SLOAS: Hearing with the Eyes

J. TOLEDO, J. TORRES, S. ALONSO, P. TOLEDO, E. J. GONZÁLEZ
Department of Fundamental and Experimental Physics, Electronics and Systems
University of La Laguna
Avda. Francisco Sánchez s/n. La Laguna, CP:38204. Tenerife (Islas Canarias)
SPAIN

Abstract: - In this paper a sensorial substitution system (SLOAS) is presented with the objective of helping deaf people. The system who is made of a microphoned glasses and a special kind of visualization system captures the environment acoustic information and encode it into visual elements, in order to have a graphic representation showed to the user though an eyesglasses-mounted display. All the process is made in real-time. The eyesglasses-mounted display has a see-through optics that show in a graphic way the information provided by the system but this information doesn't hide the normal vision space of the user. In the body of the glasses a small microphones are placed. These microphones capture the sound information that the deaf user can't hear. The signal of the microphones is the input of a system who calculates the main parameters of the sound wave. The parameters are the special localization of each sound source, the power of the sound and their spectral composition. This acustic information is converted in a visual information. This system is being tested by five deaf users, and now their opinion is being analyzed.

Key-words: - Deafness, eyeglasses, sensorial substitution.

1. Introduction

Deafness is a kind of sensorial disability. It hampers the relationship between the deaf person and the acoustic elements of his environment. For this reason, many deaf people have difficulties to learn.

One of the main problems of the deafness is the real impossibility to spatially locate an acoustic source, because it puts the deaf in risk. For example, a deaf moving along a public way cannot detect any potential risk (like a car or a bus) that gets it from behind. On the other hand, a deaf cannot hear sounds that warn or give information about the nearby environment. That is, a phone bell, a doorbell, a horn, etc.

The solution to deafness depends on the kept capacity to hear. If that capacity exists then the deaf can use audiphones, otherwise it can have some surgical treatment like the cochlear implant. With this last treatment, the deaf can hear some sound but cannot understand a spoken conversation.

The solution presented in this paper is an electronic device that stimulated by sound waves, finds the spatial localization and spectral analysis of acoustic sources on the environment. The deaf gets all this information through the sense of sight. Therefore it is a sensorial substitution device between the sense of hearing and sight.

2. Device Description

The device has two main components. First some eyeglasses, that have several (two or three) omni directional microphones and a portable eyeglass display. The microphones are located around the head to acquire the audio signals from the environment . As we can see in Figure 1, those portable eyeglass displays can take the output signal from ordinary electronic devices and project images into the glasses. It's like having your monitor right in from of you all the time, anywhere you want.

Figure 1. Portable eyeglass display

The see-through optic of the portable eyeglass display allows the user to see all the information that the

device can show without obstructing their normal eyesight. This technology has been developed by [1].

At present it has been developed a portable prototype fed by batteries (Figure 2) that communicates by means of a radio system with the central computer (architecture Intel x86) equipped with a standard sound card for the capture of the sound sources. Two radio connections exist between the computer and the eyeglasses. The first one permits to establish a communication among the microphones installed in the eyeglasses and the sound card, thus the computer receives the sound to process since the prototype. The second between the graphic output of the computer and the system of viewing (eyeglasses) of the prototype, permitting to show the processed sound and become visual in the device. Both connections are carried out by means of radio links, for which the present rank of action is limited to some hundreds of meters. On the other hand, using personal computers accelerates the development and the test of the different prototypes. This makes easy the experimentation of the users in their own home. Only some minutes are needed to prepare all the system in a specific place. The software in the device has two

Figure 2. Portable prototype

main subsystems: 1) The signal processing subsystem that makes the necessary tasks to spatially locate and spectrally analyze every sound source. 2) The graphic subsystem. It takes the information gave by the signal processing subsystem and encodes it into visual elements to generate the graphic output. Then the users can see a real time graphic representation of the acoustic environment.

2.1 Signal processing subsystem

The main task of the subsystem of the signal processing subsystem is to locate the sources of sound, reconstructing original sources and quantifying the power of each source in the different bands of frequencies. This subsystem takes the sounds mixed from the microphones mounted in the eyeglasses and should separate the most important sources.

