Automatic evaluation of eye gestural reactions to sound in video sequences

Alba Fernández [a,b], Marcos Ortega [a,b,*], Joaquim de Moura [a,b], Jorge Novo [a,b], Manuel G. Penedo [a,b],

[a] *Department of Computer Science, University of A Coruña, 15071 A Coruña, Spain*

[b] *CITIC - Research Center of Information and Communication Technologies, University of A Coruña, 15071 A Coruña, Spain*

Abstract

Hearing loss is a common disorder that often intensifies with age. In some cases, especially in the elderly population, the hearing loss may decrease in physical, mental and social well-being capacities. In particular, patients with signs of cognitive impairment typically present specific clinical-pathological conditions, which complicates the analysis and diagnosis of the type and severity of hearing loss by clinical specialists. In these patients, unconscious changes in gaze direction may indicate a certain perception of sound through their auditory system. In this context, this work presents a new system that supports clinical experts in the identification and classification of eye gestures that are associated with reactions to auditory stimuli by patients with different levels of cognitive impairment. The proposed system was validated using the public Video Audiometry Sequence Test (VAST) dataset, providing a global accuracy of 97.12% for the classification of eye gestures and a 100% for gestural reactions to auditory stimuli. The proposed system offers a complete analysis of audiometric video sequences, being applicable in daily clinical practice, improving the well-being and quality of life of the patients.

Key words: Sound responses; gestural reactions; hearing screening; movement detection; screening methods; video sequences.

1 Introduction

Hearing loss, also known as hearing impairment, is a pathological condition that is commonly characterized by the gradual loss of the ability to perceive different levels of sound. Hearing pathology is the third most common chronic physical condition, representing a serious health problem that typically affects people of all ages [1]. It is estimated that about 700 million people worldwide present some degree of hearing loss. This pathology can also be caused by physical changes in the auditory nerve, which provoke a significant reduction in the ability of the brain to perceive and process sounds, particularly those at high frequencies. Over recent years, representative relevant studies correlated the hearing loss with the aging process, showing the specially high prevalence of this pathology in the elderly population [2] [3]. Therefore, as population aging increrases rapidly in the most economically advanced countries, the elderly population with hearing loss is also significantly increased [3] [4].

According to the World Health Organization (WHO), by the year of 2050, it is estimated that over 900 million people will suffer from hearing loss or deafness [5]. Currently, approximately one-third of the people over 65 years

* Corresponding author. Department of Computer Science, University of A Coruña, 15071 A Coruña, Spain.

Email addresses: alba.fernandez@udc.es (Alba Fernández), mortega@udc.es (Marcos Ortega), joaquim.demoura@udc.es (Joaquim de Moura), jnovo@udc.es (Jorge Novo), mgpenedo@udc.es (Manuel G. Penedo).

old are affected by this pathology [6], the impact of which is significantly intensified at advanced ages [7]. In fact, the number of older people (60 years or more) is expected to more than double by 2050 and more than tripled by 2100 [8], which clearly drives the increase of the indicated rates.

Recent research indicates that hearing loss is associated with a reduction in physical, mental and social well-being capacities [3] [9]. Moreover, a person with hearing loss can be presented with significant difficulties when participating in social activities due to a restricted ability to interact and communicate with other people. This state of social isolation can lead to depression, physical and mental health problems and even mortality. A convenient identification of these patients and the corresponding adjusted diagnosis is, therefore, crucial. To do so, computational systems for the evaluation of eye gestural reactions to sound represent a valuable support for specialists, being applicable in daily clinical practice, helping the audiologist to produce a more precise analysis of this pathology, and improve the well-being and life quality of patients.

It is well known that age-related neurodegenerative diseases, such as Alzheimer's disease and Parkinson's disease, affect the aging of individuals. Alzheimer's disease is the most common progressive neurodegenerative disorder [10] and is mainly characterized by a progressive decrease of cognitive functions, which leads to a complete dependence on others for the basic activities of the daily life routines. It is estimated that there are approximately 44 million people worldwide affected by Alzheimer's disease or some degree of dementia.

Different studies associate the hearing loss with dementia and cognitive problems [11,12]. In particular, older people with hearing loss are more likely to develop Alzheimer's disease than the rest of the population [11]. The magnitude of the association between the hearing impairment and the cognitive decline is clinically significant. Older adults with hearing impairment have cognitive impairment rates of 30% to 40% compared to individuals with normal hearing, suggesting an early indicator of possible dementia [12]. As previously indicated, hearing impairment can result in important alterations in human brain function. Besides, a degraded hearing may force the brain to devote too much energy to process sounds, becoming an extremely exhausting activity that reduces the energy that remains available for other activities such as thinking or remembering [13]. Patients with dementia, especially the case of Alzheimer's disease, present specific clinical-pathological conditions, which makes the analysis and diagnosis of hearing impairments significatively difficult. Under these considerations, we note the importance of regular hearing checks, even of screening programs over the whole population, but specially among the elderly individuals.

Pure Tone Audiometry (PTA) is the most common hearing test procedure used to obtain the hearing threshold levels of an individual. This clinical test determines the softest sound that the subject can perceive at the selected frequencies. It is used to measure hearing ability and also to determine if hearing aids or surgery can restore or at least improve an individual's hearing. In particular, PTA is based on the patient's responses to the pure tone stimuli. Therefore, PTA is a subjective test as the specialist expects some type of consistent response from the patient when the different test sounds are heard. This interaction is very limited in individuals with cognitive impairment or other serious communication problems, which increases the difficulty of the hearing test process.

In detail, the process is as follows: first, during the audiometric evaluation, the stimuli of pure tones at different frequencies are presented to the individual through conventional headphones [14]. The individual must then interact with the clinical specialist when an acoustic stimulus is perceived. For individuals without cognitive decline, this interaction can be by raising a hand or with an oral response. When the patient shows some degree of cognitive decline, this common interaction could not be possible.

Normally in the PTA hearing test, patients with cognitive impairment or other severe communication disorders do not show a voluntary response to the auditory stimuli. Instead, these patients usually react unconsciously with mild or subtle facial reactions. These facial reactions are perceptible principally in the region of the eye. In particular, unconsciously changes in the gaze direction, eye opening or closing, may indicate a certain perception of sound through their auditory system. Clinical experts must, therefore, focus their attention on these unconscious gestures in the region of the eye, allowing the target measurement and evaluation of the patient's hearing thresholds. However, the analysis and interpretation of these spontaneous reactions requires a high level of expertise and experience by the audiologist. Further, the intrinsic nature of these reactions to auditory stimuli depends completely on the patient. Hence, each patient can show different gestures in reaction to different hearing evaluations, increasing the difficulty of the process. In summary, this subjectivity makes this evaluation an imprecise problem, susceptible to errors and difficult to reproduce. In fact, the main consequence of this important limitation lies in the reproducibility of the results by different clinical specialists in the evaluation of patients with cognitive impairment or serious communication disorders.

