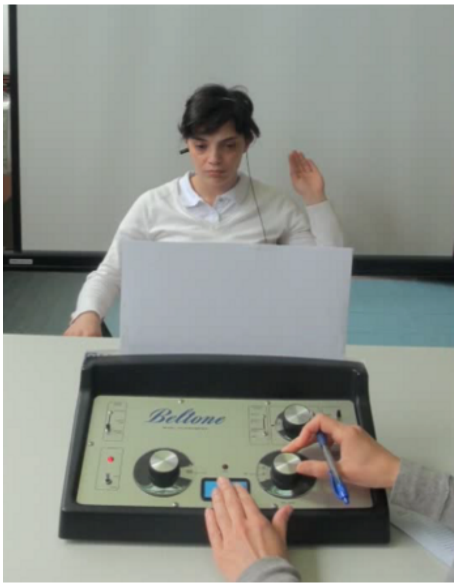


Figure A.1: Face detection layout.

The basic principle of the Viola-Jones algorithm is to scan a sub-window capable of detecting faces across a given input image. The algorithm has mainly four stages: Haar feature selection, creating integral image, Adaboost training algorithm and cascade classifiers.

Some samples of face detection can be seen in Figure A.2.

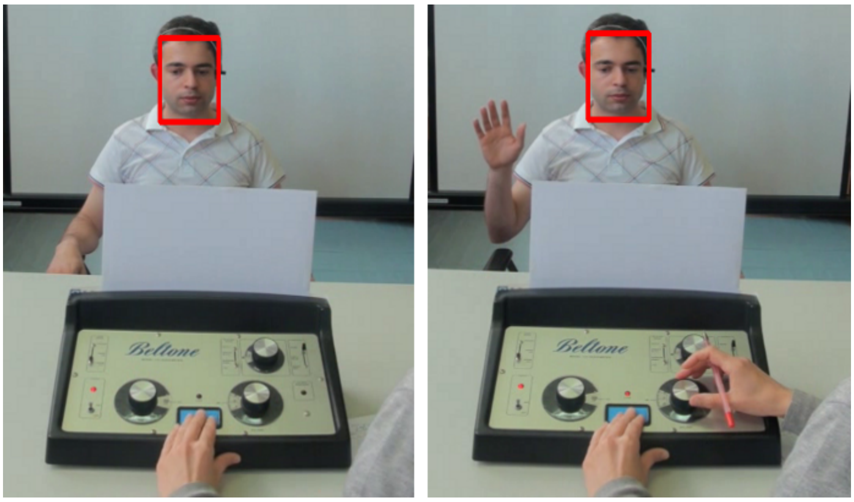


Figure A.2: Face detection samples at different times during the test.

Appendix B

TSL color space

The TSL (Tint, Saturation and Lightness) is a perceptual color space proposed by Terrillon and Akamatsu (2000) with the primarily purpose of providing an efficient detection and location of human faces in static images. TSL has been selected as better color space to extract skin color from complex backgrounds because it has the advantage of extracting a given color robustly while minimazing illumination influnce. Robustness under changing lighting conditions, viewpoint or scale is achieved by a color space that effectively separates the chrominance and luminance using a proper model for the distribution of human skin chrominance.

This color space defines color as a combination of *tint* (the degree to which a stimulus can be described as similar to or different from another stimuli that are described as red, green, blue, yellow, and white, can be thought of as hue with white added), *saturation* (the colorfulness of a stimulus relative to its own brightness), and *lightness* (the brightness of a stimulus relative to a stimulus that appears white in similar viewing conditions).

In (Terrillon & Akamatsu, 2000), authors compare the efficiency of multiple color spaces in the face detection task (among them, the normalized color space TSL and a version without illumination TS) by conducting several experiments. They conclude that the TS space is the one that provides better results in segmentation and more robust face detection. The advantages of TSL color space lie within the normalization within the RGB-TSL transform. Utilizing normalized r and g allows for chrominance spaces TSL to be more efficient for skin color segmentation. Additionally with this normalization, the sensitivity of the chrominance distributions to the variability of skin color is significantly reduced, allowing for an easier detection of different skin tones.

A normalized chrominance-luminance TSL space is a transformation of the normalized RGB into more intuitive values, close to hue and saturation in their meaning. The transformation required for the normalized TSL color space are the ones detailed in (B.1), (B.2) and (B.3). r' y g' are computed according to (B.4) and (B.5), which are based on the amounts of red and green in zero (since when the proportions of red, green and blue are equal $r = g = \frac{1}{3}$, then $r' = g' = 0$), while through equations (B.6) and (B.7) the proportions to red and green are computed. It can be noted how r y g match with the chromaticity values derived from RGB color space and that L' is obtained by the usual formula of luminance. Using these transformations values for T', S', and $L' \in [0, 1]$ are obtained, since the components R, G and B have been initially transformed to that interval.

$$L' = 0.299R + 0.587G + 0.114B \tag{B.1}$$

$$S' = \sqrt{\frac{9}{5}(r'^2 + g'^2)} \tag{B.2}$$

$$H' = \begin{cases} \frac{1}{2\pi} arctan\left(\frac{r'}{g'}\right) + \frac{1}{4} & \text{si } g' > 0 \\ \frac{1}{2\pi} arctan\left(\frac{r'}{g'}\right) + \frac{3}{4} & \text{si } g' < 0 \\ \frac{1}{2} & \text{si } g' = 0 \end{cases} \tag{B.3}$$

$$r' = r - \frac{1}{3} \tag{B.4}$$

$$g' = g - \frac{1}{3} \tag{B.5}$$

$$r = \frac{R}{R+G+B} \tag{B.6}$$

$$g = \frac{G}{R+G+B} \tag{B.7}$$

Finally, T', S' y L' are rescaled in order to use the full range offered by 8-bit through formulas (B.8), (B.9) and (B.10).

$$T = 255T' \tag{B.8}$$

$$S = 255S' \tag{B.9}$$

$$L = 255L' \tag{B.10}$$

Appendix C

Interest operators for the detection of gestural reactions

The methodology proposed in Chapter 2.1 makes use of the optical flow. Optical flow needs an interest operator as pre-step. Good Features to Track (Shi & Tomasi, 1994) is the interest operator associated to optical flow by default, anyway, we wanted to test the behavior of other interest operators in our particular domain. Since the results of the optical flow depend on the interest points that the method receives as input, choosing these interest point is a crucial step, since the following steps will be highly affected by the results of this stage of the methodology. In this Appendix, different interest operators were studied in order to find the most appropriate.

Firstly, it is important to describe the features that define an interest point. Usually, these points are define by qualities like: well-defined position on the image, mathematically well-founded, rich in terms of local information and stable to global perturbations. These properties are assigned regularly to corners or to locations where the color of the region suffers a big change.

Considering this, we want to choose those interest points that can be easily matched by the optical flow. To select them, an analysis between different interest operators was conducted. Each of these methods has different foundations, and consequently, a different way of performing, so the results that one of them provides can be very different from those provided by any of the others.

The interest operators analyzed here are: Harris corner detector (Harris & Stephens, 1988), Good Features to Track (Shi & Tomasi, 1994), SIFT (Lowe, 2004), SURF (Bay, Ess, Tuytelaars, & Van Gool, 2008), FAST (Rosten & Drummond,

2005) (Rosten & Drummond, 2006) and a particular version of Harris with a little modification. Also different classification techniques were tested, in order to find the best detector-classifier combination.

C.1 Experimental results

Experiments were conducted over a dataset of video sequences from audiometric evaluation considering only the eye region. The original video sequences (before considering only the eye region) are Full HD resolution (1080x1920 pixels) with a frame rate of 25 FPS. Despite the high resolution of the original images, it is important to take into account that the resolution of the eye region will not be as optimal (since the recorded scene is very wide), and moreover, lighting conditions will affect considerably.

Tests were conducted with 9 different video sequences, each one from a different patient. Each audiometric test takes between 4 and 8 minutes. Considering that the video sequences have a frame rate of 25 FPS, an average video sequence of 6 minutes will have 9000 frames, implying a total number of 81000 frames for the entire video set. Taking into account that reacions only occur in a timely, we finally have 128 pairs of frames to be considered. Since each eye is considered separately, the test set will consist of 256 movements. These movements are labeled into four classes depending on the movement they represent (see Table C.1) and they are classified according to the methodology presented in Chapter **??**.

Table C.1: Number of samples for each class of movement.

Eye opening	Eye closure	Gaze left	Gaze right
80	82	46	48

Four different experiments were conducted in order to find the best detector for this domain. The three experiments are:

1. Find the best classifiers.
2. Find the best configuration parameters for each interest points detector.
3. Evaluate the detector-classifier results.
4. Evaluate the classification by classes

C.1.1 Classifier selection

In this part of the survey, different classifiers were tested with the aim of selecting the three best methods for applying them on the following tests. The considered classifiers are provided by the WEKA tool (Hall et al., 2009), and they are: Naive Bayes, Logistic, Multilayer Perceptron, Random Committee, Logistic Model Trees (LMT), Random Tree, Random Forest and SVM.

In order to obtain these results, 18 test were conducted for each pair detector-classifier, where each one of these tests is the result of a 10-fold cross validation. Computing the average per method (without considering the detector used) we obtain the results shown in Figure C.1. As it can be observed on this graph, all the methods obtain an accuracy between 60 and 75%. Worst results are observed for Naive Bayes, Logistic and Random Tree. Best results are obtained for SVM, followed by Random Committee and Random Forest, so these are going to be the three classifiers considered for the next survey.

