Detection of reactions to sound via gaze and global eye motion analysis using camera streaming

Alba Fernández · Marcos Ortega · Joaquim de Moura · Jorge Novo · Manuel G. Penedo

Abstract This work is focused on the field of automatic hearing assessment for patients presenting cognitive decline or severe communication difficulties. Audiometry is a test of behavior requiring intense interaction between patient and audiologist, so it is extremely difficult to properly assess patients with severe communication disorders. However, patients with cognitive decline often make some eye gestures in reaction to auditory stimuli. These reactions are interpreted by expert audiologists. On the other hand, a manual assessment of the patient creates problems such as subjectivity or low reproducibility, to name but two. Bearing this in mind, this paper introduces a novel methodology to analyze video recordings acquired during audiometric evaluations and characterize movements of the eye so they can be interpreted as a positive gestural reaction to sound. Motion analysis in the eye region helps the human expert to establish the existence of a reaction to sound, thus increasing the reproducibility and objectivity of the test.

Keywords Auditory responses · computer vision · gestural reactions · hearing screening · movement detection · screening methods

1 Introduction

Hearing loss is an invisible condition involving a sudden or gradual decrease in hearing. Its prevalence in older adults ranks third among chronic health conditions [8] and is also one of the most widely under-treated. It can be present at any age but the loss of hearing capacity increases over the years, especially at high frequencies. In this sense, Murlow *et al.* [24] and Davis's [10] studies indicate that, of all the disabilities, hearing loss is the most closely linked to the aging process. Given the aging phenomenon in the world population nowadays [9] [15] the number of elderly people presenting cases of hearing loss is growing fast. A total of 700 million people are currently estimated to present some degree of hearing impairment.

A study by the National Institute of Deafness and Other Communication Disorders [25] shows that about 2% of people aged between 45 and 54 present some kind of disabling hearing loss. In the 55 to 64 range the rate is 8.5%, almost 25% for the 65 to 74 range and 50% for people over 75 years of age. Meanwhile, the Australian Hearing Annual Report [3] shows that over 50% of individuals in the 60-70 years of age range present a case of hearing loss, with the rate rising to 70% for individuals over 70. Generally speaking, the world population is older, as indicated in IMSERSO's [16] book of active aging.

Different studies, such as [9], have exposed the undesirable effects that hearing loss may have on the well-being of an individual not only physically but psychologically or socially. Indeed, people presenting some kind of hearing impairment may suffer communication difficulties negatively affecting their social lives which, in turn, may lead to further isolation or other problems. The use of hearing aid solutions or rehabilitation techniques is clearly relevant in terms of improvement of well-being and quality of life.

Aging also increases the probability of having neurodegenerative disorders and problems with proper com-

Alba Fernández, Marcos Ortega, Joaquim de Moura, Jorge Novo, Manuel G. Penedo
Department of Computer Science, University of A Coruña
CITIC-Research Center of Information and Communication Technologies, University of A Coruña
Campus de Elviña, S/N, P.C. 15071, A Coruña (Spain)
E-mail: alba.fernandez@udc.es, mortega@udc.es, joaquim.demoura@udc.es, jnovo@udc.es, mgpenedo@udc.es

munication. Alzheimer's disease is a very common form of neurodegenerative disorder typically affecting individuals over 65 years old [1]. The prevalence of neurodegenerative disorders and especially of Alzheimer's disease is increasingly significant. The number of people presenting symptoms of Alzheimer's disease or some other form of dementia is estimated at nearly 44 million people. With current trends this number will reach 76 million by 2030 and exceed 135 million by 2050.

Furthermore, recent research indicates that loss of hearing may be a risk factor for cognitive impairment [20]. In particular, a possible association between hearing loss and aggravation of Alzheimer's disease is established. As stated before, hearing impairment may result in isolation of the individual entailing long-term effects to brain functioning. Moreover, a hearing disability typically forces the brain to use more energy in sound processing, reducing the energy devoted to other activities including recall or thinking. The correlation between these conditions is a major drawback for the diagnosis of hearing disabilities, especially considering that most adults will develop some kind of cognitive decline with time. Also, given the association of age with both conditions, it is normal for hearing impairment and cognitive decline to appear in the same subject.

All these circumstances highlight the need to perform regular checks on the hearing capacity of older adults, this also being highly recommended for any individual if a suspicion of hearing ability issues arises.

Pure Tone Audiometry (PTA) is unequivocally described as the gold standard test for the clinical evaluation of hearing abilities. It establishes the finest tones the subject is able to perceive as sound at certain frequencies. By means of this test the audiologist is able to evaluate the hearing capacity of the patient and also to determine the prevalence of hearing loss conditions. As it relies on the responses of the patient to the stimuli, some degree of subjectivity is expected to be involved in the test. And since it is a behavioral test certain operational limitations may be present, particularly for disabled subjects.

To perform the audiometry, pure tone sounds are emitted to the patient via earphones. Then, the patient must give a sign if the stimulus is perceived [2]. This sign can be typically raising his/her hand. The test is performed manually by the audiologist who also checks for the signs. However, when the subject presents some cognitive decline, the standard procedure may not be applicable as the typical interaction between the audiologist and the patient is not feasible.

The complications of this situation may be overcome by an expert audiologist: cognitively impaired people may not show any voluntary response (e.g. hand raising). Instead, they typically react in an unconscious way via slight facial gestures which are mainly focused in the eye region. Typically, changes in gaze direction or opening of the eyes are signs that indicate the patient has perceived the emitted sound. With this in mind, the audiologist must focus on the appearance of unconscious gestures in the eye to obtain clues about a possible perception. The clinical interpretation of such reactions, however, necessarily demands a high degree of experience on the part of the clinician. Moreover, the particular nature of gestures as a reaction to sound is totally dependent on the subject, as each individual could react in a different way. Even the same patient can be expected to offer different gestures at different times, given their unconscious nature. Taking all these circumstances into account, the subjectivity of the procedure is very high. As is always the case, this is a major limitation of the reproducibility and robustness of the evaluations performed over different sessions or by different members of the clinical staff. As a consequence, inaccuracies in the assessments may arise.