Over recent years, proposals were presented for the evaluation and analysis of eye movement in different clinical scenarios. In De Santi *et al.* [15], the authors proposed an eye-tracking system for the detection of the pursuit ocular movement dysfunctions in patients with multiple sclerosis disease. In the work of Pereira *et al.* [16], the authors presented a study of different eye movement patterns as a measurement of the degree of Alzheimer's disease. Eye movement analysis is also valuable for visual tracking systems. In Raney *et al.* [17], the authors proposed a complete methodology to perform the evaluation of the cognitive processes that are involved in text understanding through eye movements. In the work of Marandi *et al.* [18], the authors presented a study of the correlation between ocular movement patterns and electrooculography (EOG) signals, using pattern recognition techniques. The current literature presents different proposals for the analysis of eye movement in different clinical scenarios. However, these proposals present a common limitation due to the lack of an in-depth analysis that covers all possible reactions of a particular individual. Further, it is very complicated to adapt existing solutions to identify and classify spontaneous reactions due to the intrinsic nature of these unconscious gestures.

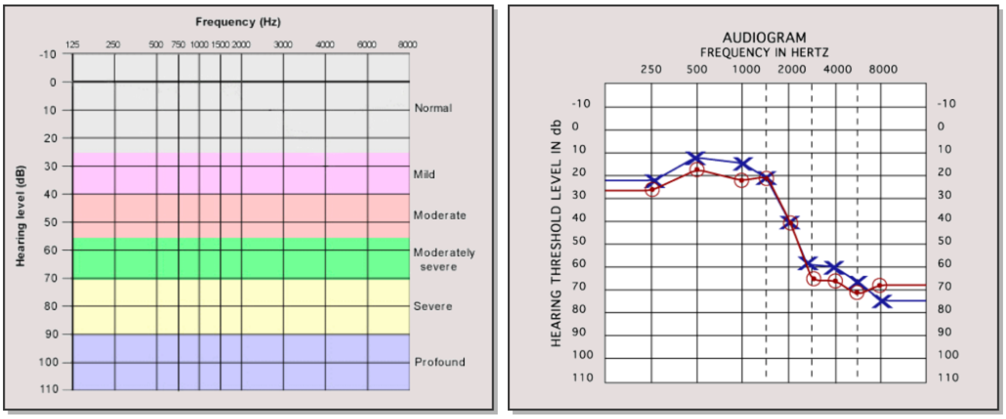
In this work, we present an automatic system for the evaluation of eye gestural reactions to sound in video sequences. In particular, this proposal combines different computer vision and image processing techniques for the identification and classification of eye gestures using audiometric video sequences. In this way, we use the information on the optical flow to identify the movement of the region of the eye and the distribution of the color of the sclera (the white region of the eyes) for the characterization of the direction of the gaze of the patient. The application of the optical flow was presented in the work of Fernández *et al.* [19] whereas the use of the color of the sclera in the work of Fernández *et al.* [20]. Taking these into account, we propose a complete system that merges these two strategies to achieve a more precise and robust way for the evaluation of the eye's gestural reactions. The main challenge of this proposal is the development of an automatic system that supports the audiologists in the identification of eye gestures associated with reactions to auditory stimuli from patients with different levels of cognitive impairment. These unconscious reactions totally depend on the clinical state of the patient, being more subtle or marked according to each individual. Moreover, different variations in the direction of the gaze or the opening/closing of the eyes are directly related to the reactions to auditory stimuli. In this context, we proposed the system to assist clinical specialists in the identification and characterization of the eye's gestural reactions in this particular group of patients. The validation of the system was performed with an extensive set of designed experiments, using the public Video Audiometry Sequence Test (VAST) [21] dataset. The VAST dataset is composed of different individuals of different ages, genders and cognitive abilities, representing a significative variability of the situations during the audiometric test. In Section 3.1, we extensively explain this used video dataset.

This work is divided into 5 main Sections. In Section 2, we explain the main details of an audiometric procedure. Section 3, Methodology, presents the methodology with the main characteristics of all the involved stages of the system. Section 4 details all the experiments that were done to validate the method as well as the discussion about the obtained results. Finally, in Section 5, Conclusions, we state the final thoughts and future lines of work.

2 Audiometric procedure

As mentioned above, regular hearing tests are important for individuals over 65 years of age. This clinical test can help audiologists to identify hearing disabilities at a very early developmental stage, which is very beneficial for the well-being of the individual. Hearing aids are designed to improve the hearing function by making the sound audible to a person with hearing loss. In particular, older adults with hearing impairment have difficulty in communicating and participating more fully and more effectively in their daily activities. The PTA clinical test is considered the gold standard for identifying a person's hearing threshold levels. It determines the softest sounds that a person can hear in the selected frequencies, progressively from low to high. The PTA hearing test constitutes a subjective measurement of the hearing threshold level, since it requires the participant to respond reliably to the auditory stimuli, when a test sound is produced. Normally, the results obtained from this screening test are graphed as an audiogram. An audiogram is a graphical representation that indicates how well a person can hear different types of sounds. These sounds are emitted at different frequencies and intensities. Figure 1(a) shows the different degrees of hearing loss: normal, mild, moderate, moderately severe, severe and profound. In Figure 1(b), we can see a representative example of an audiogram, where the frequency is displayed logarithmically on the x-axis with a linear scale dBHl on the y-axis. Each red circle (for the right ear) and the blue cross (for the left ear) represent the individual sound frequencies that have been presented.

In the PTA test, the clinical expert first describes for the individual the basic principles and the main characteristics of this hearing screening. In this



(a)

(b)

Fig. 1. A representative example of an audiogram. (a) The degrees of hearing loss are described as normal, mild, moderate, moderately severe, severe and profound. (b) The red line represents the right ear whereas the blue line corresponds the left ear.

way, the individual will interact and respond correctly to the clinical expert when receiving an auditory stimulus. It is very important that the individual comprehends the specific instructions to perform this hearing test. This necessity of understanding between the individual and the clinical specialist is what complicates the evaluation of patients with different levels of cognitive impairment. In particular, in the case of individuals without cognitive impairment, the clinical expert indicates that they will receive different types of auditory stimuli through hearing aids. Consequently, the patients must interact with the clinical expert by raising the corresponding hand when they perceive each auditory stimulus. In the particular case of individuals with some degree of cognitive impairment, the clinical expert focuses on the spontaneous movements (especially in the ocular region) to acquire some sign of response to the auditory stimulus.

The evaluation of eye gestural reactions is performed without involving significant changes in the behavior of the individual or the clinical specialist. These characteristics, in particular, constitute an important advantage for the implementation of our approach. For this, the method analyzes the images that are acquired using a conventional digital video camera during the performance of the audiometric evaluation. In the next Section, the proposed methodology is explained in more detail.

3 Methodology

The proposed methodology is composed of a set of progressive steps for the automatic evaluation of eye gestural reactions to sound using video sequences. A schematic representation of the main steps of this methodology can be seen in Figure 2. Firstly, as input, the presented methodology receives a video sequence. Then, the system localizes the region of the patient's eyes. Finally, using the identified region as reference, the method combines the information of the analysis by optical flow and the color distribution of the sclera to identify and classify the movement of the eyes. Each of these steps will be discussed below.

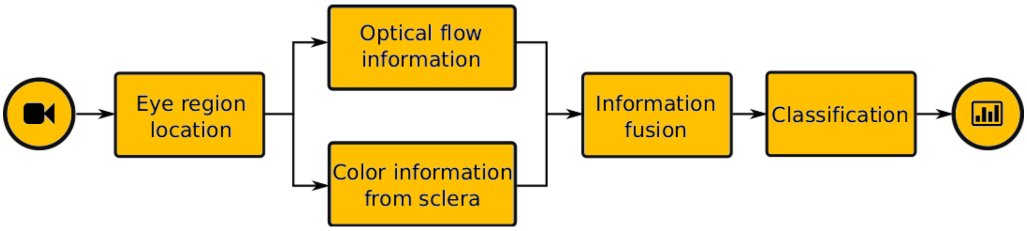


Fig. 2. Schematic representation of the main steps of the methodology.