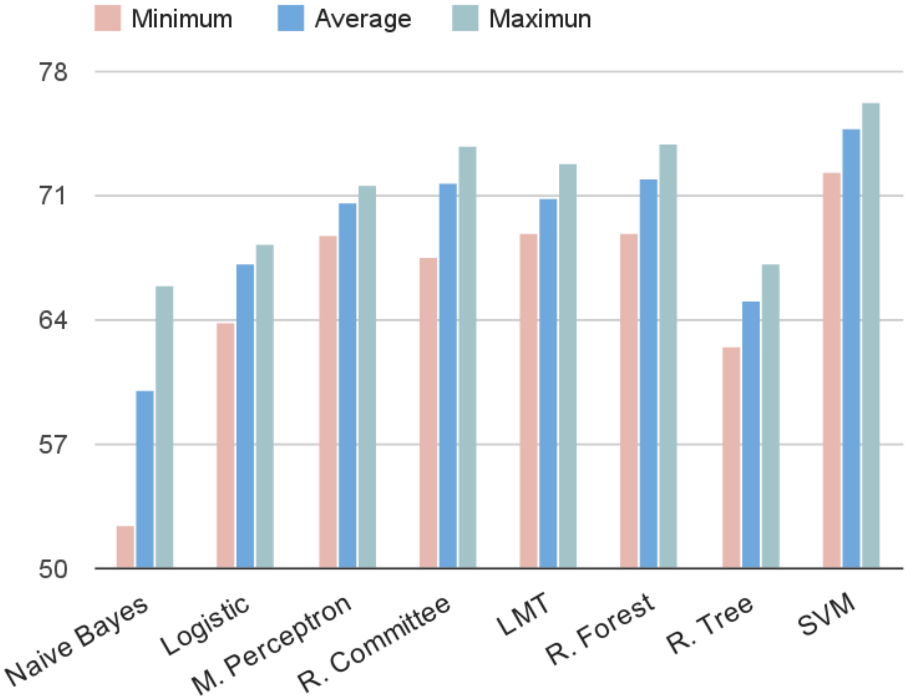


Figure C.1: Minimum, maximum and average success percentage by classifier.

C.1.2 Adjustment of parameters

The methodology proposed in Chapter **??** makes use of different parameters that are going to be adjusted for the different interest operators in this experiment. The parameter adjustment is performed dependently on the method used. The parameters studied in this experiment are:

- Number of detected points: it indicated the number of points that the detector must select. Very few point may not be enough to create a proper motion descriptors and a number too high might introduce too much noise.
- Minimum percentage of equal point to remove the movement: sometimes, the detected motion is due to global motion between two frames and not to a real movement within the region (addressed in Chapter **??** Section 4.2.2). This will imply a high number of vectors with the same direction and strength. With the aim of removing this offset component, the parameter λ is introduced. This parameter indicates the required minimum percentage of equal vectors to be considered a global motion, and consequently, discard them.
- Minimum length: very short vectors wil not be representative of movement. In order to choose the representative vectors three classes were established depending on the length of the vector (Chapter **??** Section 4.2.1): u_1 for vectors smaller than 1.5 pixels, u_2 for vectors between 1.5 and 2.5 pixels and u_3 for vector between 2.5 and 13 pixels (vectors larger than 13 pixels will be considered erroneous). Vectors in u_1 are considered too small and are not taken into account for the descriptor, while vectors in u_3 are considered relevant and are always part of the descriptors. The inclusion or not of vectors in u_2 is going to be studied on this section.

Harris corner detector

Harris has a particular behavior, it detects few points concentrated in areas with high contrast. The obtained results are represented in Figure C.2. Each line represents a classifier (Random Committee, Random Forest and SVM), distinguishing between using only u_3 vectors (green lines) and u_2 and u_3 vectors (blue lines).

It can be observed that, the higher λ is, the better the results are. Moreover, the inclusion of vectors in u_2 provides worst results. It can be noticed that in Figure C.2(c) there is a value nearly the 100% of accuracy. This value is an outliers that

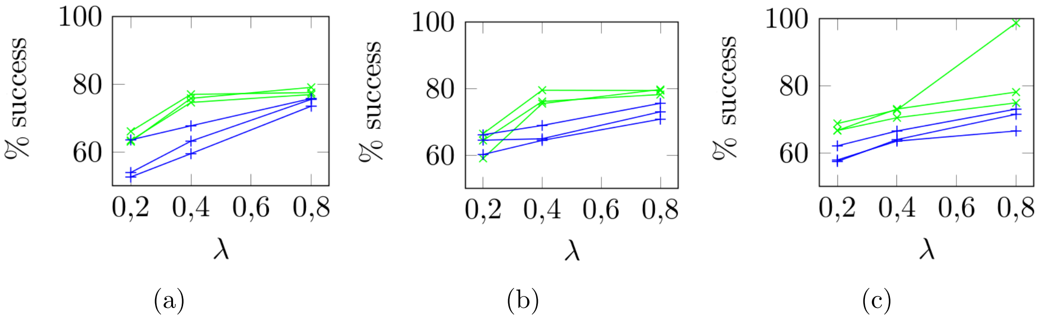


Figure C.2: Classification results for Harris. Green lines for u_3 vectors and blue lines for u_3 and u_2 vectors. Each of the three lines for each color corresponds to a different classifier. (a) For 40 points of interest. (b) 80 points. (c) 160 points.

may not be repeatable, since it breaks the tendency of other values, and for this reason it should not be considered. However, it confirms the tendency that, with higher λ values, the accuracy increases.

Good features to track

This interest operator was specifically designed for its use together with the optical flow. Figure C.3 shows the obtained results for this operator. As it can be observed, results are quite consistent regardless of the values of the parameters. The behavior is better for low values of λ, and also considering 80 points of interest. Although results are very similar, including vectors in u_2 slightly increases the success rate in some cases.

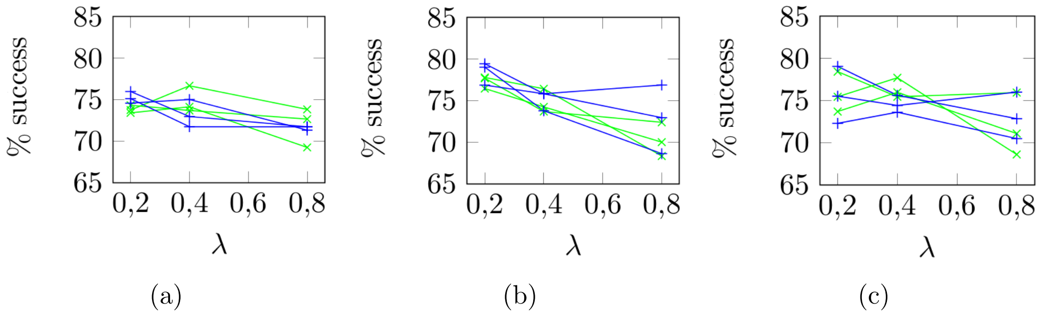


Figure C.3: Classification results for Good Features to Track. Green lines for u_3 vectors and blue lines for u_3 and u_2 vectors. Each of the three lines for each color corresponds to a different classifier. (a) For 40 points of interest. (b) 80 points. (c) 160 points.

SIFT

The SIFT detections are quite similar to the detections obtained with Good Features to Track. Its results are also broadly similar (see Figure C.4). However, unlike the previous method, in this case the obtained results for 80 points of interest are slightly worse than the ones obtained for 40 or 160 points of interest. The λ parameter does not affect the results too much. Inclusion of the intermediate vectors (u_2) offers also better results.

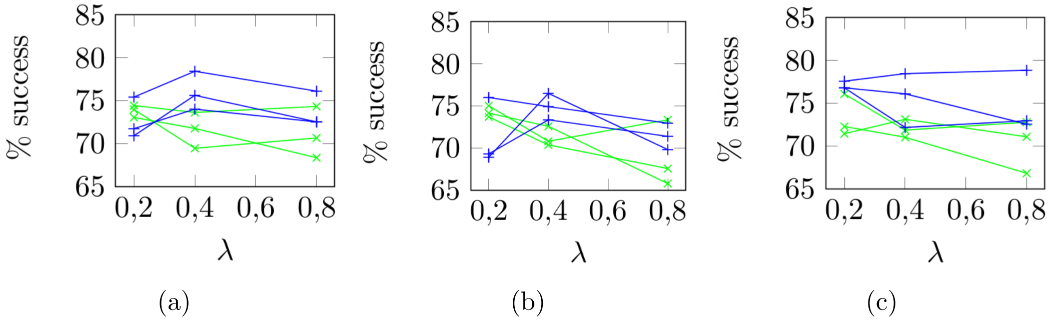


Figure C.4: Classification results for SIFT. Green lines for u_3 vectors and blue lines for u_3 and u_2 vectors. Each of the three lines for each color corresponds to a different classifier. (a) For 40 points of interest. (b) 80 points. (c) 160 points.

SURF

SURF detector is a very particular method since it is very selective about the detected points. With images from our domain it is not possible to detect more than 35-40 point, even applying very permissive thresholds. Due to this particularity, the only results obtained are the ones shown in Figure C.5. Better results are obtained when including vectors in u_2, for which the most appropriate value of λ is 0.8.

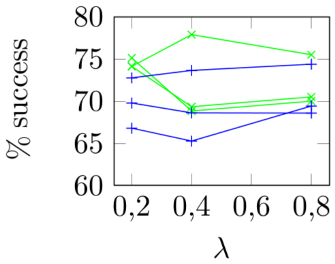
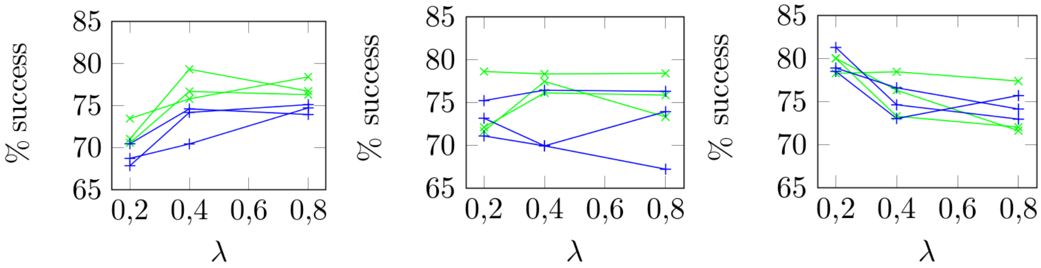


Figure C.5: Classification results for SURF. Green lines for u_3 vectors and blue lines for u_3 and u_2 vectors. Each of the three lines for each color corresponds to a different classifier.