Several works have presented methodologies for eye movement analysis in patient assessment scenarios. In [11] a video-based eye tracking solution is presented to analyze eye dysfunctions in multisclerotic patients. Also related to visual deficits is the work presented in [26] where eye movement patterns were processed to find a correlation with symptoms of Alzheimer's disease. Also, in the proposal of [5], eye tracking was used to diagnose autism spectrum disorders. In [14] the authors proposed a clinical application of eye movement as an aid to understand Parkinson's disease pathophysiology. Eye movement assessment is also useful for behavioral analysis. In [27] it is used in order to evaluate the procedure of text comprehension. Also, in [23] a study on visual memory recall combining image and sound was performed considering subjects' eye movements. A common characteristic in most of these works is the lack of analysis of the convenience of the proposed methods in order to cover all possible statuses of a patient. Also, as reactions are spontaneous, it makes it harder to apply existing solutions given the fact that the reaction must be detected and classified with no prior knowledge of when (or if) this will happen.

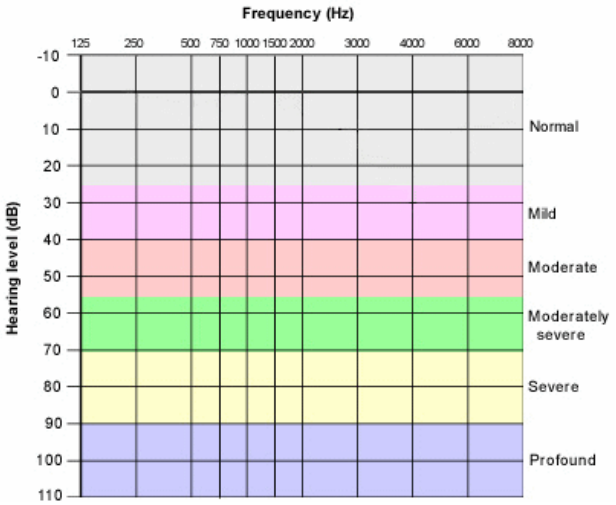
In this paper, a new approach for the classification of movements in the eye region is proposed specifically aimed at audiometric processes. Our methodology combines a series of pattern recognition and computer vision techniques to process the videos recorded during the audiometric test. This proposal is a combination of two methods; on the one hand, optical flow is used for the detection of eye region motion, whilst on the other, color information from the sclera region of the eye is consid-

ered in order to characterize the direction of the subject's gaze. The use of optical flow information has been previously studied and justified in [13], and the use of color information from the sclera was initially proposed in [12]. Considering this, we propose the combination of the two sources of information in order to obtain a more complete and robust characterization of eye movements, so they can be used in conjunction with one another. The major challenge of the proposed approach is the recognition of eye gestures emitted as reactions to the sound perceived by the patient. Most of the cases involve gaze direction changes. The main difficulties are associated with: variability in the subtlety of the eye movements, as they may be more or less subtle depending on the patient, the appearance of wrinkles or lighting changes in the scene, all of which may affect the process. It is important to clarify that the goal of this work is not the development of new eye tracking methodologies but rather the development of a single methodology valid for the characterization of spontaneous reactions in the eye region of individuals with varying degrees of cognitive damage. Therefore, as explained above, traditional eye-tracking approaches are neither applicable nor comparable since the spontaneous reactions are often global movements in the eye region or closing/opening of the eyes. Moreover, as each individual reaction is different, a further classification of the movements is performed in order to facilitate the future study of the characterization of patients by their particular reactions. With this scope in mind, the methodology proposed here shows promising results, despite the above-mentioned adverse conditions, by combining optical flow information and color information from the sclera.

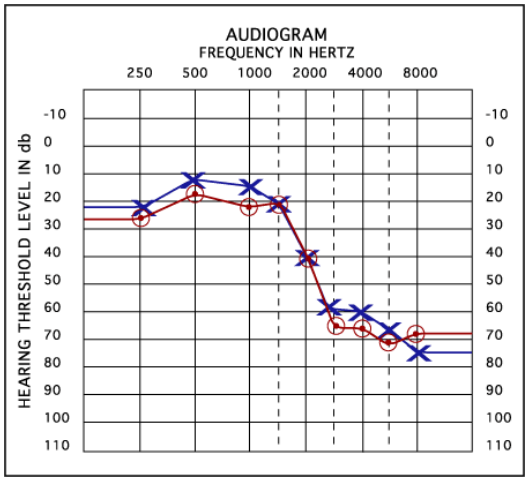
The paper is organized as follows: Section 2 details the main steps of an audiometric assessment. Section 3 introduces the proposed methodology. Section 4 is devoted to the experimental results obtained for the evaluation of the methodology. Finally, Section 5 provides a discussion and some conclusions.

2 Audiometric procedure

As stated before, conducting regular hearing checks is of particular importance for people over 60 years old. The prescription of hearing aids can improve the feelings of frustration and isolation of people with hearing loss. PTA is the standard test to identify the threshold levels of hearing for a particular individual. It allows the softest sound that a patient can perceive to be determined in a pre-established environment. As previously mentioned, PTA is a behavioral test in which auditory stimuli at certain ranges of frequencies and intensities are delivered to the patient. The test requires cooperation since the patient is asked to indicate when he or she has perceived the stimulus. The results of the hearing test are plotted on a graph called an audiogram (see Figure 1(b)), which is a two dimensional graph displaying intensity and frequency. The frequency in hertz (Hz) is displayed on 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 1(b)). The position of these symbols on the chart indicates the quietest sounds the patient can hear at different frequencies.



(a)



(b)

Fig. 1 Audiogram samples. (a) According to the results charted here, hearing can be classified as normal or with mild, moderate, moderately severe, severe or profound loss. (b) The X's and blue lines are responses for the left ear and the 0's and red lines correspond to the right ear.