3.1 Video signal acquisition - the VAST dataset

The public Video Audiometry Sequence Test (VAST) [21] dataset was used as input and validation of the presented system. The VAST database currently includes 11 audiometric video sequences of different individuals with different ages, genders and cognitive abilities. These video sequences were taken using a digital video camera, with a high resolution of 1,080 × 1,920 pixels and 25 frames per second (fps). In particular, each of the video sequences takes between 4 and 8 minutes. This study was designed as a statement of the ethical principles for medical research and following the tenets of the Declaration of Helsinki.

Each video sequence was recorded during an audiometric test. The scene of acquisition of the video sequences is simple. The digital video camera is located behind the medical specialist (the medical specialist is sitting in front of the patient) and the recorded scene shows the patient's face and the audiometer. In Figure 3, we can see 3 representative examples of video scenes that were recorded during the audiometric test.

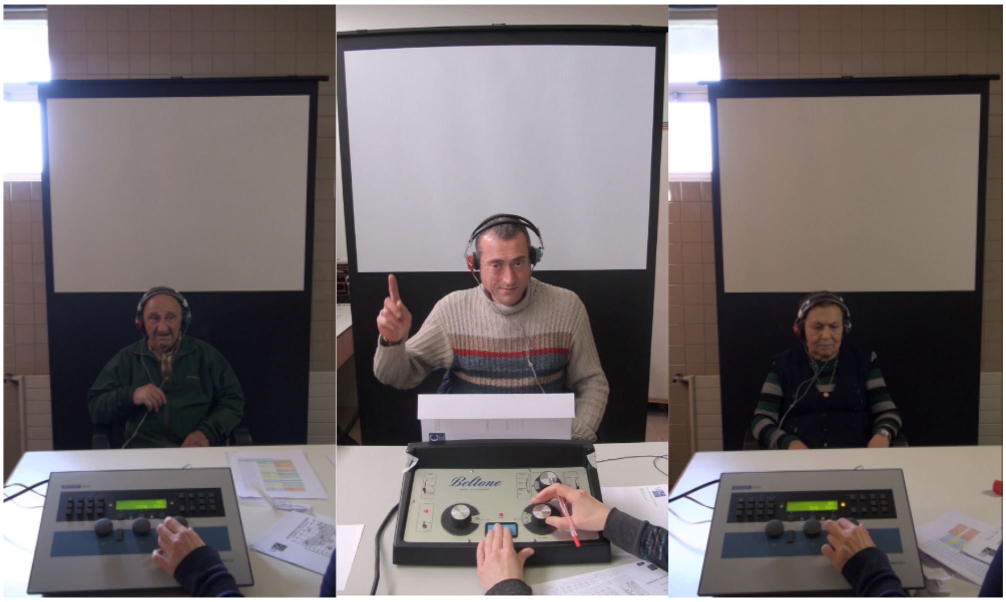


Fig. 3. Three representative examples of video scenes that were recorded during the audiometric test.

3.2 Eye region location

Using the recorded video, the system detects the region of interest (ROI), represented by the eye regions of the patients. To do so, we analyze the video sequence frame-by-frame. To reduce the search space and, therefore, the computational requirements, we firstly detect and extract the identified face region. Subsequently, the eye region is detected within the face region.

In particular, our domain is extremely stable with respect to the position of the elements of interest within the scene. The specialist is always sitting in front of the individual and the digital video camera is situated behind the specialist to ensure that the face of the individual is always registered in the frontal position. This specific and particular configuration of the audiometric test scene allows us to accurately apply the Viola-Jones detector [22], using an optimized cascade for the detection of faces in frontal position. The Viola-Jones detector [22] is a general object detection framework that is widely used in the literature to solve similar problems, such as the face detection in video sequences. This object detector algorithm uses an optimized Haar feature-based cascade classifiers and, therefore, is extremely fast and reliable. In Figure 4, we can see two representative examples of face detection using the Viola-Jones algorithm.

After the face region localization, we further need to detect the eye region. For that, the Viola-Jones object detector is used again, but in this case using a cascade of classifiers that were specifically trained for this specific domain. For the training process, more than 1,000 images were randomly selected to identify and classify the eye region in the audiometric video sequences. These images were obtained from different face images of different datasets [23] [21].

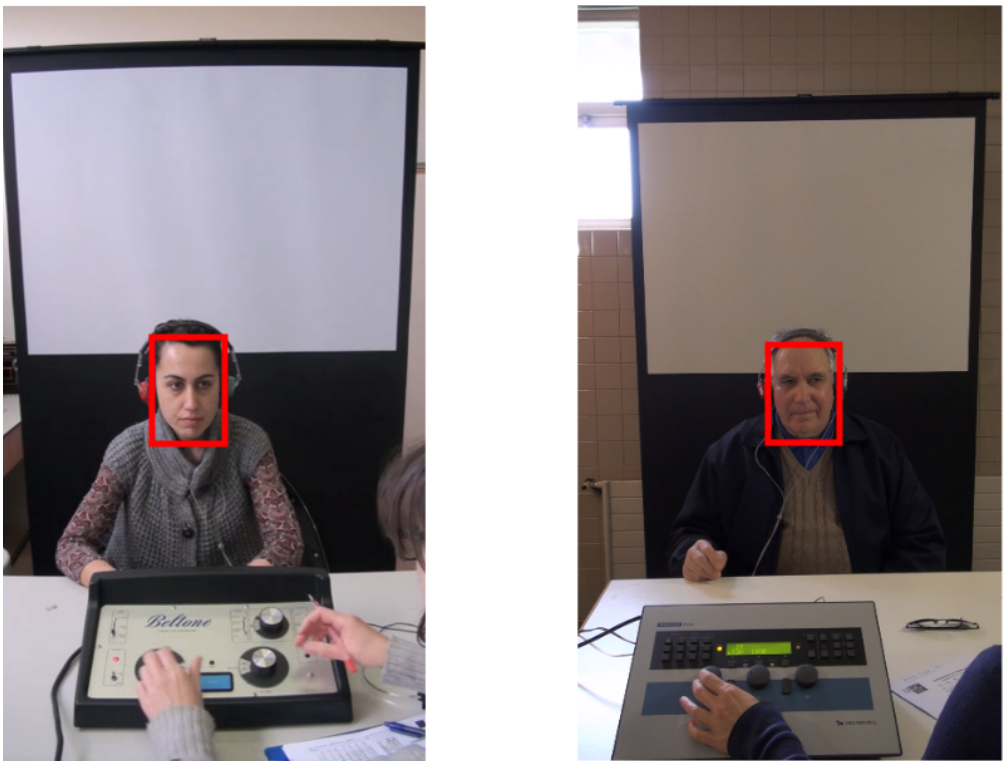


Fig. 4. Two representative examples of face detection using the Viola-Jones approach in an audiometric video sequence.

As result, an accuracy of the 98% was obtained during the evaluation of this eye detector, even when the eyes of the patient are closed. In Figure 5, we can see two representative examples of eye detection using the Viola-Jones algorithm.

Using the identified eye region as reference, the proposed method combines the optical flow information and the color distribution of the sclera to identify and classify the movements of the eyes. Each case will be discussed below.