FAST

Through a visual observation, it can be established that the interest points detected by FAST are quite significant for this domain. Charts with the results can be observed in Figure C.6. Regarding the length of the vectors, results vary according the number of considered points. For 40 and 80 points best results are obtained only considering the strong vectors (u_3), while for 160 points best results are obtained when considering vectors in u_3 and u_2. For 40 points of interest the most appropriate is a high value for λ, for 80 points the results are quite stable regardless of the value of λ, and for 160 points low values for λ offer better results.



(a) (b) (c)

Figure C.6: Classification results for FAST. Green lines for u_3 vectors and blue lines for u_3 and u_2 vectors. Each of the three lines for each color corresponds to a different classifier. (a) For 40 points of interest. (b) 80 points. (c) 160 points.

Harris modified

The original Harris detector detects few points in areas where the contrast is high. To achieve a greater separation between points, and therefore more representative points, a location of the local maximums is conducted. Also a thresholding is applied over the Harris image, and finally, the *and* operation is computed with these two images, obtaining this way more distributed interest points.

Results for this alternative version of Harris are charted in Figure C.7.

These results are similar to the ones obtained with FAST. In the general case, better results are obtained considering onlu vectors in u_3. For 80 and 160 interest points, the best behavior occurs for the lower value of λ (0.2). In the case of considering 40 points, best results occur for λ equal to 0.4.

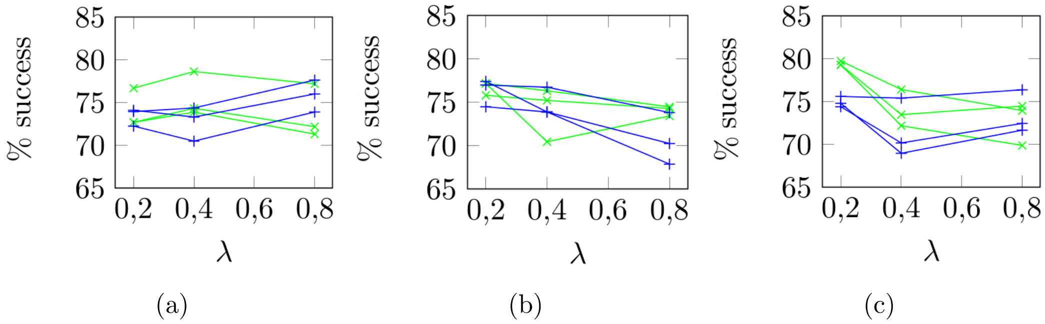


Figure C.7: Classification results for Harris modified. Green lines for u_3 vectors and blue lines for u_3 and u_2 vectors. Each of the three lines for each color corresponds to a different classifier. (a) For 40 points of interest. (b) 80 points. (c) 160 points.

C.1.3 Final evaluation of the results

Once the behavior of the different methods in relation to their configuration parameters has been analyzed, we are going to compare here the results of the different methods considering only their best configuration. The optimum configuration parameter and classifier for each one of the considered interest operators is detailed in Table C.2.

Table C.2: Optimum configuration parameters for each method.

Method	Classifier	No. points	λ	Vectors
Harris	SVM	160	0.8	u_3
Good Feature	R. Forest	80	0.2	u_2 & u_3
SIFT	SVM	160	0.8	u_2 & u_3
SURF	SVM	40	0.4	u_3
FAST	R. Forest	160	0.2	u_2 & u_3
Harris mod.	SVM	160	0.2	u_3

Results are shown graphically for better understanding. In order to assess the capacity of each one of the interest operators in the detection of relevant movements, the obtained descriptors are compared with the ground truth of the movements previously labeled by the experts.

The graph presented in Figure C.8 shows the true positive and false positive rate ($T_{tp}(d)$ and $T_{fp}(d)$ respectively). It can be noted that SURF has a good value for the false positive rate, but a poor value for the true positive rate. SIFT is the opposite case, it has a good value for the $T_{tp}(d)$ but poor for the $T_{fp}(d)$. The same happens with Harris, which offers intermediate values for both rates. Instead, FAST, Good Features to Track and Harris modified show good values for both rates. Good Features and FAST offer almost equivalent results, while Harris modified has a worst $T_{tp}(d)$ but it is compensated with a optimum $T_{fp}(d)$ rate.

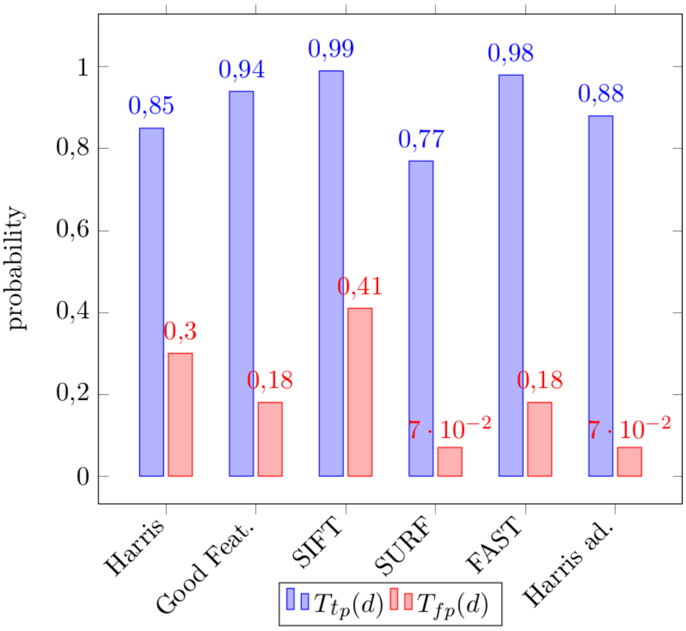


Figure C.8: True positive rate ($T_{tp}(d)$) and false negative rate ($T_{fp}(d)$) for the different methods.

Given the previous results, only FAST, Good Features to Track and Harris modified are considered for the last evaluation. Figure C.9 shows the true positive rate in detection ($T_{tp}(d)$), the specificity ($1 - T_{fp}(d)$) and the true positive rate in classification ($T_{tp}(c)$).

All the methods have a similar value for the true positive rate in classification ($T_{tp}(c)$). FAST offers better results than Good Feature for the three evaluated measures; so between these two methods, FAST would be chosen. Comparing between FAST and Harris modified, it can be observed that the $T_{tp}(c)$ is quite similar, while the $T_{tp}(d)$ and the specificity are slightly opposite. FAST offers better results for the $T_{tp}(d)$ while with Harris better results are obtained for the specificity. The decision of choosing one or another depends on the suitable results for this domain. If we want to reduce the number of false positives Harris is the best solution, while if the true positive detections are more important, FAST is the method that should be

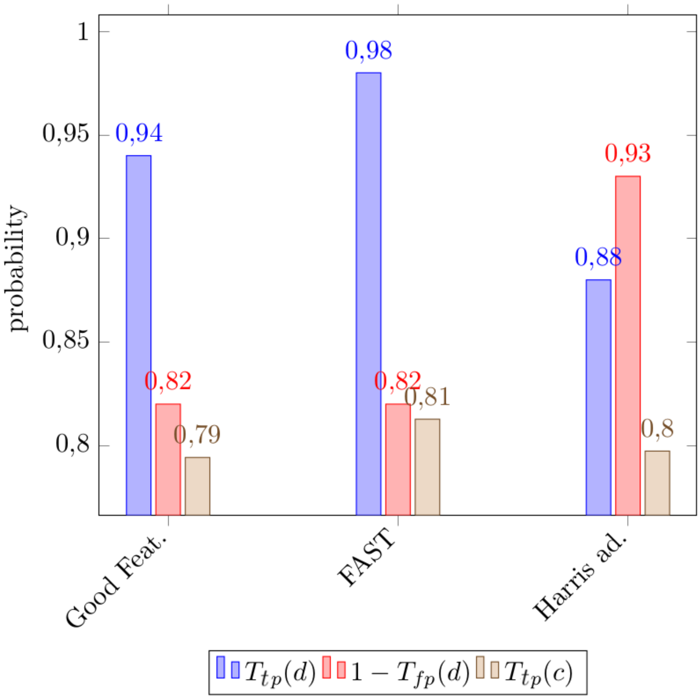


Figure C.9: True positive rate ($T_{tp}(d)$) and specificity ($1 - T_{tp}(d)$) for detection and true positive rate for classification ($T_{tp}(c)$).

chosen. Anyway, the obtained results for Good Features to Track are considerably good so it would be a suited option too.

C.1.4 Classification by classes

A more detailed survey about the results was conducted in order to analyze the classification results by classes. Tables C.3, C.4 and C.5 represent the confusion matrices for the three best interest operators selected from the previous results combined with the best classifier for each one of them, i.e., *FAST* - *Random Forest* (Table C.3), *Good Features to Track* - *Random Forest* (Table C.4) and *Harris ad.* - *SVM* (Table C.5). Where *Open*, *Close*, *Left* and *Right* correspond with the considered movement categories: eye opening, eye closure, gaze shift to the left and gaze shift to the right.

As it can be inferred from these confusion matrices, the classification accuracy for the eye closure movement is always over the 90%. Instead, the worst results are obtained for the movements of gaze shift to the right and gaze shift to the left. None of them is able to achieve the 80% of accuracy; which which worsens in the case of *Harris ad.* and *Good Features to Tack* where the accuracy falls to 60%.