2.1.1 Localization and separation method

The device can use two or three microphones. By using two microphones, the device can only calculate the azimuth of a sound source between –90 and 90 degree. That is, it cannot remove the front-back confusion discretion. With more than two microphones, the device can calculate the azimuth of a sound source over the whole azimuth region. Figure 3 shows the position of three microphones and the whole azimuth region.

The blind source separation method used by the signal processing subsystem is presented in [2]. The method assume the signal are W-disjoint orthogonal. That is, the method applies when the support of the windowed Fourier transformations of any two signals in the mixture are disjoints sets. This assumption allows the partition of the time-frequency representation of the mixtures to recover the original sources and to estimate the spatial localization of every source. The technique is valid even in the case when the number

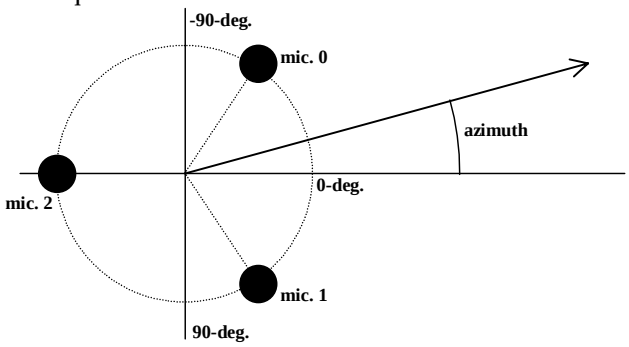

Figure 3. Three microphones over the azimuth region

of sources is larger than the number of microphones.

As we can see in [2], we can write the model for the i^{th} and j^{th} mixed signals as,

$$\begin{bmatrix} X_i(\omega,t) \\ X_j(\omega,t) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 e^{-i\omega\delta_1} & \cdots & a_N e^{-i\omega\delta_N} \end{bmatrix} \begin{bmatrix} S_1(\omega,t) \\ \vdots \\ S_N(\omega,t) \end{bmatrix}, \quad (1)$$

where $X_i(\omega,t)$ is the Fourier transformation of the mixture $x_i(t)$ and $S_i(\omega,t)$ is the Fourier transformation of the source signal $s_i(t)$, in the time interval between $t-\tau$ and t. The amplitude a_k and the delay δ_k are the mixing parameters of the k^{th} source in the mixed signal $x_j(t)$ respect to $x_i(t)$.

For W-disjoint orthogonal source most one of the N sources will be non-zero for a given (ω,t), that is,

$$\begin{bmatrix} X_i(\omega,t) \\ X_j(\omega,t) \end{bmatrix} = \begin{bmatrix} 1 \\ a_k e^{-i\omega\delta_k} \end{bmatrix} S_k(\omega,t), \quad (2)$$

Keeping in mind the Equation (2), it is clear that the mixing parameters can be obtained as,

$$(a,\delta) = \left(\left\| \frac{X_j(\omega,\tau)}{X_i(\omega,\tau)} \right\|, \mathrm{Im}\left(\log\left(\frac{X_i(\omega,\tau)}{X_j(\omega,\tau)} \right) \right) / \omega \right). \quad (3)$$

The signal processing subsystem can use Equation (3) to calculate amplitude-delay estimates from a number of (ω,t) points on each mixture pair. Since a time-frequency point maps to one or more points in the azimuth region, it calculates one or more azimuth candidates for each (ω,t) point. The subsystem uses some basic restrictions between the time-frecuency points to get valid azimuth candidates.

The signal processing subsystem makes an azimuth histogram for each band in each mixture pair, where the histogram maximums found are the azimuth estimates associated with each source. It partitions the time-frequency plane using one of the mixtures and the azimuth estimates. Thus, for each time-frequency point, it can determine which of the M peaks on the azimuth histogram is the closest.

2.1.2 Echo rejection enhancements

The signal processing subsystem has an enhanced version of [2] to get a better echo rejection. These enhancements are based in the precedence effect model that we can see in [3].

The precedence effect shows that the human auditory system can detect the beginning of a sound and inhibit the subsequent reverberation portion . It is the adaptation of the human auditory system to reverberant environment.