3.3 Optical flow information

In this step of the proposed methodology, we estimate and characterize the movement of the eye region using the information provided by the optical flow. A schematic representation of the main constituent steps can be seen in

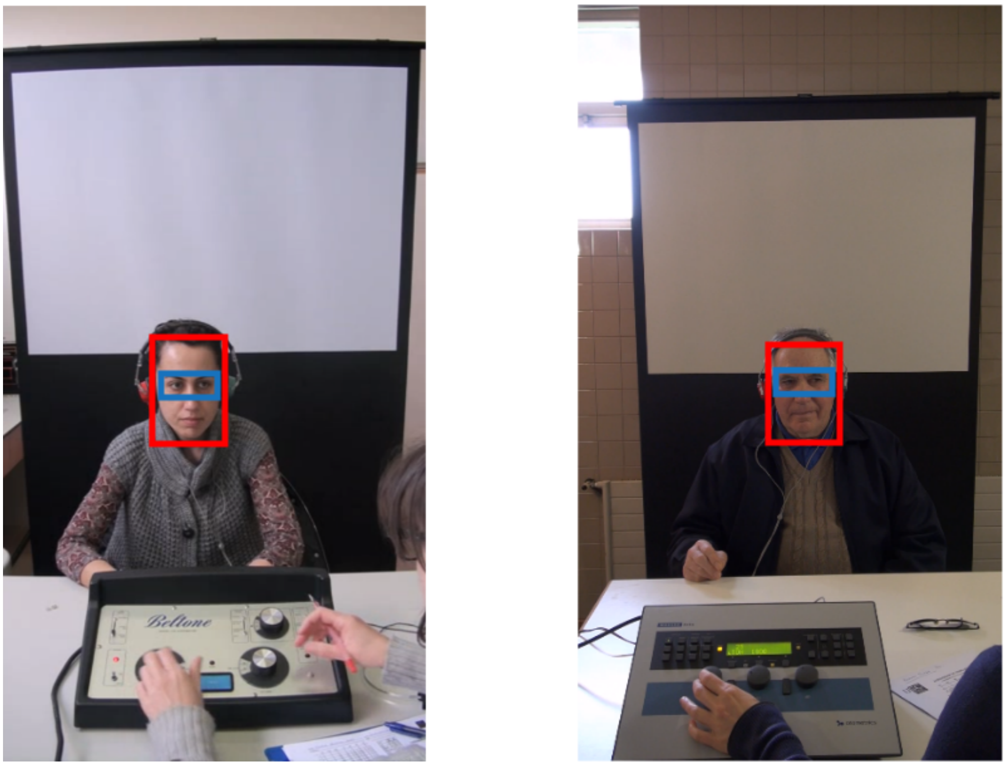


Fig. 5. Two representative examples of eye detection using the Viola-Jones approach in an audiometric video sequence.

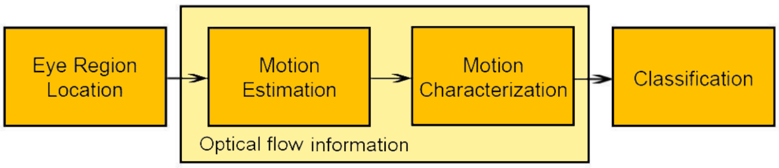


Fig. 6. Schematic representation of the optical flow information analysis.

Figure 6.

3.3.1 Motion estimation

Once the region of the eye is localized, the movements or changes in the expressions that are due to the reaction to sound stimuli are estimated only within this particular region. To do so, we use the iterative Lucas-Kanade [24] optical flow method with pyramids [25]. This method has demonstrated its suitability for the identification and analysis of the movements that can be generated by facial expression changes [19].

The recorded audiometric video sequences have associated with them a particular value of the frame rate, expressed in frames per second or fps. This term, also known as frame frequency, is the frequency with which consecutive frames appear in a video sequence. If the frame rate is high, it is possible that the analysis between a specific frame and the next one does not show significative changes, since facial expression changes cannot occur so quickly. With the purpose of allowing the facial expression changes to be notable enough, we consider a time window (t) between 2 compared frames. In this way, the optical flow is calculated between the frame (i) and the frame (i+t), being related to the frame frequency. For our audiometric video sequences, with a value of frame rate of 25 fps, the time window (t) value was empirically established to a value of 3, covering a 12% of the frame rate.

The optical flow estimation is based on the detection of different feature points. For that, we apply the Good Features to Track [26] over the reference frame (i), being its corresponding feature points extracted in the frame (i+t). Normally, these feature points are determined by a well-defined position in the image. The information provided by these feature points is rich in terms of local properties and stable to global perturbations. In Figure 7, we can see a representative example of eye movements detected within the optical flow.

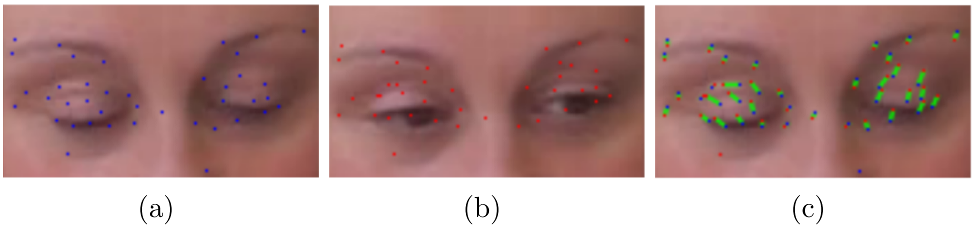


Fig. 7. A representative example of eye movements that were detected with optical flow. (a) Reference frame (i) with the feature points represented in blue. (b) Second frame (i+t) with the corresponding feature points obtained by the optical flow represented in red. (c) Movement vectors between the frame (i) and the frame (i+t) represented in green.

3.3.2 Motion characterization

The information provided by optical flow serves as reference to characterize the movements of the eyes produced in the audiometric video scene. For this purpose, a set of relevant features is used to obtain descriptor vectors that can be classified into one of our considered movements. The set of features that are used to obtain these descriptor vectors are related with the dispersion, the strength and the orientation of the motion vectors. Each motion of the eye in the video scene generates 2 different descriptors of movement, one for the right eye and one for the left eye. For the determination of these movement descriptors, the set of features is distributed into 8 ranges that are calculated according to their azimuth angles, as detailed in Equation (1).