Table C.3: Confusion matrix for *FAST* - *Random Forest*

	Open	Close	Left	Right
Open	0.772	0.050	0.076	0.101
Close	0.012	0.927	0.060	0
Left	0.068	0.205	0.727	0
Right	0.174	0	0.065	0.761

Table C.4: Confusion matrix for *Good Features to Track* - *Random Forest*

	Open	Close	Left	Right
Open	0.818	0.039	0.026	0.117
Close	0.049	0.89	0.061	0
Left	0.075	0.3	0.625	0
Right	0.25	0	0.023	0.727

Table C.5: Confusion matrix for *Harris ad.* - *SVM*

	Open	Close	Left	Right
Open	0.811	0.04	0.027	0.122
Close	0.012	0.901	0.086	0
Left	0.033	0.2	0.767	0
Right	0.357	0	0.048	0.595

This negative effect for *Harris ad.* and *Good Features to Tack* is offset by a better behavior for the eye opening movements, where they outperform *FAST*.

It can be observed that the main problem in classification occurs when distinguishing between the classes eye opening and gaze shift to the right, in this case, the miss-classification occurs in both directions, sometimes samples of the class eye opening are classified as gaze shift to the right and sometimes the opposite happens. It can be also observed that there is a considerable number of miss-classifications between classes eye closure and gaze shift to the left, but in this case, it only occurs when classifying samples of the class gaze shift to the left, and not in the opposite direction. Between the remaining classes there are also some errors in the classification, but they are almost negligible.

For a more general understanding, in Table C.6 we show the true positive rate in classification $T_{tp}(c)$ for each class (the diagonal of the confusion matrices previously presented) and the false positive rate in classification $T_{fp}(c)$ computed from the miss-classifications. Likewise, we show the ROC area value, which provides a measure of how good the classification is.

Table C.6: Results by classes for *FAST*, *Good Features to Track* and *Harris ad.*

		Open	**Close**	**Left**	**Right**	**Average**
FAST	$T_{tp}(c)$	0.772	0.927	0.727	0.761	0.813
	$T_{fp}(c)$	0.07	0.077	0.068	0.039	0.066
	ROC area	0.896	0.969	0.88	0.934	0.924
Good Features	$T_{tp}(c)$	0.818	0.89	0.625	0.727	0.794
	$T_{fp}(c)$	0.108	0.093	0.039	0.045	0.08
	ROC area	0.928	0.966	0.929	0.941	0.943
Harris ad.	$T_{tp}(c)$	0.811	0.901	0.767	0.595	0.797
	$T_{fp}(c)$	0.111	0.062	0.056	0.049	0.0075
	ROC area	0.923	0.974	0.915	0.907	0.94

As it can be inferred from Table C.6 all the three methods show successfully results for the ROC area, near to the ideal, where the best of them is *Good Features to Track* with a 0.943 value. From the evaluation of the false positive rate $T_{fp}(c)$ the results confirm the same idea that was deducted from the evaluation of the confusion matrices, which is that the class eye opening is the worst in terms of classification, since it has the highest $T_{fp}(c)$ value. In general terms, the obtained results are promising for both the detection and the classification for any of the interest operators presented here. This indicates that any of these three interest operators could provide optimal results when integrated with in the final methodology.

C.2 Conclusions

The survey here conducted allow us to determine that FAST, Good Features to Track and Harris modified are the most appropriate interest operators for our domain. Since the differences between them are not clear enough, Good Features to Track is the interest operator chosen for the proposed methodology detailed in Chapter **??**. This decision is supported by its accurate behavior (with values nearly to the ones obtained with FAST and Harris modified) and also because it is the interest operator defectively associated with the optical flow.

Appendix D

Interest operators for the detection of candidate points

In this experiment three different interest operators are analyzed with the aim of choosing the one that offers better result in the detection of candidate points to correspond with eyes' corners. This task is accomplished in the methodology presented in Chapter 5 Section 5.1.2. The three interest operators considered for this study are: Harris (Harris & Stephens, 1988), Shi-Tomasi (Shi & Tomasi, 1994) and FAST (Rosten & Drummond, 2005) (Rosten & Drummond, 2006). Each one of them has different foundations, so the provided results may be different.

For this experiment we use three different video sequences recorded during the audiometric evaluation. From each one these videos, 100 frames are selected. If we consider each eye separately, consequently we have 200 samples per video sequence. This way, our test dataset is composed by a total number of 600 samples, 300 for the right eye and 300 more for the left eye.

For the evaluation of this experiment, it is necessary to compute the average and the standard deviation of the detection error. To that end, the expected points are previously labeled. The expected points P_e are compared with the candidate points P_c in order to compute the detection error. Between all the candidate points P_c, the one with the minimum distance to the expected point P_e is labeled as P_{cmin}, the point that better represents and eye corner. With these considerations, we can compute the error the module of the difference between the expected point P_e and the best obtained point P_{cmin}, according to (D.1), where i represents each one of the four eyes' corners.

$$error_i = |P_e - P_{cmin}| \; \forall i \in \{1, 2, 3, 4\} \tag{D.1}$$

Figure D.1 represents how the error calculation is computed. Points represented in red correspond with the expected point P_e and points represented in yellow represent the candidate points P_c provided as a result by the interest operator. It can be observed that for this error calculation only the nearest candidate point to the expected point is considered.

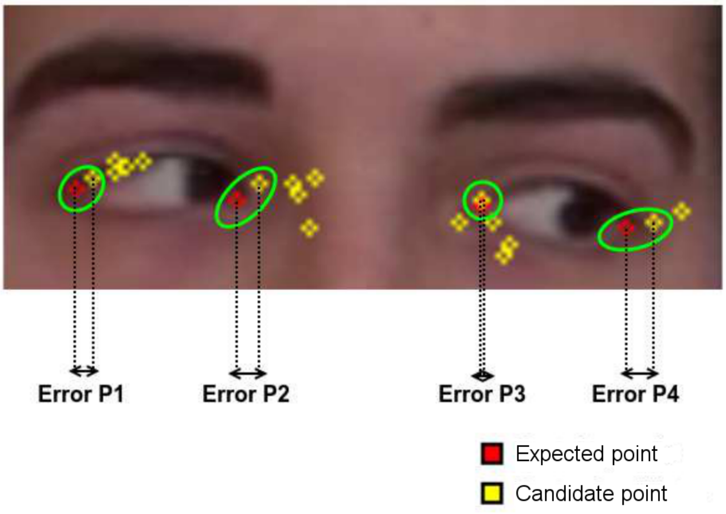


Figure D.1: Error calculation for the interest operators, following (D.1).

Next, the obtained results for each one of the interest operator are going to be presented.

D.1 Harris results

The Harris method is based on corner detection, by understanding as corners those points where the near horizontal and vertical gradients are significant. The main advantage of this approach is that the computations involved do not require high performance times.

Figure D.2 shows the interest points detected by the Harris operator for a sample image from our domain, the obtained interest points are represented in red.

In Table D.1 the average and the standard deviation of the detection error for each one of the eyes' corners (P_1, P_2, P_3 and P_4) are presented. The last column presents the global results (both for average and standard deviation) computing together the four considered corners.

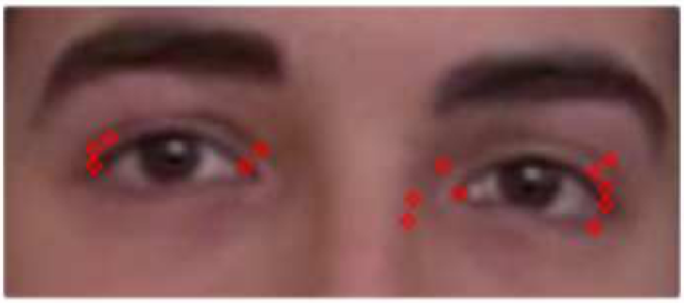


Figure D.2: Sample interest points detected by Harris.

Table D.1: Detection error results for Harris.

	Error	Error	Error	Error	Error
	P_1	P_2	P_3	P_4	(P_1, P_2, P_3, P_4)
Average	1.9545	4.7121	3.2526	2.2400	2.7703
Std. dev.	1.2238	1.9150	2.4984	2.2246	2.1150

D.2 Shi-Tomasi results

The Shi-Tomasi algorithm is based on the Harris corner detector. It has as basis to modify the criteria for choosing the characteristic point.

Figure D.3 shows as red points the interest points detected by the Shi-Tomasi operator for a sample image from our domain.

In Table D.2 the calculation error is showed in terms of average and standard deviation for each one of the eyes' corners. Last column represent the global results when considering all the corners simultaneously.

Table D.2: Detection error results for Shi-Tomasi.

	Error	Error	Error	Error	Error
	P_1	P_2	P_3	P_4	(P_1, P_2, P_3, P_4)
Average	1.0000	0.9481	1.1403	1.0113	1.0227
Std. dev.	0.4444	0.5109	0.6410	0.6336	0.5456

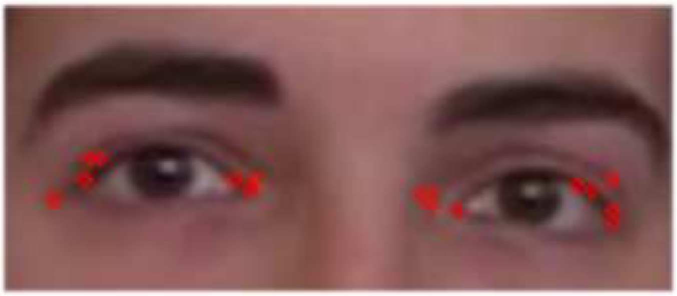


Figure D.3: Sample interest points detected by Shi-Tomasi.