Prior to any exploration, the audiologist gives the patient an explanation of the protocol that must be

followed during the audiometry. At this stage, it is crucial that the patient understands the given instructions. This requirement of understanding between patient and audiologist is what makes the test performance difficult when subjects present some cognitive decline or patients with significant hearing loss are not wearing a hearing aid. For patients without impairments, they are typically required to respond to the perceived auditory stimuli by clearly raising their hand. In the case of patients with cognitive decline, understanding of the procedure or hand raising protocol may not be a possibility or it is not informative enough. In that case, the specialist focuses on spontaneous eye region gestures in order to obtain some indication of reaction to sound.

Our proposed approach for eye-based reaction analysis complies with the clinical protocol as it does not involve any change in relation to the actions of the specialist or the patient. This is especially important since, as far as audiologists are concerned, it is necessary to respect the environment of the test as far as possible without disturbing the concentration of the patient. This is a clear advantage of the approach since it only requires the capture of audiometry procedures with a conventional video camera without influencing the audiometry protocol. Our methodology is presented in the next section.

3 Methodology

A flowchart of the main stages of the proposed methodology can be seen in Figure 2. As can be observed, starting from the video signal acquired the region of interest (i.e. the eye region) is located. Focusing on the selected region of interest, the proposed methodology combines optical flow information about changes in the region of interest and color information from the sclera in order to analyze eye gestural reactions. The following sections will explain each step of the schema.

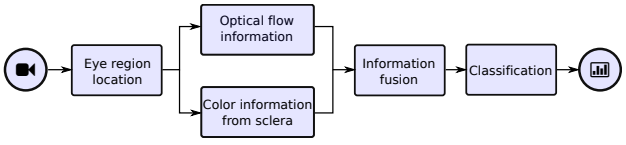


Fig. 2 Main steps of the proposed methodology.

3.1 Video signal acquisition

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 in different situations, under different lighting conditions and from a wide range of patients. They were annotated by audiologists in order to test the proposed automated assessments. All images were acquired and annotated by audiologists from the Faculty of Optics and Optometry, University of Santiago de Compostela (Spain).

The recorded video sequences, a total of 11 with a duration ranging from 4 to 8 minutes, were obtained using a conventional video camera. The only requirements for the video sequences used with this methodology are that they must be high resolution in order to cover the whole scene with sufficient definition and have a frame rate of 25 frames per second.

In order to maintain maximum possible stability of the clinical protocol and not disturb the patient's concentration, the video camera should be located in a discrete position behind the audiologist (the audiologist will be seated facing the patient). This location will allow us to record not only the patient (who will be seated in a specific position) but also the audiometer through which the audiologist will be delivering the auditory stimuli that the patient perceives via earphones. The approximate distance from the camera lens to the patient should be two meters. A schematic representation of this scene can be observed in Figure 3.

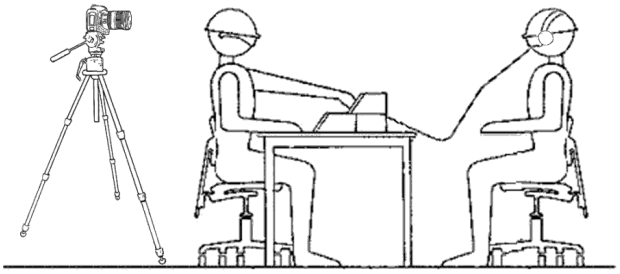


Fig. 3 Schematic representation of the scene.

The reason for such a general perspective is that it is necessary to study not only the patient's responses (whether expressed by hand raising or eye-based gestural reactions) but also the delivery of the stimuli that the expert sends via the audiometer. The need for such a general perspective in conjunction with the precision required to evaluate eye gestural reactions is the reason why HD video is required for the development of this methodology. A sample of the recorded scenes is presented in Figure 4. Of course, the development of an automated methodology for the recognition of eye gestural reactions is of major relevance in order to im-

prove objectivity and reproducibility in the evaluation of these patients.

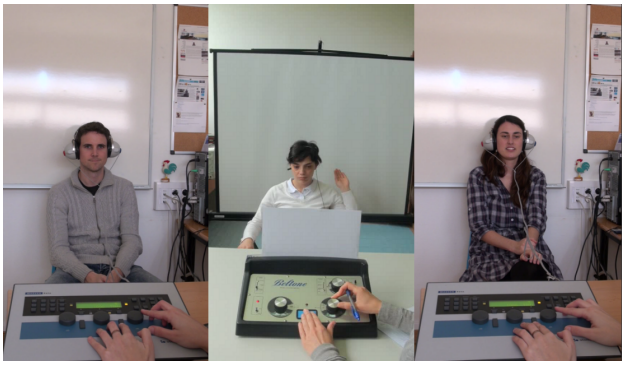


Fig. 4 Sample of the recorded images.

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

Changes in lighting conditions during the performance of the audiometric evaluation can cause shadows or other situations that substantially modify the appearance of the scene and can lead to detection inaccuracies. For this reason, maintaining adequate and homogeneous lighting conditions is recommended when recording the scenes to guarantee quality recordings and to avoid disturbances such as shadows or occlusions.

In this work, the audiometric evaluations are conducted in a controlled environment with restricted conditions. In this way, the impact of the changes in the illumination conditions as well as other intensity characteristics are reduced, without significantly affecting to the performance of the proposed method.

3.2 Eye region location

Proper location of the eye region is the first step of the methodology, since the recorded scene is much larger and contains a great amount of information (the image of the patient from the waist up, the audiometer, the background, etc.). We therefore 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, the specific region of the eye is located within the general region of interest. The initial location of the face allows for a drastic reduction of the search area, greatly reducing the number of errors and the computational cost of the following stage of the methodology.

The particular setup for this domain is stable regarding location of the items of interest: the specialist sits in front of the patient while the camera is placed behind the specialist in order to ensure that the patient's face is recorded with a frontal pose. Ensuring a frontal pose for the patient enables the application of the approach developed by Viola and Jones [31] [17]. This approach is a framework for object detection in a scene providing high rates of precision in low computation time. Although it can be trained for any particular kind of object, it was initially designed for face detection. As a consequence, an optimized frontal face detector is available in the Open CV library. This detector is not a flexible solution, but due to the specific and stable conditions of our domain its use is highly appropriate. In this context, the used method provided accurate results, returning a 100% of correctly detected faces for our video dataset.