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

where i takes different values in the range between 0 to 7, and $* \in \{L, R\}$ represents the left (L) and right (R) eyes, respectively. The descriptors are arranged according to their azimuth angle, where the 8 first values represent the number of descriptors in each particular range, as we can see in Equation (2).

$$n_i^* = |R_i^*| \tag{2}$$

The magnitude of the vector provides relevant information about the descriptor of the movement of the eye, in terms of displacement intensity values. The following 8 values of the movement descriptor are associated with the magnitude of the vector, which allows to distinguish the different types of movements in accordance with their displacement intensity values. The average of the magnitude is computed according to Equation (3).

$$m_i^* = \frac{1}{n_i^*} \cdot \sum_{v \in R_i^*} |v| \tag{3}$$

The optical flow vector allows us to differentiate between localized movements (gaze changes) and global movements (head motions). The computation of the optical flow vector is done by range, according to the respective vector angles ($v = \overrightarrow{AB}$). In this way, the centroid of the vector angles is determined according to Equation (4), where $B = (B_x, B_y)$ represents the final coordinates of this vector.

$$c_i^* = \left(\frac{1}{n_i^*} \cdot \sum_{v=\overrightarrow{AB}, v \in R_i^*} B_x, \frac{1}{n_i^*} \cdot \sum_{v=\overrightarrow{AB}, v \in R_i^*} B_y \right) \tag{4}$$

The dispersion of the optical flow vector d_i^* is calculated by means of the average distance to each obtained center c_i^*. Equation (5) represents the mathematical formulation, where $d(B, c_i^*)$ indicates the euclidean distance between the final coordinate B and the obtained center c_i^*.

$$d_i^* = \frac{1}{n_i^*} \cdot \sum_{v=\overrightarrow{AB}, v \in R_i^*} d(B, c_i^*) \tag{5}$$

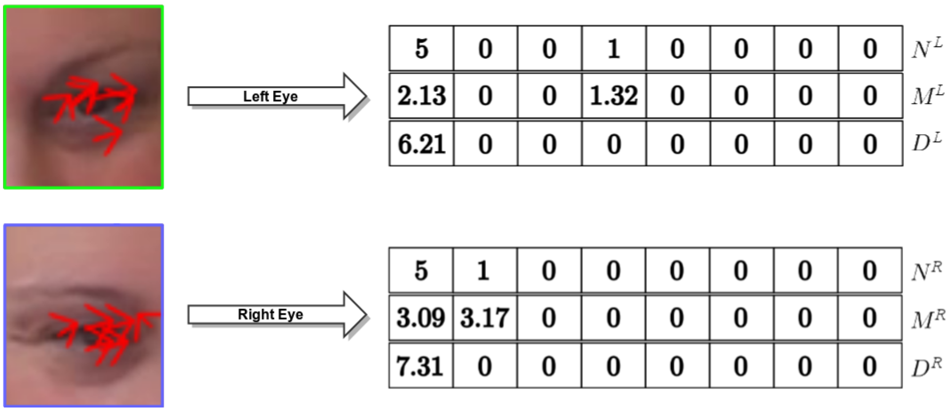
Consequently, the descriptor of the eye movement is composed of a vector of 24 calculated values. The 8 first values represent the orientation, (N^*), the next 8 values represent the magnitude, (M^*), whereas the last 8 values provide information about the dispersion (D^*), as indicated in Equations (6), (7) and (8), respectively.

$$N^* = \{n_i^* | i \in \{0...7\}\} \tag{6}$$

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

$$D^* = \{d_i^* | i \in \{0...7\}\} \tag{8}$$

In Figure 8, we can see 2 representative examples of movement descriptors, one for the right eye and other for the left eye.



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

Fig. 8. Representative examples of movement descriptors that indicate a gaze shift to the right, one for the left eye and other for the right eye. First row represents the orientation, the second one the magnitude whereas the last one the dispersion.

3.4 Color information of the sclera

In this step of the proposed methodology, we estimate and characterize the movement of the eye region using the color information of the sclera. A schematic representation of the involved main steps can be seen in Figure 9.

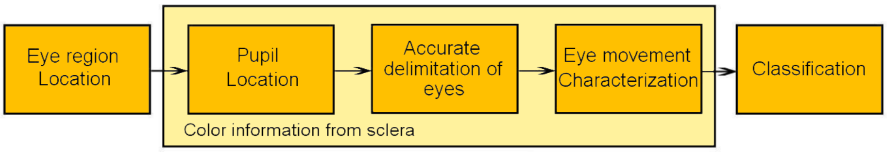

Eye region Location
Pupil Location
Accurate delimitation of eyes
Eye movement Characterization
Classification
Color information from sclera

Fig. 9. Schematic representation of the analysis of the color information of the sclera.

3.4.1 Pupil location

Once the eye region is localized, the eye movements are identified and characterized using the color information of the sclera. To do that, in the first place, we localize the center of the pupil in both eyes of the patient in an audiometric video sequence. To achieve this, a gradient-based algorithm is used [27]. In Figure 10, we can see representative examples of the obtained results, where the green points represent the detected pupil center locations during different eye motions.

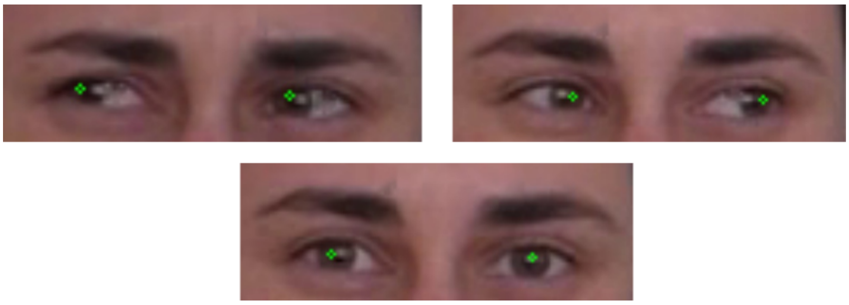


Fig. 10. Representative examples of the obtained results, where the green points represent the detected pupil center location in different eye motions.

Further, two other different approaches were studied with the same purpose: one using a gradient-based approach [28] and a second one using a Starburst approach [29].

For the validation of this stage, we used 10 audiometric video sequences from the VAST dataset. In particular, 20 frames of each audiometric video sequence were selected, which were manually labeled by a clinical expert, identifying a total of 400 samples (200 from the left eye and 200 from the right eye).

This stage was evaluated by computing the difference between the labeled center (P_e) and the calculated center (P_c), according to the results obtained from the proposed approaches. Mathematically, this metric is formulated as showed in Equation (9).

$$error = |P_e - P_c| \tag{9}$$

Table 1 summarizes the results that were obtained from each approach in terms of standard deviation and average of the obtained errors, metrics that are expressed in pixels. In general terms, all the approaches showed a satisfactory performance with the images from the VAST dataset. In particular, the best results were achieved with the gradient-based approach [27], as presented in Table 1.

3.4.2 Accurate delimitation of the eyes

Using as reference the region of the eyes and the pupil's location, we built a method that identifies the corners of the eyes in the audiometric video sequences. This stage is divided into 3 phases: the identification of the candidate points, the extraction of the reference points and, finally, the selection of the best candidates. Each one of these phases will be discussed below. A schematic representation of this stage of accurate delimitation of the eyes can be seen in Figure 11.

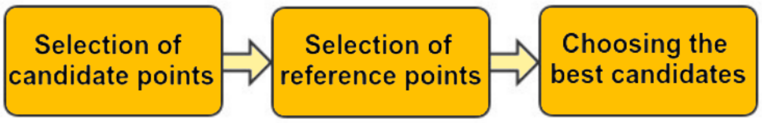

Selection of candidate points
Selection of reference points
Choosing the best candidates

Fig. 11. Schematic representation of the accurate delimitation of the eyes.

Firstly, we identify different points that are considered as candidates for the

Table 1
Results for the pupil location approaches (expressed in pixels).