D.3 FAST results

The FAST (Features from Accelerated Segment Test) approach appears with the aim of providing a interest operator with low computational without compromising the quality of the detected points. It is highly suited for real time applications regardless of their complexity.

Figure D.4 shows the interest points provided by the FAST approach for a sample image. These points are represented in red.

Table D.3 presents the detection errors for FAST. First four columns represent eye one of the eyes' corners, and the last column represents the global error in terms of average and standard deviation.

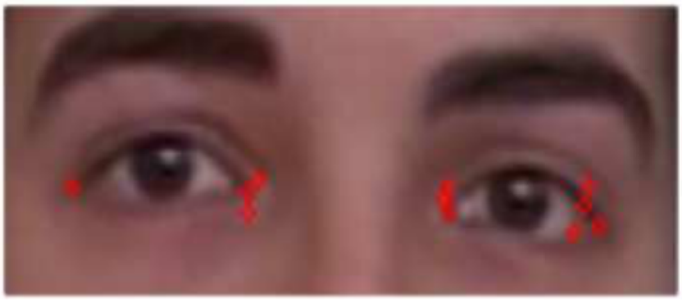


Figure D.4: Sample interest points detected by FAST.

Table D.3: Detection error results for FAST.

	Error P_1	Error P_2	Error P_3	Error P_4	Error (P_1, P_2, P_3, P_4)
Average	1.1324	1.3251	1.6103	1.1754	1.3082
Std. dev.	0.5990	0.8054	1.2853	0.7021	0.8515

D.4 Comparison between interest operators

From the previous results it was built a global table in order to summarize the results. Table D.4 allow to establish a global analysis in order to conclude which one of the three proposed interest operators is most suited for this particular domain. From the results presented in this table it can be concluded that the Shi-Tomasi method is the one that better approximate the eyes' corners, since the detection errors for this method are the smaller ones. For this reason, as a result of this survey, the Shi-Tomasi is the interest operator applied in the methodology presented in Chapter 5 Section 5.1.2.

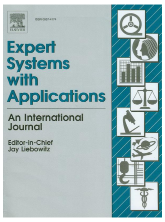
Table D.4: Global results for the detection of candidate points.

	Harris	Shi-Tomasi	FAST
Average	2.7703	1.0227	1.3082
Std. dev.	2.1150	0.5456	0.8515

Appendix E

Publications and other mentions

JCR Journals



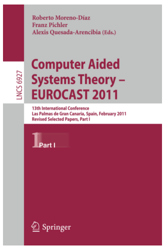
A. Fernández, M. Ortega, B. Cancela, M. G. Penedo, C. Vazquez, LM Gigirey. Automatic processing of audiometry sequences for objective screening of hearing loss. Expert Systems with Applications, 39, 12683-12696, 2012.

A. Fernández, M. Ortega, M. G. Penedo, C. Vazquez, LM Gigirey. A Methodology for the Analysis of Spontaneous Reactions in Automated Hearing Assessment. IEEE Journal of Biomedical and Health Informatics, 2014.

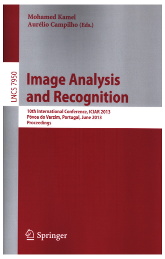
Other Journals

A. Fernández, M. Ortega, B. Cancela, M. G. Penedo. Contextual and Skin Color Region Information for Locating Human Body Parts. Journal of Computer and Information Technology, 1(1), 1-16, 2011.

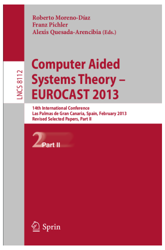
Chapters in Book Series



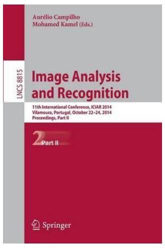
A. Fernández, M. Ortega, B. Cancela, M. G. Penedo. Contextual and Skin Color Region Information for Face and Arms Location. Lecture Notes in Computer Science: Computer Aided Systems Theory, Revised Selected Papers EUROCAST 2011, 6927, 616-623, 2012.



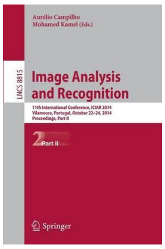
A. Fernández, M. Ortega, M. G. Penedo, B. Cancela, LM Gigirey. Automatic Eye Gesture Recognition in audiometries for patients with cognitive decline. Lecture Notes in Computer Science: International Conference on Image Analysis and Recognition (ICIAR), 7950, 27-34, 2013.



A. Fernández, M. Ortega, B. Cancela, LM Gigirey. Automatic Eye Movement analysis for the assessment of hearing in patients with cognitive impairment: A preliminary study. Lecture Notes in Computer Science: Computer Aided Systems Theory, Revised Selected Papers EUROCAST 2013, 8112, 133-139, 2013.



A. Fernández, M. Ortega, M. G. Penedo. Computer aided hearing assessment: detection of eye gesture reactions as a response to the sound. Lecture Notes in Computer Science: International Conference on Image Analysis and Recognition (ICIAR), 8815 (Part II), 39-47, 2014.

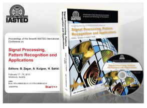


A. Fernández, M. Ortega, M. G. Penedo. Computer aided hearing assessment: detection of eye gesture reactions as a response to the sound. Lecture Notes in Computer Science: International Conference on Image Analysis and Recognition (ICIAR), 8815 (Part II), 39-47, 2014.

cover
not
available

A. Fernández, J. Marey, M. Ortega, M. G. Penedo. Influence of the interest operators in the detection of spontaneous reactions to the sound. Lecture Notes in Artificial Vision: International Conference on Agents and Artificial Intelligence (ICAART) - Revised and Extended Papers, 2014. (Pending of pubblication).

International Conferences



A. Fernández, N. Barreira, M. G. Penedo, L. Lado. Evaluation of the color space influence in face detection. Signal Processing, Pattern Recognition and Applications (SPPRA), 241-247, Innsbruck, Austria, February 2010.

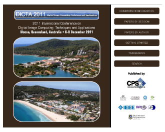
Conference Programme

BARCELONA, 26-29 October 2011

CogART

ISABEL

A. Fernández, M. Ortega, B. Cancela, M. G. Penedo, C. Vazquez, LM Gigirey. Measuring response times to auditory stimuli during an audiometry. 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL), Barcelona, Spain, October 2011.



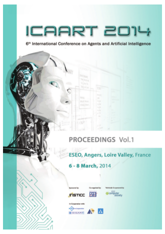
A. Fernández, M. Ortega, M. G. Penedo, B. Cancela, LM Gigirey, C. Vazquez. Automatic analysis of the patient's conscious responses to the emission of auditory stimuli during the performance of an audiometry. International Conference on Digital Image Computing: Techniques and Applications (DICTA), 291-296, Noosa, Australia, December 2011.

A. Fernández. B. Remeseiro. Automatic System for Hearing Loss. II Workshop Internacional en IMAGEN MÉDICA y NUEVAS TECNOLOGÍAS (Invited speaker), A Coruña, Spain, June 2012.

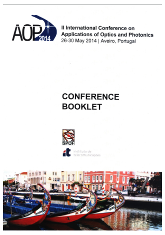
C. Vazquez, A. Fernández, LM Gigirey, M. Ortega, M. G. Penedo. A new method for measuring hearing loss. Part I: a study among a sample of standard population. AHS 2012 - 2nd International Conference on Adult Hearing Screening, Cernobbio (Lake Como), Italy, 2012.

A. Fernández, B. Remeseiro. Aplicación de apoyo al diagnóstico basado en vídeo para pruebas de audiometría tonal liminar. III Workshop Internacional en IMAGEN MÉDICA y NUEVAS TECNOLOGÍAS (Invited speaker), A Coruña, Spain, June 2013.

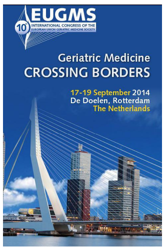
A. Fernández, C. Vazquez, LM Gigirey, M. Ortega, M. G. Penedo. Automatic Eye Gesture Detection and Classification during Hearing Assessment for Patients with Cognitive Decline. 11th Congress of the European Federation of Audiology Societies (EFAS), 59(2), 58, Budapest, Hungary, June 2013.



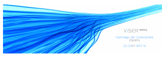
A. Fernández, J. Marey, M. Ortega, M. G. Penedo. Interest operator analysis for automatic assessment of spontaneous gestures in audiometries. 6th International Conference on Agents and Artificial Intelligence, 1, 221-229, Angers, France, March 2014.


AOP 2014
II International Conference on
Applications of Optics and Photonics
26-30 May 2014 | Aveiro, Portugal
CONFERENCE
BOOKLET

A. Fernández, M. G. Penedo, J. L. Doncel, C. Vazquez, LM Gigirey. Detection of spontaneous eye-gestural reactions during an audiometric evaluation. II International Conference on Applications of Optics and Photonics (AOP), 41, May 2014.



C. Vazquez, LM Gigirey, A. Fernández, M. Ortega, M. G. Penedo. Detection of unconscious eye movements to check audiometric thresholds (Pilot study). International Congress of the European Union Geriatric Medicine Society (EUGMS), Rotterdam, The Netherlands, September 2014.



A. Fernández, M. Ortega. Detection of unconscious eye gestural reactions as a response in audiometric evaluation. Vision Sciences and Eye Research Meeting (ViSER), Santiago de Compostela, Spain, November 2014.

cover not available

A. Fernández, J. de Moura, M. Ortega, M. G. Penedo. Detection and characterization of the sclera: evaluation of eye gestural reactions to auditory stimuli. 10th International Conference on Computer Vision Theory and Applications (VISAPP 2015), Berlin, Germany, March 2015. (Pending of pubblication).

cover not available

V. Bolón-Canedo, A. Fernández, A. Alonso-Betanzos, M. Ortega, M. G. Penedo. On the use of machine learning techniques for the analysis of spontaneous reactions in automated hearing assessment. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2015), Brugues, Belgium, April 2015. (Pending of pubblication).