Once the face region has been located, the specific eye region must be located within it. In this way, the proposed method only retrieves one identification of the eye region, given that in the restricted region of the face we can assume the existence of a single one. Therefore, we can obtain two different results for the eye detector method: the identification of a single region or no region. In order to locate eye region, a cascade of classifiers was specifically trained using over 1000 different eye region samples. All these images were manually selected in order to delimit our ROI. The training set was built by cropping facial images from different image databases. Testing the trained system in our video sequences, an accuracy of 98% is obtained for the eye region detection task. In our tests, the proposed method returned no false positives for eye region. An important feature to note is that this approach is able to detect the eye region regardless of facial expression and even if the eyes are closed in a particular frame. This is a desirable property in light of the unpredictability of target patients' expressions. Figure 5 shows some examples of the eye region location method.

After the eye region has been properly located, the proposed methodology makes use of two different sources of information (as depicted in Figure 2): optical flow information and color information from the sclera. Each of these two branches will be addressed below.

3.3 Optical flow information

In order to compute the optical flow information several steps need to be taken. Each of them is represented in Figure 6, and further explained and justified in [13].

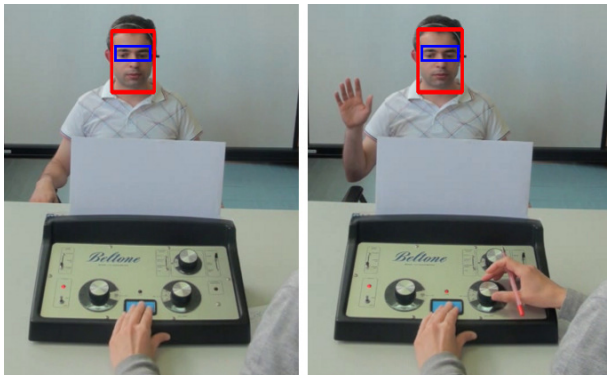


Fig. 5 Examples of eye region location in an audiometric test sequence.

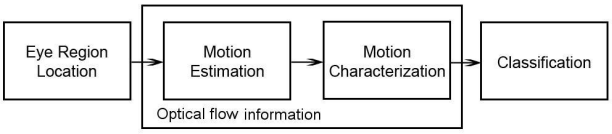


Fig. 6 Main steps of the optical flow information element.

3.3.1 Motion estimation

After the eye region has been located, the motion of the eye region is estimated by using the iterative [21] optical flow with a pyramidal approach [6]. Optical flow offers optimal results for the detection of movements produced by facial expression dynamics.

The recorded video sequences present a given frame rate. If it is high then matchings between one particular frame and the next may not offer significant changes as the evolution of expression in that minimal quantity of time is very small. In order to detect sufficiently notable changes in facial gesture, a time window (t) is considered between compared 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 avoiding unimportant movements and maintaining all the relevant ones. For our particular sequences, with a frame rate of 25 frames per second, t was empirically assigned a value of 3, which approximately corresponds to a time lapse of $0.125s$.

Computation of the optical flow is performed on interest points of the scene. An interest operator (in our particular case, Good Features to Track [29]) is applied over the first frame that serves as the image of reference. Then, their corresponding points are located on the frame $i+t$.

For a better visual representation of the motion field of the scene, motion vectors will be depicted as arrows, where the arrow illustrates the point displacement from one instant to the other. Furthermore, motion vectors are categorized attending to their magnitude and only strong movements will be considered for the next steps (since we are only interested in considering relevant movements). Figure 7 shows several samples where vectors are represented as arrows and non-significant vectors are removed.

(a) (b)

Fig. 7 Some sample movements detected with the optical flow.

3.3.2 Motion characterization

Once the motion has been detected and quantified in the scene, the next step is to characterize the type of motion taking place in it. In order to achieve a robust classification of the motion, its characterization is required based on a series of relevant properties that will be used to build a descriptor vector of the motion. To capture all the relevant properties of a particular movement in our domain a descriptor is proposed based on three main properties: orientation, magnitude and dispersion. The motion in each eye is considered separately so each scene will contain two descriptors, one per eye. As stated before, one of the main properties considered is the orientation of the movement. The orientation vector gives information about the direction of the movement produced within the eye region. This orientation is different in a case of change in gaze direction as opposed to a movement of eye closure or eye opening, as can be observed in Figure 7, illustrating a gaze shift to the right (a) and an opening of the eyes (b). In order to build the descriptors, vectors are divided into eight different equally distributed ranges according to their angle. This classification can be represented mathematically as 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 $* \in \{L, R\}$, 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 first 8 values of the descriptor correspond to the number of vectors in each range (see Equation (2)).

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

where n_i^*, represents the cardinality of the set of vectors that belong to each descriptor R_i^*.

It is also important to include the magnitude of the motion vector as it provides information about the intensity of the movement. Keeping this in mind, the subsequent eight values for the motion descriptor are computed based on the magnitude of the movement. With vectors grouped by ranges (the angle ranges previously defined), the average of the module of the vector is calculated according to Equation (3). This feature provides information about the intensity of the movement, making it possible to distinguish between strong and weak movements.

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

Finally, the dispersion of the optical flow vector contributes with the other eight values to the descriptor. The dispersion of the optical flow allows us to discriminate between local motion due to gaze changes and global motion of the whole frame due to head movement. The computation is considered by range, in other words according to the angle of the vectors. From each of the vectors ($v = \overrightarrow{AB}$) the destination point is taken $B = (B_x, B_y)$ and their center is calculated according to Equation (4) locating the centroid.

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

Once the centroid has been calculated, the dispersion is computed through the calculation of the average distance to that center, according to Equation (5).