		Error approach 1	Error approach 2	Error approach 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
Global	Average	3.4479	2.7711	2.7380
	Std.dev	2.0645	2.7467	2.7154

corners of the eyes. For the detection of these points, we apply the Shi-Tomasi [30] algorithm. Then, we must select the points that best represent the eyes' corners. For that, the edge detector is employed to segment the boundaries of the eyelids. To facilitate this detection, we improve the contrast of the frames by enhancing the boundaries of the eyelids. To achieve this, the frame that is obtained after the application of the erosion filter $S_f(x,y)$ is subtracted from the saturation frame $S(x,y)$, as presented in Equation (10).

$$R(x,y) = S(x,y) - S_f(x,y) \tag{10}$$

Then, a threshold value (th_s) is used for the binarization of the frame, as indicated in Equation (11), where μ represents the average from the difference image (I_{diff}) and σ indicates the standard deviation.

$$th_s = \mu(I_{diff}) + 0.75 * \sigma(I_{diff}) \tag{11}$$

And finally, the binarization image is computed using the Equation (12), where $I_{ths}(x,y)$ represents the thresholded frame.

$$I_{ths}(x,y) = \begin{cases} 1, & \text{if } I_{diff}(x,y) > th_s; \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

Taking into account the different anthropometric restrictions of the eyes, we may consider the corners of the eyes as points of union between the ellipses that correspond to the eyelids. Our method uses these points as reference to eliminate possible false candidates. In this way, we select the candidate point that is closest to the reference point. In Figure 12, we can observe the different points that are obtained in this stage of the methodology. The yellow points are the candidate points, the red points represent with the reference points and the green points correspond the final points that were selected as eyes' corners.

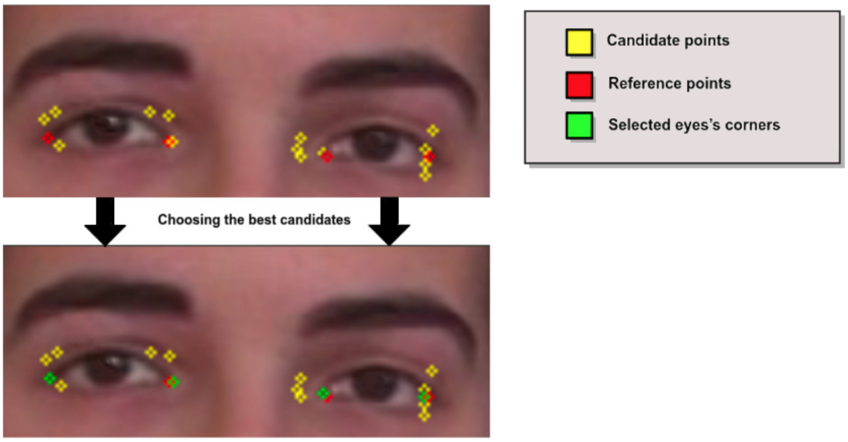


Fig. 12. Representative example of the obtained points for the eyes delimitation, where the yellow points are the candidate points, the red points represent with the reference points and the green points correspond the final points that were selected as eyes' corners.

3.4.3 Eye movement characterization

Once the corners of the eyes are localized, the eye motion is characterized using the color information provided by the sclera (the white region of the eyes). To achieve this, we estimate the sclera region using the corners of the eyes, previously detected. Firstly, the input frame is transformed to the grayscale space, applying the histogram equalization technique across it. Subsequently, a distribution of the gray level is calculated to represent the intensity values of each pixel that belongs to the line that joins the corners of both eyes, as we can see in the illustrative example of Figure 13.

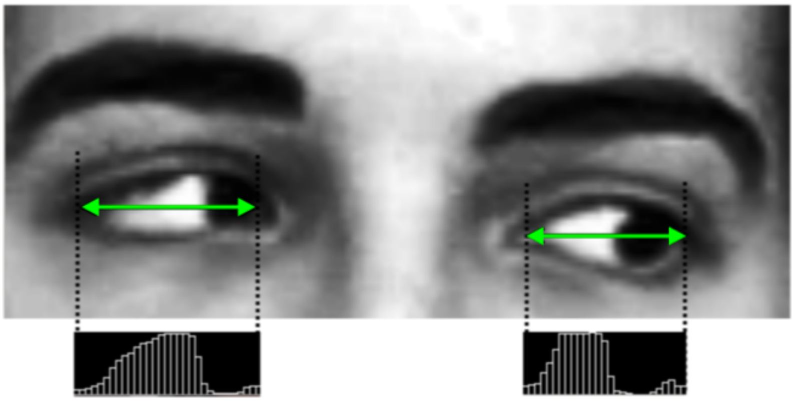


Fig. 13. Representative example of the gray level distribution for a gaze shift to the right.

Once the gray level distribution is calculated, it is possible to delimit it into 3 regions of interest: the right side of the sclera, the left side of the sclera and the iris region. To that end, we make use of the information that is provided by the center of the pupil, the corners of the eyes and the diameter of the iris, as we can see in the example of Figure 14.

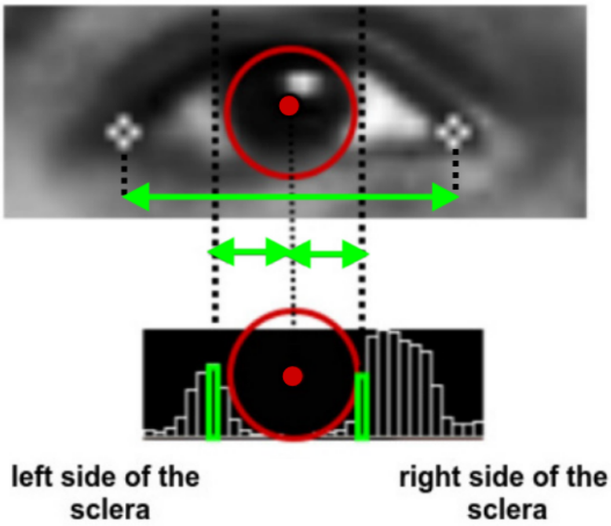


Fig. 14. Representative example of the delimitation of the 3 regions of interest over the gray level distribution.

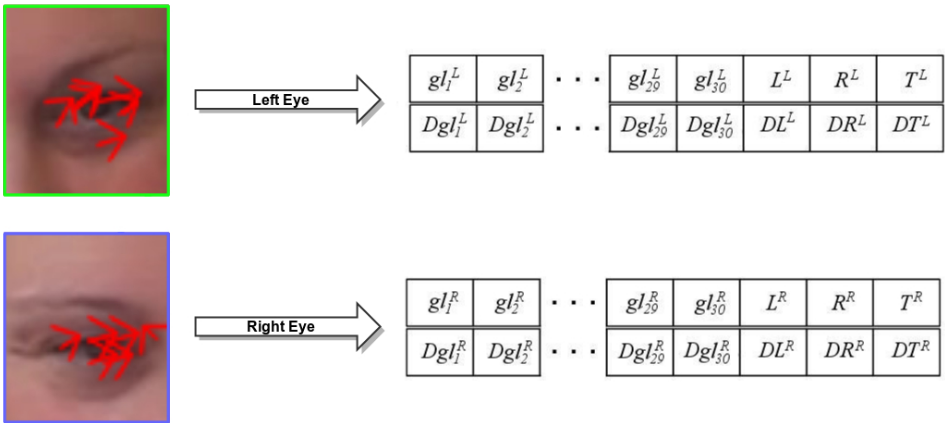
For the motion characterization, we use a movement descriptor composed of a feature vector of 66 values. In this, 33 values represent the current state and other 33 values represent the previous state. In addition, considering the defined sclera regions, we included 3 additional features to the movement descriptor: the sum of the gray intensities for each sclera region and the global sum of both regions. Thus, the subtraction of the movement descriptors of the current frame i and the frame i+t is calculated. In this way, we add 33 new features to the final movement descriptor, as we can see in Figure 15. First row represents the first 30 values of the gray-level distribution, where L describes the sum of the gray intensities in the left sclera region, R represents the sum of the gray intensities in the right sclera region and T indicates the sum of both sclera sides (left and right). Second row represents the first 30 values of the gray-level subtraction of the frame i and the frame i+t (Equation 13), DL corresponds to the subtraction of the sum of gray intensities of the left sclera region (Equation 14), DR represents the subtraction of the sum of gray intensities of the right sclera region (Equation 15), and DT represents the subtraction of the sum at both sclera regions (Equation 16).