National Conferences

A. Fernández, M. Ortega, C. Mariño, M. G. Penedo, B. Cancela. Skin region analysis for face detection. Congreso Español de Informática (CEDI)- Asociación Española de Reconocimientos de Formas y Análisis de Imágenes (AERFAI), 101-108, Valencia, Spain, September 2010.

BIOINTEGRASAÚDE 2013
Integramos persoas cun obxectivo: a saúde da cidadanía
Santiago de Compostela, 25 de abril de 2013
Aula Magna José Varela de Montes do Pabillón Docente Novoa Santos
Complexo Hospitalario Universitario de Santiago de Compostela
BIOINTEGRASAÚDE 2013
CAMPUS VIDA
inibic
ibi
idis

A. Fernández, M. Ortega, M. G. Penedo, C. Vazquez, LM Gigirey. Sistema automático para la evaluación auditiva de pacientes con deterioro cognitivo. BioIntregaSaúde (BIS), 42, Santiago de Compostela, Spain, April 2013.

Software Registration

Software registration of the product *Herramienta para la medición de tiempos de respuesta en audiometrías.* Authorship shared by Universidade da Coruña, Manuel Francisco González Penedo, Marcos Ortega Hortas, Noelia Barreira Rodríguez and Alba Fernández Arias.

Software registration of the product *Herramienta para la detección de reacciones espontáneas en audiometrías.* Authorship shared by Universidade da Coruña, Manuel Francisco González Penedo, Marcos Ortega Hortas, Noelia Barreira Rodríguez and Alba Fernández Arias.

Mentions

First award to the best doctoral proposal for the article *Sistema de apoyo en la evaluación audiométrica para la detección de respuestas al estímulo.* Escuela de Verano de Inteligencia Artificial (EVIA) of the Asociación Española para la Inteligencia Artificial (AEPIA).

Appendix F

Resumen

El sentido del oído permite la percepción de sonidos; nos informa de los sonidos y ruidos que se producen a nuestro alrededor y de nuestra posición en el espacio. La pérdida de audición es una afección invisible que implica una disminución repentina o gradual de la audición. Aunque la pérdida de audición es un fenómeno global que afecta indistintamente a todos los grupos de edad, también es cierto que existe una mayor incidencia en la población de más edad; concretamente en la pérdida progresiva de audición para las altas frecuencias según aumenta la edad. La pérdida auditiva es la tercera enfermedad crónica más presente entre los adultos de avanzada edad, siendo también una de las menos tratadas.

Por otra parte, el envejecimiento de la población es un fenómeno demográfico que afecta al mundo entero. Este aumento de la esperanza de vida lleva asociado consigo un aumento de los años vividos con incapacidad e invalidez. En relación con la audición, este envejecimiento tiene como consecuencia no solo una mayor prevalencia de problemas auditivos, sino también una mayor severidad en sus efectos. La pérdida auditiva asociada a la edad, también llamada presbiacusia, se caracteriza por una elevación de los umbrales auditivos, la dificultad para entender el habla en entornos ruidosos y reverberantes y las interferencias en la percepción de los cambios rápidos del habla. A mayores, con el aumento de edad también aumenta la posibilidad de sufrir trastornos neurodegenerativos u otras limitaciones comunicativas. Esta problemática implicará una serie de limitaciones a la hora de evaluar la capacidad auditiva que se comentarán más adelante.

Diferentes estudios han demostrado los efectos negativos que la no correción de la pérdida auditiva pueda tener sobre el bienestar físico, psicológico, social y cognitivo de quienes las sufren. El deterioro auditivo puede derivar en comunicaciones

distorsionadas o incompletas. De hecho, aquellos que sufren pérdida auditiva pueden llegar a experimentar una comunicación tan incompleta que afecte de manera negativa a su vida social, abocando en ocasiones en aislamiento, retraimiento y falta de independencia.

Como se comentó anteriormente, con la edad no solo aumentan las posibilidades de sufrir pérdida auditiva sino también el riesgo de aparición de trastornos neurodegenerativos. Una de las manifestaciones más comunes de entre las enfermedades neurodegenerativas es la enfermedad de Alzheimer, que tiende a afectar a personas mayores de 65 años. La prevalencia de trastornos neurodegenerativos, y especialmente de la enfermedad de Alzheimer, es cada vez más significativo en nuestra sociedad actual. Además, investigaciones recientes demuestran que la pérdida de audición es un factor de riesgo potencial para el deterioro cognitivo. Existen evidencias cientf́icas que relacionan la pérdida auditiva con el aumento de la enfermedad de Alzheimer. Los mayores con pérdida auditiva tienen una tasa de deterioro cognitivo que es hasta un 40% más rápida que la tasa de personas con audición normal.

La razón para esta relación entre pérdida auditiva y deterioro cognitivo puede ser debida al aislamiento social que sufren los individuos que tienen la audición afectada, ya que este aislamiento social tiene consecuencias a largo plazo sobre el funcionamiento del cerebro. Además, la pérdida de audición puede forzar al cerebro a dedicar demasiada energía al procesado del sonido, reduciendo el gasto de energía dedicado a la memoria o al pensamiento. La coexistencia de estas patologías supone una importante complicación en la evaluación de la capacidad auditiva. Casi todos los mayores desarrollarán algún grado de deterioro cognitivo con el paso del tiempo. Dado que el envejecimiento está altamente relacionado tanto con la pérdida de audiciń como con el deterioro cognitivo, la coexistencia de estas dos problemáticas es altamente probable.

El empleo de audífonos y programas de rehabilitación auditiva tiene como consecuencia una mejora en el estado social, emocional, psicológico y físico de las personas con problemas auditivos. Los audífonos modernos mejoran la inteligibilidad del habla y, por lo tanto, la comunicación. Todas estas consideraciones ponen de relieve la importancia de la realización de controles regulares de audición, especialmente entre la población mayor o a cualquier edad en caso de que se observen dificultades auditivas de algún tipo.

La audiometría tonal liminar (ATL) ha sido descrita de manera inequívoca como la prueba estándar para la evaluación clínica de la sensibilidad auditiva. Esta prueba determina los tonos más bajos que una persona puede oír a determinadas frecuencias.

Con ella, los audiólogos evaluan la capacidad auditiva y pueden diagnosticar la presencia de problemas auditivos y su severidad. Es una medición conductual de los umbrales de audición, puesto que se basa en la respuesta que el paciente muestre de manera voluntaria a los estímulos sonoros. Por lo tanto, la cooperación del paciente es necesaria para poder llevar la prueba a cabo, lo cual puede implicar ciertas limitaciones operativas que se discutirán pronto.

F.1 Tesis

Durante la evaluación audiométrica se envían al paciente sonidos puros a través de unos auriculares conectados a un dispositivo llamado audiómetro. Manejando este dispositivo el experto irá modulando diferentes frecuencias a intensidades que enviará al paciente. El paciente deberá responder afirmativamente cuando sea capaz de percibirlos (típicamente se le pedirá que levante la mano para mostrar esta respuesta positiva).

La medición de los tiempos de respuesta es una evaluación adicional que los especialistas llevan a cabo durante la realización de la evaluación auditiva. Esta medición es de especial relevancia para lograr la identificación de pacientes con tiempos de respuesta anormalmente lentos, ya que esto puede ser un síntoma que indique algún tipo de problema asociado y que deberá ser estudiado por el especialista correspondiente. El problema aquí es que el experto necesita haber tratado y estudiado a un elevando número de pacientes con el fin de ser capaz de discriminar cual es el tiempo de respuesta medio de los pacientes “normales”. Incluso a pesar de las habilidades del especialista ésta no deja de ser una tarea cargada de subjetividad, lo que implica que sea una evaluación proclive a errores e imprecisiones. Una alternativa más precisa consistiría en grabar las pruebas en vídeo para una posterior medición manual de cada unos de los tiempos de respuesta, pero esta solución consumiría mucho tiempo por parte de la persona encargada de ello. Es por todo ello que el desarrollo de una solución automática para la medición de estos tiempos de respuesta sería muy útil para los expertos puesto que aceleraría la tarea y proporcionaría mediciones precisas y reproducibles.

En la Parte I se presenta un método que mide de forma automática los tiempos de respuesta a partir de secuencias de vídeo grabadas durante la realización de las pruebas audiométricas. Se utilizan técnicas basadas en el uso del color para la detección de la mano cuando ésta es levantada por el paciente. Puesto que la mano es una zona de piel que durante la realización de la prueba permanecerá expuesta, si se consigue detectar adecuadamente su presencia y movimiento se podrán detectar las respuestas positivas del paciente. Puesto que los tiempos de respuesta se deben medir como el tiempo transcurrido desde el momento en que se inicia el envío del estímulo hasta el momento en que se inicia la respuesta del paciente, es necesario identificar en el audiómetro los momentos en los que se envía estímulo auditivo con el fin de combinar ambas informaciones. Debido a que los expertos están trabajando con dispositivos analógicos, la detección del envío de estímulos se realizará utilizando propiedades de color buscando el indicador luminoso que se enciende cuando se está enviando estímulo. Los resultados obtenidos por la metodología propuesta son altamente positivos, alcanzando valores superiores al 99% de acierto en la detección de las reacciones del paciente. Además, el método proporciona una medición precisa y objetiva de los tiempos de respuesta, medición que es obtenida a partir de la combinación de la detección de la respuesta del paciente y de la detección del envío de estímulos por parte del experto. A través de los resultados experimentales se ha demostrado la capacidad del método para permitir la discriminación entre pacientes con tiempos de respuesta normales y pacientes con tiempos de respuesta anormalmente lentos. De este modo, la propuesta aquí presentada permite la identificación de los pacientes con tiempos de respuesta significativamente más lentos de lo normal (lo cual era una de las metas de este trabajo) y se facilita la medición precisa de cada uno de estos tiempos facilitando a los expertos poder llevar a cabo diferentes estudios clínicos.