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

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

Finally, the motion descriptor consists of a total of 24 computed values where 8 are related to orientation values, N^* (Equation (6)), 8 represent the histogram for magnitude, M^* (Equation (7)), and the final 8 include information on the dispersion of the motion vectors in the scene, D^* (Equation (8)):

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

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

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

Figure 8 shows an example of the motion descriptor.

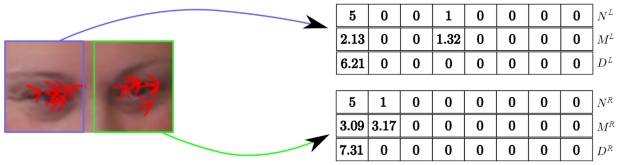


Fig. 8 Descriptors sample. The left-hand image shows the movement vectors for each eye and the tables on the right the corresponding descriptors for each eye. The first row represents orientation, the middle one magnitude and the last one dispersion.

As a result of this characterization we obtain a feature vector consisting of 24 features. This feature vector will then be combined with the feature vector provided by the analysis of the color information from the sclera.

3.4 Color information from the sclera

The main steps of this branch of the methodology are represented in Figure 9.

3.4.1 Pupil location

After the location of the eye region, this step is aimed at the location of the center of both pupils, to enable this information to be used as a reference point in the subsequent steps of the methodology. To that end, a method based on gradients [30] is applied. The yellow points in Figure 10 correspond to the pupil's locations provided as a result by the proposed method for some eye region samples from different patients.

A further two different methods aimed at locating the pupil's center were also considered: another gradient-based method [18] and the Starburst method [19]. However, the [30] approach was the one that showed the most accurate results for our domain in our test. In this test three different alternatives were compared for the pupil location task. The three methods for locating the pupil's center analyzed in this study are: method 1 (Starburst [19]) and methods 2 and 3 (both gradient-based, [18] and [30]).

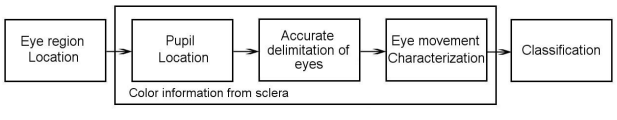


Fig. 9 Main steps of the color information from the sclera.

Fig. 10 The yellow points represent the center of the pupil obtained at this step.

The test set 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. In each of these frames the eye region of each eye was labeled, thus obtaining a total number of 40 samples for each video sequence. Considering that we have 10 video sequences, a total of 400 samples will then be evaluated, 200 for the right eye and 200 more for the left eye.

In order to conduct this experiment the expected pupils' centers were previously labeled so a comparison with the results obtained 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 location as indicated in Equation (9). Since both P_e and P_c are pixel locations, the obtained error is also measured in pixels.

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

Table 1 contains the average and the standard deviation (expressed in number of pixels) of the obtained errors for the complete dataset. Each row corresponds to one of the ten video sequences.

Generally speaking, the results provided by the three different alternatives are quite similar in terms of error. Since the results provided by the [30] approach are slightly better than the others, this is the method chosen for inclusion in our methodology.

Table 1 Individual and global results for the pupil location methods.

		Error Starburst [19]	Error Gradient [18]	Error Gradient [30]
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

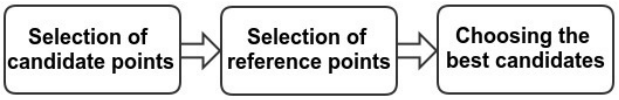


Fig. 11 Phases of the delimitation of the eyes stage.

3.4.2 Accurate delimitation of eyes

Using the eye region and the pupils' location provided by the previous steps for information, we designed a method that locates the eyes' corners in three steps (see Figure 11).

First, we detect points that can be considered as candidates to be the eyes' corners by the use of an interest operator. More particularly, after studying several interest operators, we have applied the Shi and Tomasi [29] method. Between these candidate points, it is necessary to choose those that better represent the eye's corners. Edge information is used at this step in order to obtain the edges associated to the limits of the eyelids, so they can be used as a reference for the eye's limits. Several optimizations are applied in order to increase the enhancement of the eyelid. First, we convert the images from the RGB color space to HSV color space, in order to use the saturation channel S. Next, an erosion filter considering the radius of the iris is applied. The image obtained as a result of applying the erosion filter $S_f(x, y)$ is subtracted from the saturation images $S(x, y)$), thereby obtaining the subtraction images $R(x, y)$ (as indicated in Equation (10)). Next, a threshold for the binarization of the image is computed using some features of the image according to Equation (11) as a reference, where μ is the average value of the pixels from the difference image I_{diff} and σ is the standard deviation. The binarization is computed according to Equation (12) where I_{ths} is the thresholded image.

Considering the anthropometric constraints presented by the human eye, we can define the eyes' corners as the intersections between the ellipses that represent the eyelid; these intersection points are considered as the

reference points. In the third step, the reference points are validated according to anthropometric references related to the average size of the eye. Once the validity of the reference points has been checked, this information is used to compute the distances between the candidate points and the associated reference points, finally choosing the candidate point nearest to the reference point. The different points obtained throughout these steps are represented in Figure 12.

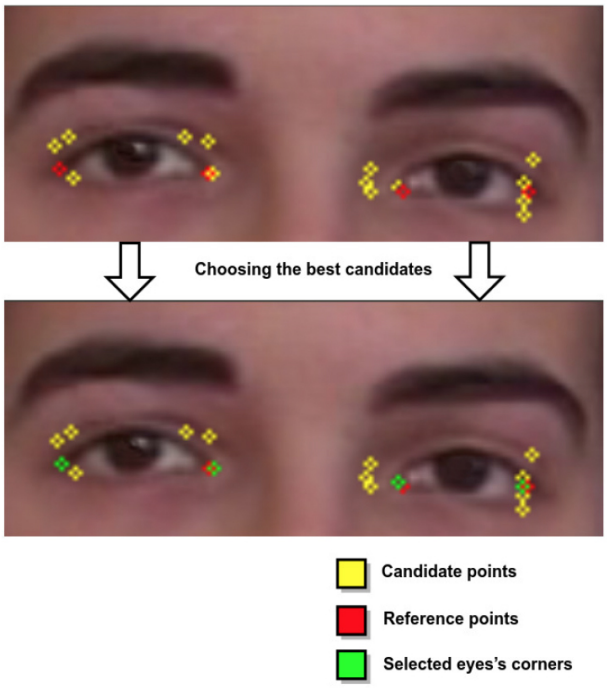


Fig. 12 Choosing the best candidates: yellow for candidate points, red for reference points and green for the selected eyes' corners.