$$Dgl_n = gl_{i+t,n} - gl_{i,n}, \mathit{where}\ n \in [1, 30] \tag{13}$$

$$DL = L_{i+t} - Li \tag{14}$$

$$DR = R_{i+t} - Ri \tag{15}$$

$$DT = T_{i+t} - Ti \tag{16}$$



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

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

Fig. 15. Representative examples of movement descriptors, one for the left eye and other for the right eye. First row represents the first 30 values of the gray-level distribution whereas the second one the first 30 values of the gray-level subtraction of the frame i and the frame i+t.

3.5 Final classification

Finally in this step, we classify the eye movements into the 4 considered categories: Eye Closure (EC), Eye Opening (EO), Gaze shift to the Left (GL) and Gaze shift to the Right (GR). To achieve this, we use the combined movement descriptors that were calculated from both branches of the proposed methodology, the optical flow information and the color information from the sclera. An experiment was conducted in order to determine the most appropriate model learning. In particular, a training step was performed using the following classify models: Random Forest, Naive Bayes, Random Committee, C4.5, Random Tree, Logistic, Multilayer Perceptron, Logistic Model Tree (LMT) and Support Vector Machines (SVM).

4 Experimental results

We conducted different experiments to validate the suitability of the proposed method using the public Video Audiometry Sequence Test (VAST) [21] dataset that was presented in Section 3.1. As indicated, this dataset includes a significant variability of conditions with 11 audiometric video sequences of different individuals as well as different ages, genders and cognitive abilities. In the experiments, we compared the results of the proposed method with the manual labeling of a clinical expert.

For the validation of the method, the different movement descriptors were associated with different types of eye movements. This correlation was established by the clinical experts for this specific domain. In particular, as indicated, 4 typical eye movements were identified as the most relevant: Eye Closure

(EC), Eye Opening (EO), Gaze shift to the Left (GL) and Gaze shift to the Right (GR). The audiometric video sequences were analyzed frame-by-frame, identifying all the relevant eye movements. In this way, 1,180 significant eye movements were labeled and classified in their appropriate category. Table 2 shows the distribution of the eye movements that are assigned to different relevant categories. In particular, we can observe a small imbalance for the classes GL and GR.

Table 2
Distribution of the significative eye movements that are assigned to the different relevant categories.

	Eye open (EO)	Eye close (EC)	Gaze left (GL)	Gaze right (GR)
Nº of samples	408	310	230	232

Using all the eye movement descriptors, a set of representative classifiers was used to test the potential of the proposed method. In particular, we use the previously indicated classifiers: Random Forest, Naive Bayes, C4.5, Random Committee, Random Tree, Logistic, Multilayer Perceptron, Logistic Model Tree (LMT) and Support Vector Machines (SVM). The classification step was performed by a 10-fold cross-validation, being calculated the mean accuracy to illustrate the overall performance of the proposed system. Table 3 details the results that were achieved by each trained model. Further, we present a comparison between the optical flow descriptors and the optical flow descriptors combined with the color from the sclera descriptors.

Analyzing the results, we can see that the information provided by the sclera movement descriptors globally improves the results of the proposed method. In general, all the models showed a satisfactory performance with the considered dataset. In particular, the best results were achieved with the SVM classifier,

Table 3
Classification accuracy comparative: left column, optical flow descriptors; right column, optical flow descriptors combined with the 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%**

reaching an accuracy of 97.46%. Table 4 shows a more detailed analysis, presenting the true-positive rates for the optical flow descriptors combined with the color information from the sclera.

Table 4
Classification results by eye movements classes for the combined approach using a 10-fold cross validation for different classifiers.

		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

Regarding the true-positive rates, we can see that the learning models are often biased towards the learning of the majority class. Furthermore, we emphasize that the minority classes, GR and GL, are the most relevant for this specific domain, so we will try to improve their classification. To address this imbalance, we apply the SMOTE (Synthetic Minority Oversampling TEchnique) [31][32] algorithm, which represents an oversampling technique that adds samples to the original dataset until the class distribution becomes balanced. In this way, we apply oversampling values of 100%, 200%, 300% and 400% over the number of samples in the minority classes. Table 5 details the results that were achieved with the SVM classifier applying the different progressive levels of oversampling.

Table 5
Classification results using the SVM classifier applying the different progressive 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**

As we can see, in this experiment, the proposed method achieved the best results for the GR and GL classes using an oversampling rate of 400%. Nevertheless, using an oversampling rate of 300%, the method maintains the global accuracy values with an important increase in the true-positive rates for the detection of gaze shifts. Besides, we wish to highlight that when using these configurations, the true-positive rates for the identification of EO and EC decrease. Despite this, these values can be considered acceptable, since these classes have a lower incidence in the audiometric tests.

4.1 Detection of reactions to sound

As mentioned above, the significative contribution of the proposed system is to provide an adequate detection of eye gestural reactions to sound in audiometric video sequences. Therefore, for the evaluation of these eye gestural reactions, we used 3 audiometric video sequences from the VAST dataset. In particular, each video sequence was analyzed frame-by-frame. The selected frames were manually labeled by a clinical expert, identifying a total of 108 gestural reactions to sound. For the manually labeled groundtruth, as gold standard, we consider that an eye gestural reaction can be determined when 3 or more consecutive frames are assigned to the same class in the classification process. Table 6 details the results that were achieved for the detection of the most relevant reactions to the auditory stimulus, considering the agreement between the proposed system and the clinical experts.

Table 6
Results for the detection of eye gestural reactions to sound in audiometric video sequences.

	Sound reactions	
	Expected	Detected
Video 1	32	32
Video 2	38	38
Video 3	38	38
Agreement	100%	

As we can see, the agreement between the proposed method and the clinical experts is complete for the audiometric video sequences that were tested in this experiment, being able to detect the 100% of gestural reactions to sound. In particular, the obtained results are motivated by the characteristic reaction to sound, that typically lasts between 5 and 15 frames and, consequently, even the misclassification of one particular frame does not significantly influence the correct final identification of the gesture reaction.

5 Discussion and conclusions

In this paper, we propose a complete methodology for the automatic evaluation of eye gestural reactions to sound by the analysis of audiometric video sequences. To do so, the proposed system used the information provided by two different and independent approaches. The first approach was performed using optical flow information and machine learning algorithms. The second approach addresses an alternative solution using the information that is provided by the color distribution of the sclera. Then, a combination of both strategies was proposed to improve the overall accuracy of the proposed system.