Por otra parte, en el caso de pacientes con deterioro cognitivo u otros trastornos severos que afecten a la comunicación, el protocolo estándar en el cual el experto le explica al paciente la prueba y le pide que levante la mano cuando perciba en estímulo auditivo es prácticamente inaplicable. En estos casos, la existencia de una interacción activa entre paciente y experto es altamente improbable, ya que este tipo de pacientes tiene importantes limitaciones a la hora de mantener una interacción estable, limitaciones que se ven agravadas según el deterioro cognitivo empeora. Si bien es cierto que la evaluación auditiva de estos pacientes se vuelve mucho más compleja, todavía es posible llevarla a cabo si el audiólogo centra su atención en la detección de sutiles reacciones espontáneas faciales (reacciones que se centrarán principalmente en la región de los ojos). La subjetividad asociada a esta intepretación de gestos y la sutileza que pueden tener estas reacciones faciales hace de esta tarea un problema cargado de imprecisión, proclive a errores y difícil de reproducir. Todas estas razones ponen en relieve las mejoras que una solución automatizada podría aportar al dominio, ayudando a los audiólogos en la detección e interpretación de estas reacciones. Esta problemática se estudiará a lo largo de la

Parte II de esta tesis.

Una primera aproximación para la detección de estas reacciones gestuales se presenta a lo largo del Capítulo 4. La detección de estas reacciones espontáneas el llevada a cabo utilizando como base el flujo óptico sobre la región de interés, que en este caso se corresponde con la región de los ojos. El flujo óptico analiza el movimiento existente entre una serie de puntos detectados como significativos en los dos momentos a comparar. Se calcula la correspondencia entre esos puntos, y el movimiento detectado (en caso de que éste exista) se caracterizará en función de su orientación, magnitud y dispersión. Una vez caracterizado, se clasificará en una de las categorías de movimiento determinadas como revelantes para este dominio por los expertos. Los resultados obtenidos muestran la capacidad del método propuesto para la correcta identificación y clasificación de dichas reacciones, abriendo el camino para el desarrollo de una herramienta automática adaptada a este dominio.

En el Capítulo 5 la detección de las reacciones gestuales se realiza a partir de información sobre la distribución de color en la esclerótica (la parte blanca del ojo). La distribución de color de la esclerótica permite obtener una aproximación de la dirección de la mirada, puesto que si la distribución de esta región está caracterizada por una zona clara, seguida de una zona oscura (que se corresponde con el iris), para luego terminar de nuevo con una zona clara, se podrá deducir que la mirada está dirigida a un punto central. En cambio, si se tiene primero una zona oscura y luego una gran zona clara, se podrá deducir que en este caso la mirada está dirigida hacia un lado. Estas distribuciones serán clasificadas en función de su forma en las distintas categorías consideradas. Los resultados obtenidos son positivos y apuntan hacia la suficiencia de este método para la detección de reacciones faciales.

Finalmente, en el Capítulo 6 se propone la combinación de las dos técnicas anteriores. En esta nueva propuesta, la detección de las reacciones faciales se lleva a cabo a partir de la combinación de la información aportada por el flujo óptico y por el análisis de la distrubición de color en la esclerótica. En el caso del flujo óptico, como en la propuestas inicial ya se trabajaba con un vector descriptor, el proceso se mantiente igual. En cambio, en el caso de la distribución de color de la esclerótica es necesario caracterizar la información en forma de vector de características. Además, esta aproximación será considerada tanto para el momento actual de la distribución, como en comparación con el momento con el cual estemos tratando de observar el cambio. Los resultados obtenidos muestran que la combinación de ambas técnicas consigue mejores resultados que si se considera cada una de ellas por separado. La mayor tasa de acierto obtenida es del 97,46%, lo cual es un valor muy positivo teniendo en cuenta las dificultades del dominio. A modo de mejora, en el Capítulo 6.5 se propone la aplicación técnicas de aprendizaje máquina para la mejora de la tasa de clasificación de las dos categorías más relevantes: movimiento de mirada hacia la izquierda y movimiento de mirada hacia la derecha. Para ello, se aplicaron técnicas de oversampling, en particular, el método SMOTE. Incrementando las dos categorías que queremos potenciar hasta un oversampling del 300%, se obtuvieron tasas de acierto por encima de 0.986 y 0.966 para ambas clases, con un acierto global de clasificación superior al 97%.

F.2 Conclusiones

A lo largo de esta tesis se han propuesto diferentes técnicas para facilitar la evaluación automática de la capacidad auditiva. Las dos principales metas eran la medición de los tiempos de respuesta y la detección de reacciones faciales como respuesta a los estímulos auditivos para el caso de pacientes con dificultades comunicativas graves. Estas soluciones automatizadas no pretenden sustituir el juicio del experto, pero podrán ser de gran ayuda en la realización de la rutina clínica y en investigación.

Una de las principales premisas bajo las que se tuvo que desarrollar este trabajo fue la de no alterar, en la medida de lo posible, el protocolo estándar que siguen los expertos para la evaluación de la audición. El silencio y la ausencia de distracciones son vitales para el correcto desarrollo de esta prueba. Es importante también crear un ambiente tranquilo, en el que el paciente se sienta cómodo y actúe de forma natural y espontánea. Es importante intentar no condicionar la respuesta natural del paciente, por lo que no es conveniente evitar que se sienta observado. Las metodologías aquí presentadas no implican ningún cambio en el protocolo estándar de realización de las pruebas audiométricas, el único requisito es la colocación de un video cámara para la grabación de las pruebas. Esta video cámara se colococará detrás del audiólogo, que estará sentado frente al paciente. Es una localización discreta que perturbará mínimamente el comportamiento del paciente.

Primeramente se propone un método capaz de medir de forma automática y precisa los tiempos de respuesta del paciente. Para ello es necesario detectar el momento en que se envía el estímulo y el momento en que el paciente levanta la mano. La precisión obtenida es óptima y la precisión de las mediciones obtenidas permitirá el desarrollo de estudios clínicos.

El siguiente paso consistía en resolver la detección de las reacciones gestuales espontáneas mostradas por los pacientes con dificultades comunicativas severas. Se propusieron dos formas alternativas de abordarlo: el análisis de movimiento a partir del flujo óptico, y la localización de la dirección de la mirada a través del análisis de la distribución de color de la esclerótica. Aunque ambas aproximaciones ofrecían resultados aceptables, se optó por combinarlas con el fin de dotar al método de mayor robustez. Los resultados obtenidos de la combinación de ambas técnicas avalan la decisión.

Esta aproximación final permite presentar una metodología para la detección y clasificación de movimientos gestuales como respuesta a los estímulos auditivos durante la realización de evaluaciones auditivas en pacientes con deterioro cognitivo u otras dificultades comunicativas severas. Cuando no existe interacción por parte del paciente el experto focalizaba su atención en busca de este tipo de reacciones, pero la disponibilidad de una herramienta automática que se encargue de esta labor facilitará su trabajo, evitará la subjetividad y lo hará menos proclive a errores.

En términos clínicos, el anális manual llevado a cabo por los audiólogos puede ser automatizado con el principal beneficio de no verse afectado por factores subjetivos. Además del hecho de que la propuesta automática produce resultados imparciales, también ahorra tiempo para los expertos y proporciona una identificación detallada de las reacciones gestuales. En este sentido, el audiólogo logrará valoraciones más objetivas, lo cual será de gran ayuda para conseguir una evaluación adecuada de la capacidad auditiva. El correcto diagnóstico correcto de la pérdida de audición permitirá la prescripción de audífonos apropiados, mejorando así la calidad de vida de estos pacientes.

Por otra parte, el hecho de que las secuencias de vídeo puedan ser visualizadas después de haber sido analizadas por la metodología propuesta facilitará el entrenamiento de nuevo expertos en la detección e interpretación de estas reacciones. Como este tipo de reacciones son bastante inconsistentes y sutiles, su adecuada interpretación requiere amplia experiencia por parte del audiólogo. Si se le proporciona a los audiólogos inexpertos en esta tarea una herramienta en la que puedan observar evaluaciones previas donde las reacciones gestuales estén adecuadamente etiquetadas, se estará facilitando su entrenamiento, pues podrán ir aprendiendo de casos previos que se encuentren archivados.

Aunque en esta tesis nos hemos centrado en el dominio de la audiología, la metodologías propuestas para la detección de movimientos de la mirada podrían ser fácilmente adaptadas a otros dominios donde estos movimientos proporcionasen algún tipo de información relevante.

Bibliography

Acton, Q. (2013). *Dementia: New insights for the healthcare professional: 2013 edition*. ScholarlyEditions.

Agrawal, Y., Platz, E., & Niparko, J. (2008, Jul). Prevalence of hearing loss and differences by demographic characteristics among us adults. *Archives of Internal Medicine*, *168*, 1522-30.

Akakin, H. C., & Sankur, B. (2011, June). Robust classification of face and head gestures in video. *Image and Vision Computing*, *29*(7), 470-483.

Asaari, M., & Suandi, S. (2010, December). Hand gesture tracking system using adaptive kalman filter. In *10th international conference on intelligent systems design and applications (ISDA)* (p. 166 -171).

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004, June). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorarations*, *6*(1), 20–29.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008, June). Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, *110*(3), 346–359.

Bouguet, J.-Y. (2000). Pyramidal implementation of the Lucas-Kanade feature tracker: Description of the algorithm. *Intel Corporation, Microprocessor Research Labs*.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In T. Theeramunkong, B. Kijsirikul, N. Cercone, & T.-B. Ho (Eds.), *Advances in Knowledge Discovery and Data Mining* (Vol. 5476, p. 475-482). Springer Berlin Heidelberg.