The precedence effect model assumes that the reflections of an impulse sound will delay a certain time more than, the direct sound this will depends on the distance from the sound source to walls, comparing to. It also assumes that the strength of reflections will decrease exponentially over time. The model estimates echoes as the maximum effects of their previous sound by the amplitude pattern of typical impulse response, regardless of the type of sound and the environment conditions .

The model of precedence assumes that the reflections of an impulsive sound are delayed a certain time before arriving at the ears (microphones in the system described). The retard depends on the distance of the source of sound to the walls, compared with the direct sound of the source. The model of precedence also

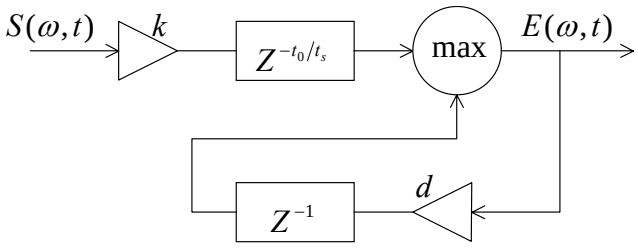

Figure 4. Echo estimation algorithm

assumes that the amplitude of the reflections will be reduced exponentially through the time. The model estimates the echoes as the maximum power of a previous sound multiplied by the main amplitude of an standard impulse, without keep in mind the type of sound and the class of environment.

Using the exponential decay as the typical impulse response the echo estimation algorithm can be implemented as shown in Figure 4, where t_s is the sampling time interval and,

$$d = e^{-t_S/\tau}. \quad (4)$$

In Figure 4, the attenuation factor k, the delay time t_0 and decay factor τ must be chosen to match the most general cases in an ordinary environment.

The signal processing subsystem mix every input channel and estimates the echo for the mixture. For

each (ω,t) point, it calculates a ratio between the mixture and the estimated echo. This ratio is used to weight the azimuth candidates of the (ω,t) point in the histogram.

In Figure 5, we can see a block diagram of the sound localization method with the echo rejection enhancements.

2.1.3 Real-Time method

The signal processing subsystem uses a version in real time of [4]. The algorithm used is presented in [3]. It is a kind of gradient descent method on the maximum probability to follow the parameters of the mixture, the method is based on the execution of a first calculation of the parameters. The objective is to enlarge the velocity of the algorithm to be able to carry out the calculations in real time. As it has just mentioned in the first place the general algorithm is utilized to calculate the sound sources for the first time. The result is used as starting point for the calculation. Thus continuing the equation (4) that minimizes the function of cost (5), applies the algorithm in real time during a period of time predetermined (in which is supposed that there is not new sonorous sources). After passing a certain interval of time the main algorithm is applied again so a new number of separated sources are calculated. The proposed strategy permits a commitment among the velocity of the algorithm in real time and the adaptation of the variation of the number of sources in the general algorithm.

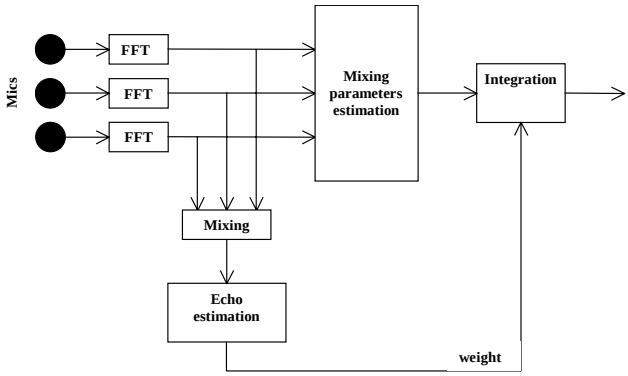

Figure 5. Sound localization method with rejection enhancements

$$a_j[k] = a_j[k-1] - \beta\alpha_j[k]\frac{\partial J(\tau_k)}{\partial a_j}$$
$$\delta_j[k] = \delta_j[k-1] - \beta\alpha_j[k]\frac{\partial J(\tau_k)}{\partial \delta_j} \quad (4)$$

$$J(\tau) = \min_{a_1\delta_1,\ldots,a_N\delta_N} \sum_w -\frac{1}{\lambda}\ln\left(e^{-\lambda p_1}+\ldots+e^{-\lambda p_N}\right). \quad (5)$$

2.2 Graphic output subsystem

The location subsystem provides to the graphic subsystem, the spatial position and the frequency analysis of the sonorous sources previously located. This information is utilized for the device to obtain a visual representation of the sound or present sounds in every single moment. The device has different ways of representation, each one of them is adjusted to the different environments in which the user can be found: in the public way, at home. ... In some cases these representations can be combined to be adjusted in the best way to the needs of the users.