Regarding the results, we tested the performance of the proposed method using the public Video Audiometry Sequence Test (VAST) dataset. This dataset contains 11 audiometric video sequences of different individuals as well as different ages, genders and cognitive abilities. The system achieved satisfactory results, reaching a best accuracy of 97.46% of classification into the considered movement categories using an SVM classifier that was trained with a 10-fold cross-validation. In addition, the oversampling technique SMOTE was applied to improve the accuracy of the minority classes, GL and GT. Using an oversampling rate of a 300%, the proposed system achieved satisfactory results reaching true positive rates of 0.986 and 0.966 for the GL and GR classes, respectively, with a global accuracy of 0.971. For the automatic detection of eye gestural reactions to auditory stimuli, the proposed method was able to detect the 100% of the gestural reactions in the audiometric video sequences. In conclusion, the proposed system is able to identify and classify the different eye gestural reactions to sound with reasonable detection rates. Therefore, this facilitates the work of the clinical experts in the analysis and diagnosis of hearing impairments of patients with different degrees of cognitive impairment or other serious communication problems.

As future work lines, further validations could be implemented by increasing the dimensionality of the VAST dataset to provide a more exhaustive validation of the proposed method. Besides, the information provided by the audiometer could be analyzed and correlated with the gestural responses to the sound. Therefore, it will be possible to follow the evolution of the hearing impairment more accurately and, consequently, improve the life quality and well-being of the patients.

Acknowledgements

This work is supported by the Instituto de Salud Carlos III, Government of Spain and FEDER funds of the European Union through the DTS18/00136 research projects and by the Ministerio de Economía y Competitividad, Government of Spain through the DPI2015-69948-R research project. Also, this work has received financial support from the European Union (European Regional Development Fund - ERDF) and the Xunta de Galicia, Centro singular de investigación de Galicia accreditation 2016-2019, Ref. ED431G/01; and Grupos de Referencia Competitiva, Ref. ED431C 2016-047.

References

[1] Collins, J.G. and National Center for Health Statistics (U.S.), Prevalence of Selected Chronic Conditions: United States, 1986-88, DHHS publication, U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Health Statistics, 1993.
URL `http://books.google.es/books?id=fYbFnQEACAAJ`

[2] C. Murlow, C. Aguilar, J. Endicott, R. Velez, M. Tuley, W. Charlip, J. Hill, Asociation between hearing impairment and the quality of life of elderly individuals, Vol. 38, J Am Geriatr Soc, 1990, pp. 45–50.

[3] A. Davis, The prevalence of hearing impairment and reported hearing disability among adults in great britain, Int J Epidemiol. 18 (1989) 911–917.

[4] IMSERSO, Las personas mayores en España, in: Instituto de Mayores y Servicios Sociales, 2008.

[5] W. H. Organization, et al., Prevention of blindness and deafness, Global initiative for the elimination of avoidable blindness. Geneva: WHO.

[6] National Institute of Deafness and Other Communication Disorders, Quick statistics, 2014.
URL `http://www.nidcd.nih.gov/health/statistics/Pages/quick.aspx`

[7] Australian Hearing Annual Report, 2009.
URL `http://www.hearing.com.au/australian-hearing-annual-reports`

[8] IMSERSO, Libro blanco del envejecimiento activo (in Spanish), 2010.

[9] K. Tambs, Moderate effects of hearing loss on mental health and subjective well-being: results from the nord-trøndelag hearing loss study, Psychosomatic medicine 66 (5) (2004) 776–782.

[10] Q. Acton, Dementia: New Insights for the Healthcare Professional: 2013 Edition, ScholarlyEditions, 2013.
URL http://books.google.es/books?id=pkE0Yv1Mf0QC

[11] F. R. Lin, Hearing loss and cognition among older adults in the united states, in: The Journals of Gerontology: Series A, 2011.

[12] F. Lin, K. Yaffe, J. Xia, Q. Xue, T. Harris, E. Purchase-Helzner, L. Ferrucci, E. Simonsick, Hearing loss and cognitive decline among older adults, in: Gerontologist, Vol. 52, 2012, pp. 508–508.

[13] F. R. Lin, K. Yaffe, J. Xia, Q.-L. Xue, T. B. Harris, E. Purchase-Helzner, S. Satterfield, H. N. Ayonayon, L. Ferrucci, E. M. Simonsick, et al., Hearing loss and cognitive decline in older adults, JAMA internal medicine 173 (4) (2013) 293–299.

[14] B. S. of Audiology, Recommended Procedure Pure-tone air-conduction and bone-conduction threshold audiometry with and without masking, 2015.

[15] L. De Santi, P. Lanzafame, B. Spanò, G. D'Aleo, A. Bramanti, P. Bramanti, S. Marino, Pursuit ocular movements in multiple sclerosis: a video-based eye-tracking study, Neurological Sciences 32 (1) (2011) 67–71.
URL http://dx.doi.org/10.1007/s10072-010-0395-1

[16] M. L. Pereira, M. v. Camargo, I. Aprahamian, O. V. Forlenza, Eye movement analysis and cognitive processing: detecting indicators of conversion to Alzheimer's disease, Neuropsychiatr Dis Treat 10 (2014) 1273–1285.

[17] G. E. Raney, S. J. Campbell, J. C. Bovee, Using eye movements to evaluate the cognitive processes involved in text comprehension, J Vis Exp (83) (2014) e50780.

[18] R. Z. Marandi, S. H. Sabzpoushan, Using eye movement analysis to study auditory effects on visual memory recall, Basic Clin Neurosci 5 (1) (2014) 55–65.

[19] A. Fernández, M. Ortega, M. Gonzalez Penedo, C. Vazquez, L. Gigirey, A methodology for the analysis of spontaneous reactions in automated hearing assessment, Biomedical and Health Informatics, IEEE Journal of 20 (1) (2016) 376–387.

[20] A. Fernández, J. de Moura, M. Ortega, M. G. Penedo, Detection and characterization of the sclera: evaluation of eye gestural reactions to auditory stimuli, in: 10th International Conference on Computer Vision Theory and Applications (VISAPP) - Vol.2, 2015, pp. 313–320.

[21] Video Audiometry Sequence Test (VAST)- 2018, accessed: 2019-01-14.
URL `http://www.varpa.org/research audiology.html`

[22] P. Viola, M. Jones, Robust real-time object detection, in: International Journal of Computer Vision, 2001.

[23] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, Y. Yao, Interestingness prediction by robust learning to rank, in: European conference on computer vision, Springer, 2014, pp. 488–503.

[24] B. D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2, IJCAI'81, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1981, pp. 674–679.

[25] J.-Y. Bouguet, Pyramidal implementation of the Lucas-Kanade feature tracker: Description of the algorithm, Intel Corporation, Microprocessor Research Labs.

[26] J. Shi, C. Tomasi, Good features to track, in: Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on, 1994, pp. 593–600.

[27] F. Timm, E. Barth, Accurate eye centre localisation by means of gradients., in: L. Mestetskiy, J. Braz (Eds.), VISAPP, SciTePress, 2011, pp. 125–130.

[28] R. Kothari, J. Mitchell, Detection of eye locations in unconstrained visual images, in: Image Processing, 1996. Proceedings., International Conference on, Vol. 3, 1996, pp. 519–522.

[29] D. Li, D. Winfield, D. Parkhurst, 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, 2005, pp. 79–79.

[30] J. Shi, C. Tomasi, Good features to track, in: 1994 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94), 1994, pp. 593 – 600.

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

[32] V. Bolón-Canedo, A. Fernández, A. Alonso, M. Ortega, M. G. Penedo, On the use of machine learning techniques for the analysis of spontaneous reactions in automated hearing assessment, in: European Symposium on Artificial Neural Networks, 2015, pp. 355–360.