Caglar, M., & Lobo, N. (2006, June). Open hand detection in a cluttered single image using finger primitives. In *Conference on Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06* (p. 148).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002, June). SMOTE: Synthetic Minority Over-sampling TEchnique. *Journal of Artificial Intelligence Research.*, *16*(1), 321–357.

Chew, S., Rana, R., Lucey, P., Lucey, S., & Sridharan, S. (2012). Sparse temporal representations for facial expression recognition. In Y.-S. Ho (Ed.), *Advances in Image and Video Technology* (Vol. 7088, p. 311-322). Springer Berlin Heidelberg.

Ciorba, A., Bianchini, C., Pelucchi, S., & Pastore, A. (2012). The impact of hearing loss on the quality of life of elderly adults. *Clinical Interventions in Aging*, *7*, 159–163.

Collins, J. (1997). *Prevalence of selected chronic conditions: United States 1990-1992* (Vol. 10) (No. 194). Hyattsville, MD: National Center for Health Statistics, Public Health Service, National Center for Health Stadistics.

Davis, A. (1989, December). The prevalence of hearing impairment and reported hearing disability among adults in Great Britain. *International Journal of Epidemiology*, *18*, 911-17.

Dawod, A., Abdullah, J., & Alam, M. (2010, August). A new method for hand segmentation using free-form skin color model. In *Advanced computer theory and engineering (ICACTE), 2010 3rd international conference on* (Vol. 2, pp. 562–566).

del Río, S., López, V., Benítez, J. M., & Herrera, F. (2014). On the use of MapReduce for imbalanced big data using Random Forest. *Information Sciences*, *285*, 112 - 137.

Domingos, P. (1999). MetaCost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 155–164). New York, NY, USA: ACM.

Fernández, A., Ortega, M., Cancela, B., Penedo, M., Vázquez, C., & Gigirey, L. (2012). Automatic processing of audiometry sequences for objective screening of hearing loss. *Expert Systems with Applications*, *39*(16), 12683 - 12696.

Fernández, A., Ortega, M., Penedo, M. G., Vázquez, C., & Gigirey, L. (2014). A methodology for the analysis of spontaneous reactions in automated hearing assessment. *Biomedical and Health Informatics, IEEE Journal of*, *PP*(99), 1-1.

FG-NET Aging Database. (n.d.). http://fipa.cs.kit.edu/index.php.

García, S., & Herrera, F. (2009, September). Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary Computation*, *17*(3), 275–306.

Gatehouse, S., Naylor, G., & Elberling, C. (2003). Benefits from hearing aids in relation to the interaction between the user and the environment. *International journal of audiology*, *42*(S1), 77–85.

Geetha, A., Ramalingam, V., Palanivel, S., & Palaniappan, B. (2009). Facial expression recognition - A real time approach. *Expert Systems with Applications*, *36*(1), 303 - 308.

Graf, H., Cosatto, E., Gibbon, D., Kocheisen, M., & Petajan, E. (1996). Multi-modal system for locating heads and faces. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on* (p. 88-93).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009, November). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, *11*(1), 10–18.

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-SMOTE: A New Over-sampling Method in Imbalanced Data Sets Learning. In *Proceedings of the 2005 International Conference on Advances in Intelligent Computing - Volume Part I* (pp. 878–887). Berlin, Heidelberg: Springer-Verlag.

Happy, S., George, A., & Routray, A. (2012). A real time facial expression classification system using Local Binary Patterns. In *Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on* (p. 1-5).

Harris, C., & Stephens, M. (1988). A Combined Corner and Edge Detector. In *Proceedings of the 4th Alvey Vision Conference* (pp. 147–151).

Hearing, A. (2009). *Australian hearing annual report.* Retrieved from `http://www.hearing.com.au/annual-reports`

IMSERSO. (2008). Las Personas Mayores en España. In *Instituto de Mayores y Servicios Sociales.*

IMSERSO. (2010, October). Libro Blanco del Envejecimiento Activo..

Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI)* (pp. 111–117).

Jorge, F., Carvalho, S., Manuel, J., & Tavares, R. (2007). Eye detection using a deformable template in static images. In *in vipimage - i eccomas thematic conference on computational vision and medical image processing* (pp. 209–215).

Kakumanu, P., Makrogiannis, S., & Bourbakis, N. (2007). A survey of skin-color modeling and detection methods. *Pattern Recognition*, *40*, 1106-1122.

Kalache, A., Barreto, S., & Keller, I. (2005). Global ageing: The demographic revolution in all cultures and societies. *The Cambridge Handbook of Age and Ageing*, 30-46.

Kawato, S., & Tetsutani, N. (2004). Detection and tracking of eyes for gaze-camera control. *Image and Vision Computing*, *22*(12), 1031 - 1038.

Kelly, D., Walsh, F., Norman, G., & Cunningham, A. (1999, May). The effects of midazolam on pure tone audiometry, speech audiometry, and audiological reaction times in human volunteers. *Anesthesia & Analgesia*, *88*(5), 1064-1068.

Kochkin, S., & Rogin, C. (2000). Quantifying the Obvious: The Impact of Hearing Aids on Quality of Life. *The Hearing Review.*, *7*(1), 8-34.

Kothari, R., & Mitchell, J. (1996, Sept). Detection of eye locations in unconstrained visual images. In *Image Processing, 1996. Proceedings., International Conference on* (Vol. 3, p. 519-522).

Kubat, M., & Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 179–186). Morgan Kaufmann.

Kumano, S., Otsuka, K., Yamato, J., Maeda, E., & Sato, Y. (2009, June). Pose-Invariant Facial Expression Recognition Using Variable-Intensity Templates. *Intenational Journal of Computer Vision*, *83*(2), 178-194.

Lam, K.-M., & Yan, H. (1996). Locating and extracting the eye in human face images. *Pattern Recognition*, *29*(5), 771 - 779.

Laurikkala, J. (2001). Improving Identification of Difficult Small Classes by Balancing Class Distribution. In *Proceedings of the 8th Conference on AI in Medicine in Europe: Artificial Intelligence Medicine* (pp. 63–66). London, UK, UK: Springer-Verlag.

Li, D., Winfield, D., & Parkhurst, D. (2005, June). Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on* (p. 79-79).

Lin, F. R. (2011). Hearing Loss and Cognition Among Older Adults in the United States. In *The Journals of Gerontology: Series A.*

Lin, F. R., Yaffe, K., Xia, J., Xue, Q., Harris, T., Purchase-Helzner, E., ... Simonsick, E. (2013). Hearing loss and cognitive decline in older adults. *JAMA Internal Medicine*, *173*(4), 293-299.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141.

Lowe, D. G. (2004, November). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2* (pp. 674–679). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Mulrow, C., Aguilar, C., Endicott, J., Tuley, M., Velez, R., Charlip, W., & et al. (1990, Aug). Quality-of-life changes and hearing impairment. a randomized trial. *Annals of Internal Medicine*, *113*, 188-94.

Murthy, K. N. B., & Natarajan, S. (2011). Feed Forward Neural Network Based Eye Localization and Recognition Using Hough Transform. *IJACSA - International Journal of Advanced Computer Science and Applications*, *2*(3), 104–109.

of Deafness, N. I., & Disorders, O. C. (2009, March). *Quick Statistics.* Retrieved from `http://www.nided.uih.gov/health/statistics/hearing.asp`

Peng, S.-Y., Wattanachote, K., Lin, H.-J., & Li, K.-C. (2011, July). A Real-Time Hand Gesture Recognition System for Daily Information Retrieval from Internet. In *Ubi-Media Computing (U-Media), 2011 4th International Conference on* (p. 146-151).

Rosten, E., & Drummond, T. (2005). Fusing points and lines for high performance tracking. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* (Vol. 2, p. 1508-1515).

Rosten, E., & Drummond, T. (2006). Machine Learning for High-speed Corner Detection. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part I* (pp. 430–443). Berlin, Heidelberg: Springer-Verlag.

Sandeep, K., & Rajagopalan, A. N. (2002). Human Face Detection in Cluttered Color Images Using Skin Color and Edge Information. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'02)* (p. 230-235).

Santos, G., & Proenca, H. (2011, Oct). A robust eye-corner detection method for real-world data. In *Biometrics (IJCB), 2011 International Joint Conference on* (p. 1-7).

Shi, J., & Tomasi, C. (1994). Good Features to Track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on* (p. 593-600).

Sobel, I., & Feldman, G. (1968). *A 3x3 Isotropic Gradient Operator for Image Processing.* (Never published but presented at a talk at the Stanford Artificial Project)

Spruyt, V., Ledda, A., & Geerts, S. (2010, Sept.). Real-time multi-colourspace hand segmentation. In *Image Processing (ICIP), 2010 17th IEEE International Conference on* (p. 3117 -3120).

Terrillon, J.-C., & Akamatsu, S. (2000). Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images. In *Proceedings of the 12th Conference on Vision Interface (VI '99)* (p. 180-187).

Timm, F., & Barth, E. (2011). Accurate Eye Centre Localisation by Means of Gradients. In L. Mestetskiy & J. Braz (Eds.), *VISAPP* (p. 125-130). SciTePress.

Viola, P., & Jones, M. (2001). Robust Real-time Object Detection. In *International Journal of Computer Vision.*

Yang, J., & Waibel, A. (1996). A real-time face tracker. In *WACV '96: Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96)* (p. 142).

Yoo, T.-W., & Oh, I.-S. (1999). A fast algorithm for tracking human faces based on chromatic histograms. *Pattern Recognition Letters*, *20*, 967-978.

Zhu, Z., & Ji, Q. (2005, April). Robust Real-time Eye Detection and Tracking Under Variable Lighting Conditions and Various Face Orientations. *Computer Vision Image Understanding*, *98*(1), 124–154.