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

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

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

3.4.3 Eye movement characterization

Eye movement is going to be characterized using color information from the sclera. To that end, it is necessary to estimate the amount of white in the eye, using the characteristic points previously obtained as a reference. First, the input image is converted to gray scale and a histogram equalization is applied over it. Then, a gray level distribution is computed representing the gray level for each of the pixels located in the line connecting both eyes' corners. This step can be observed in Figure 13.

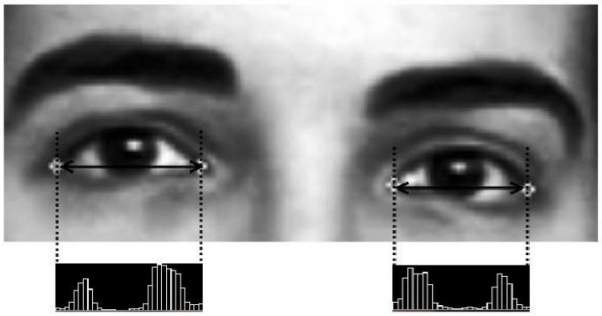


Fig. 13 Sample of the gray level distribution.

Then, the gray level distribution is divided into three areas of interest: the iris, the left side of the sclera and the right side of the sclera. The distribution of the delimitation of these three areas is shown in Figure 14. To achieve this delimitation we use information about the pupil's center and an estimation of the radius of the iris.

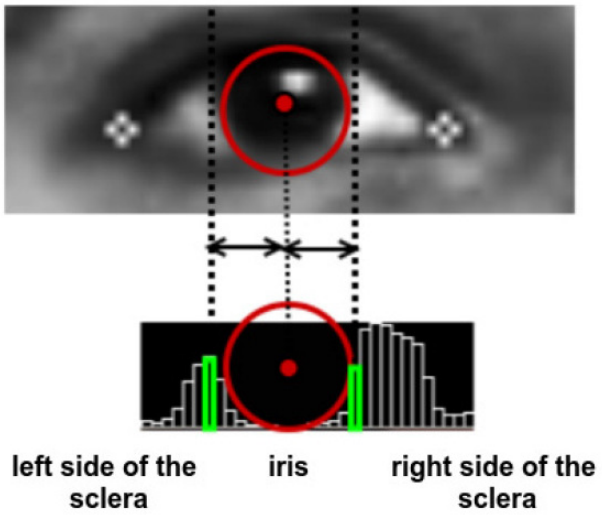


Fig. 14 Delimitation of the three areas of interest over the gray level distribution.

From this information, a movement descriptor composed of 66 values is defined; 33 values for the current state and 33 more values considering previous information. First, the gray level distribution is normalized to

30 values, so each of these 30 new 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 of the sides is computed. These two values, along with their sum, are added as 3 additional values to the movement descriptor. Furthermore, since we are also interested in the movement with respect to previous frames, we compare the current descriptor with a previous descriptor. With the intention of allowing sufficiently notable expression changes, 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 at 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, thereby adding 33 new values to the final descriptor. Figure 15 shows an example. In the first row, the first 30 values correspond to the normalized gray level distribution, L represents the summation of the gray levels in the left side of the sclera, R corresponds to the summation of the gray levels on the right-hand side of the sclera and T is the summation of both sides. In the second row, the first 30 values correspond to the subtraction of the normalized gray level distribution of frame i and frame i+t (Equation (13)), DL represents the subtraction of the summation of gray levels in the left side of the sclera (Equation (14)), DR corresponds to the subtraction of the summation of gray levels in the right side of the sclera (Equation (15)), and DT is the subtraction of the summations of both sides (Equation (16)).

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

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

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

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

Once the vector descriptors have been computed for each movement, the next step is to combine them with the optical flow information and classify them according to the movement categories determined for this domain.

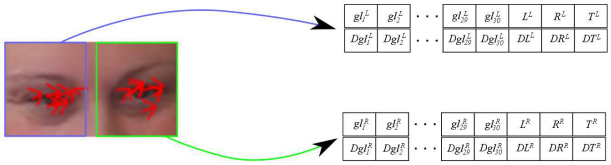


Fig. 15 Color descriptors sample. The left-hand image shows the movement vectors for each eye and the tables on the right the corresponding color descriptors for each eye. The first row represents the normalized gray level distribution. The second row represents the subtraction of the normalized gray level distribution of frame i and frame i+t.

3.5 Final classification

Combining the descriptors that are obtained from both branches of the methodology, the last step of the methodology is their classification into the considered movement categories. An experiment on classification was carried out in order to find the most suitable classifier. Supervised training was 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 section.

4 Experimental results

To perform this experiment, different video sequences were analyzed and the eye movements produced during them were manually labeled. The video sequences considered for the experiment had full HD resolution (1080 x 1920 pixels), and 25 frames per second (FPS). Images were recorded using a conventional video camera with Full HD resolution and no other particular hardware requirements. The only requirement was to try to maintain favorable and constant lighting conditions.

Our dataset consists of a total of 11 video sequences corresponding to male and female patients of different ages and cognitive statuses. Each of the video sequences takes between 4 and 8 minutes, and each 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 audiologists. The test video sequences were analyzed frame by frame, and each of the relevant movements was labeled into the corresponding category. A total of 1180 descriptors were classified as significant movements and allotted to the corresponding

category, giving the descriptors distribution shown in Table 2.

Table 2 Distribution of significant movements between the considered categories.