The form of the representation we choose (Figure 6) part of the supposition that the screen placed on the eyeglasses is represented the environment in which the user moves. The sound sources are represented like graphic images and the angle with which are represented respect to the centre indicates the azimuth angle of origin of the sound. The chosen form for the representation of the sonorous sources are bells of Gauss. This mathematical function permits us easily indicate the angle of sound source origin since resembles an arrow indicating the direction of the sonorous source, besides allow us to hide the smaller

Figure 6. Representation seen by the user

visual space to the user. The power of the sound corresponds with the size of the Gaussian, the more amplitude have the sound, greater will be the area of the Gaussian and will attract the attention of the user. The components in frequencies of the sound sources are represented by means of bands of colours that divide the Gaussian. Thus the cold colours in the base of the Gaussian represent the low frequencies (serious) and the hot in the upper part, they represent sharper frequencies. Moreover the ordering of these bands is always the same one, so the locating of a sonorous source and its changes in frequency can be seen in real time. This way to represent the graphic environment allow us to represent two important aspects.: On one hand the simplicity of understanding, the form to represent the sounds is quite intuitive and easily understandable by the user, by another a wealth of information that can be very useful after the necessary period of learning of the user.

In the example of the Figure 6 a normal situation in a house is represented. It can be verified that there is different sonorous sources emitting simultaneously, each one with different spectral components and with different positions (the Gaussians have different colours, positions and sizes). The user can verify easily which is the origin of each one of the sonorous sources. This is a capture in a given moment, the parameters will be changing through the time.

The representations can be combined with more complex elements. The system can detect critic sound sources like sirens or bells, and inform through a warning system. The detection of danger sources is based on the high power sounds and concrete frequencies that are used to coinciding with dangerous circumstances in the environment of the user.

It is also possible to combine the visual representation with other system of tactile stimulation. In this case part of the information obtained by the user will be assimilated through the sense of touch.

3. TESTS

The device is being tested on a sample of five deaf subjects. The main task is to evaluate the effectiveness of the graphic interface, as this is a subjective matter. Several points must be taken into account for this purpose. The interface must avoid a distracting effect and integrate correctly with subject eyesight. The amount of information presented must be graduated to be informative without being a burden. The subject experiences will be used to resolve these issues. The subject responses in the preliminary tests are promising.

4. CONCLUSIONS

A sensorial substitution device has been presented. The device has two main components, an eyeglasses with microphones that grasp the sounds of the environment and a viewfinder that shows the information of the sound by means of intuitive images on the other hand a desktop computer takes charge of the calculation of the acoustics signs and a graphic generation. Computer and eyeglasses are communicated using a radio link which permit at the moment the utilization of the system in environments as the home, the office… In the following version intends to manufacture a completely portable device of such form that can be utilized in any environment. The system has been tested by an set of users to evaluate its effectiveness and to adjust its efficiency.

Acknowledge

To the Science and Technology Spain Ministry by its financing in the development of this Project through the PROFIT program.

References

[1] The MicroOptical Corporation http://www.microopticalcorp.com/

[2] Alexander Jourjine, Scott Rickard and Özgür Yilmaz in Proceedings of the 2000 *IEEE Conference on Acoustics, Speech, and Signal Processing* (ICASSP2000), Volume 5, Pages 2985-2988 (Istanbul, Turkey, June 2000)

[3] Jie Huang, Noboru Ohnishi and Noboru Sugie, Modeling the precedence effect for sound localization in reverberant environment" in Proceedings. *IEEE Instrum. Meas. Technol. Conf.* (IMTC'96), pp.633-636, (Brussels, June 1996).

[4] Scott Rickard, Radu Balan and Justinian Rosca. Real-time time-frecuency based blind source separation in Proceedings of *ICA2001 Conference* (San Diego CA, December 2001).