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

As can be observed in Table 2, there is a certain degree of imbalance between the categories gaze shift to the right (GR) and gaze shift to the left (GL). With all the 1180 labeled descriptors, 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 3, where a comparison between applying only the optical flow descriptors vs. the optical flow and color from the sclera descriptors is presented.

As can be observed from these results, classification accuracy is improved in all cases with the introduction of the complementary information from the color distribution of the sclera. The worst results were obtained with the C4.5 and Random Tree classifiers, in both cases with an accuracy rate of under 90%. The rest of the classifiers achieved classification accuracy rates of over 90%. The best results were obtained by the Support Vector Machines, with an accuracy rate of 97.46%, which is a highly accurate result for this domain. For a more in-depth study, Table 4 presents the true positive rate by classes for the combined approach.

Table 3 Classification accuracy comparison: left-hand column for optical flow descriptors and right-hand column for optical flow & color from the sclera descriptors.

	% Accuracy	
Method	Optical flow	Optical flow & 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%**

It is easy to note that these algorithms are often biased towards learning the majority class, leading to higher misclassification rates for the minority class instances (gaze left and gaze right). In order to deal with this imbalance we tested oversampling methods, aimed at obtaining a more balanced distribution among classes by increasing the number of samples in the smaller class. In particular, the SMOTE algorithm (Synthetic Minority Oversampling TEchnique) [7] [4] is an oversampling approach that inserts synthetic samples from the minority class into the initial dataset. This is done until all classes are equally sized. To achieve this goal, this technique builds the new synthetic samples based on the class existing ones: it finds the k nearest neighbors of the example in the minority class that will serve as the base for the new synthetic ones. Once this has been obtained, the synthetic sample is computed as an intermediate point between the original sample and one of its neighbors.

Initially, we applied an oversampling rate of 100% on the classes gaze left (233 samples) and gaze right (239 samples). In the second row of Table 5 we can see the results of this experiment, in which the true positive rates for the two classes that are the most important ones in this problem (gaze left and gaze right) have increased, while maintaining global accuracy. In this work, we performed several experiments, increasing the oversampling rate to 200%, 300% and 400%, as can be seen in the remaining rows of Table 5. The used parameter values were based on the works [22] and [28]. For the sake of comparison, the first row shows the classification results when no oversampling technique is applied.

The best results for the gaze shifts were obtained when the level of oversampling was 400%. In this case, we achieve maximum true positive rates for the detection of the most important classes (gaze left and gaze right) whilst global accuracy has been reduced by less than 0.3%. However, with 300% oversampling, we can maintain the global accuracy with a noticeable increase in the detection of gaze shifts. Note that with these configurations TP rates for the detection of closure and opening decrease, but nevertheless remain within the range of acceptable values considering their lower incidence in the detection of gestural reactions.

4.1 Detection of reactions to the sound

As commented before, the significant contribution that we can offer audiologists is the proper detection of eye movements associated with reactions to 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 being the reason why we are not overly concerned about obtaining a high success rate in

Table 4 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

Table 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**

classification, because a typical reaction lasts between 5 and 15 frames and therefore the misclassification 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. An annotation procedure by an expert audiologist in three video sequences was performed in order to determine the baseline for reactions to sound stimuli. Stimuli reaction recordings were equally distributed between three population segments: young people (under 30 years old) with no disorders, middle-aged people (around 45 years old) with mild signs of disorder and older people (over 65 years old) with cognitive disorder. The results are detailed in Table 6, where we evaluate the extent of agreement between the methodology and the audiologists based on the number of reactions to the stimuli detected by each of them. 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, and despite the existing classification error of eye movements, the detection of eye gestural reactions is optimal (100% 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 the proper detection of the reaction.

Table 6 Evaluation in the detection of reactions to the sound. Results are expressed in number of reactions.

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

5 Discussion and conclusions

This approach presents a methodology for the detection and classification of eye gestural reactions as a response to auditory stimuli during the performance of audiometric evaluations. This methodology will allow automatic detection of positive unconscious gestural reactions which can be interpreted as positive responses in the hearing assessment of patients with cognitive decline or other severe communication difficulties. The development of an automated tool will facilitate the detection of these specific reactions, avoid subjectivity and make hearing assessment less error prone.

One of the premises of this work is to modify the traditional protocol of audiometric assessment as little as possible. When working with this particular group of patients, it is extremely 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.

One major issue when dealing with this kind of patient is the variety of spontaneous responses that each

particular individual offers. Although a reaction related to the eye region is to be expected, the particular nature of this reaction cannot be predicted given the heterogeneity of patients' conditions. Bearing this in mind, in this novel proposal the detection of eye-based gestural movements is achieved through a combination of optical flow information and analysis of the color distribution of the sclera, effectively fusing global and localized motion information for the eye region. The most accurate results were obtained with the SVM classifier, with a classification accuracy of 97.46%, which is a highly accurate result for this domain. Furthermore, an oversampling technique (in particular, SMOTE) was applied in order to improve the accuracy of the classification of gaze shift movements. This oversampling approach is particularly appropriate for problems where minority classes are of special interest, as is our case given the range of eye gestures for our target patients. Moreover, given the low dimensionality of our data, we work in a range where the SMOTE performs adequately. With oversampling rates of 300% and 400% the true positive rates for the gaze shift categories are increased with no noticeable consequences for the remaining categories. Considering this, the proposed methodology has shown encouraging and positive results, paving the way to its inclusion in a fully automated tool.

Regarding the computational cost, the only global analysis that is performed using the entire image is in the face detection stage. This stage uses the Viola and Jones strategy that is relatively simple and efficient. Subsequently, the method performs the analysis only in the face region, firstly, and in the eye region, finally. This region represents a significantly small portion of the entire image, for what the rest of proposed methodology imply an insignificant computational cost in addition to the face detection stage.

In clinical terms, the manual analysis conducted by audiologists can be automated with the main benefit of being unaffected by subjective factors when evaluating patients with cognitive decline or severe communication difficulties. As well as producing unbiased results, the automatic proposal also saves time for the experts and provides a detailed identification of eye-based gestural reactions. In this sense, the audiologist can obtain a more objective detection of the patient's unconscious positive reactions to sound, greatly assisting the evaluation of 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 groups of patients.

With an eye to the future, it would be necessary to obtain more video sequences in order to provide a more comprehensive validation of the methodology. In addition, eye gestural movements should be correlated with the information about auditory stimuli delivery so they can be analyzed as gestural reactions to sound. We could also consider the additional study of the audiometric evaluations in less controlled environments combined with the analysis of different face detection methods. Finally, this methodology could be adapted to other related tests as well as other procedures for general purposes.

Acknowledgements This work is supported by the Instituto de Salud Carlos III, Government of Spain and FEDER funds of the European Union through the PI14/02161 and the DTS15/00153 research projects. 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. Acton, Q.: Dementia: New Insights for the Healthcare Professional: 2013 Edition. ScholarlyEditions (2013). URL http://books.google.es/books?id=pkE0Yv1MfOQC
2. of Audiology, B.S.: Recommended Procedure Pure-tone air-conduction and bone-conduction threshold audiometry with and without masking (2015). URL http://www.thebsa.org.uk/resources/pure-tone-air-bone-conduction-threshold-audiometry-without-masking/
3. Australian Hearing Annual Report: (2009). URL http://www.hearing.com.au/australian-hearing-annual-reports
4. Bolón-Canedo, V., Fernández, A., Alonso, A., Ortega, M., Penedo, M.G.: On the use of machine learning techniques for the analysis of spontaneous reactions in automated hearing assessment. In: European Symposium on Artificial Neural Networks, pp. 355–360 (2015)
5. Boraston, Z., Blakemore, S.J.: The application of eye-tracking technology in the study of autism. The Journal of physiology **581**(3), 893–898 (2007)
6. Bouguet, J.Y.: Pyramidal implementation of the Lucas-Kanade feature tracker: Description of the algorithm. Intel Corporation, Microprocessor Research Labs (2000)
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling TEchnique. Journal of Artificial Intelligence Research. **16**(1), 321–357 (2002)
8. 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
9. Davis, A.: The prevalence of hearing impairment and reported hearing disability among adults in great britain. Int J Epidemiol. **18**, 911–917 (1989)
10. Davis, A.: Prevalence of hearing impairment. In: Hearing in adults, chapter 3, pp. 46–45. London: Whurr Ltd (1995)

11. De Santi, L., Lanzafame, P., Spanò, B., D'Aleo, G., Bramanti, A., Bramanti, P., Marino, S.: Pursuit ocular movements in multiple sclerosis: a video-based eye-tracking study. Neurological Sciences **32**(1), 67–71 (2011). DOI 10.1007/s10072-010-0395-1. URL http://dx.doi.org/10.1007/s10072-010-0395-1
12. Fernández, A., de Moura, J., Ortega, M., Penedo, M.G.: 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, pp. 313–320 (2015)
13. Fernández, A., Ortega, M., Gonzalez Penedo, M., Vazquez, C., Gigirey, L.: A methodology for the analysis of spontaneous reactions in automated hearing assessment. Biomedical and Health Informatics, IEEE Journal of **20**(1), 376–387 (2016). DOI 10.1109/JBHI.2014.2360061
14. Fukushima, K., Fukushima, J., Barnes, G.R.: Clinical application of eye movement tasks as an aid to understanding parkinsons disease pathophysiology. Experimental brain research **235**(5), 1309–1321 (2017)
15. IMSERSO: Las personas mayores en España. In: Instituto de Mayores y Servicios Sociales (2008)
16. IMSERSO: Libro blanco del envejecimiento activo (in Spanish) (2010)
17. Jain, V., Learned-Miller, E.: Fddb: A benchmark for face detection in unconstrained settings. University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009 **2**(7), 8 (2010)
18. Kothari, R., Mitchell, J.: Detection of eye locations in unconstrained visual images. In: Image Processing, 1996. Proceedings., International Conference on, vol. 3, pp. 519–522 (1996)
19. Li, D., Winfield, D., Parkhurst, D.: 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, pp. 79–79 (2005)
20. Lin, F.R.: Hearing loss and cognition among older adults in the united states. In: The Journals of Gerontology: Series A (2011)
21. Lucas, B.D., Kanade, T.: 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, pp. 674–679. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1981)
22. Lusa, L., et al.: Smote for high-dimensional class-imbalanced data. BMC bioinformatics **14**(1), 106 (2013)
23. Marandi, R.Z., Sabzpoushan, S.H.: Using eye movement analysis to study auditory effects on visual memory recall. Basic Clin Neurosci **5**(1), 55–65 (2014)
24. Murlow, C., Aguilar, C., Endicott, J., Velez, R., Tuley, M., Charlip, W., Hill, J.: Asociation between hearing impairment and the quality of life of elderly individuals. pp. 45–50. J Am Geriatr Soc (1990)
25. National Institute of Deafness and Other Communication Disorders: Quick statistics (2014). URL https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing
26. Pereira, M.L., Camargo, M.v., Aprahamian, I., Forlenza, O.V.: Eye movement analysis and cognitive processing: detecting indicators of conversion to Alzheimer's disease. Neuropsychiatr Dis Treat **10**, 1273–1285 (2014)
27. Raney, G.E., Campbell, S.J., Bovee, J.C.: Using eye movements to evaluate the cognitive processes involved in text comprehension. J Vis Exp (83), e50,780 (2014)
28. del Río, S., López, V., Benítez, J.M., Herrera, F.: On the use of MapReduce for imbalanced big data using Random Forest. Information Sciences **285**, 112 – 137 (2014)
29. Shi, J., Tomasi, C.: Good features to track. In: Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on, pp. 593–600 (1994). DOI 10.1109/CVPR.1994.323794
30. Timm, F., Barth, E.: Accurate eye centre localisation by means of gradients. In: L. Mestetskiy, J. Braz (eds.) VISAPP, pp. 125–130. SciTePress (2011)
31. Viola, P., Jones, M.: Robust real-time object detection. In: International Journal of Computer Vision (2001)