Journal of New Music Research
2001, Vol. 30, No. 3, pp. 271–277

0929-8215/01/3003-271$16.00

3-D Audio with Video Tracking for Multimedia Environments

José Javier López and Alberto González

Universidad Politécnica de Valencia, Departamento de Comunicaciones, Grao de Gandia, Spain

Abstract

This paper deals with a 3-D audio system that has been developed for desktop multimedia environments. The system has the ability to place virtual sources at arbitrary azimuths and elevations around the listener's head. It is based on HRTF binaural synthesis. A listener seated in front of a computer and two loudspeakers placed at each side of the monitor have been considered. Transaural reproduction using loudspeakers has been used to render the sound field to listener ears. Furthermore the system can cope with slight movements of the listener head. Head position is monitored by means of a simple computer vision algorithm. Four head position coordinates (x,y,z,ϕ) are continuously estimated in order to allow free movements of the listener. Cross-talk cancellation filters and virtual source locations are updated depending on these head coordinates.

1. Introduction

At the present time a strong tendency exists to increase the realism in the sound and music reproduction systems for space sensation and simulation of acoustic environments.

The multichannel sound systems, well established in the industry of the cinema, try to re-create this type of acoustic sensations. From the first systems of rendering sound until the current 5.1, 6.1 and 7.1 systems, a great evolution of signal processing techniques has taken place. These signal processing techniques have been used mainly in the digital compression of the multichannel sound, giving place to different standards, which became property of different companies. However, these systems, although they are suitable for the cinema do not provide a true periphonic space sensation appropriate for the accurate reproduction of music, and a good localization of the instruments in the sound space.

On the other hand, the systems based on the HRTF, binaural systems and transaural systems using cross-talk cancellation filters, have also experienced a considerable evolution. These systems have the advantage that only two channels are required for their transmission and recording, making them very appropriated for the music distribution.

The evolution of multimedia technologies together with the increasing computational power of the personal computers allow the implementation of more advanced applications in the field of sound reproduction. Incorporation of 3-D sound and digital image processing to the computer represent a real fact due not only to the mentioned increasing power, but also to the reduction of the prices of associated peripherals.

A typical multimedia environment is usually composed of a personal computer with central unit and monitor, and a stereo sound system using a pair of loudspeakers placed at each side of the monitor. Over the screen, a video camera looking at the user is placed. This camera is commonly used for videoconference purposes over telephone lines or Internet. Figure 1 shows all the aforementioned elements. A similar situation of a semi-static listener is produced when people listen to the music seated on a chair or sofa (typical hi-fi environment).

Using this equipment, normally found in most of the computers installed nowadays, a 3-D audio system through loudspeakers has been developed for desktop multimedia environments. Moreover, the system is robust to slight movements of the listener head.

In the following sections, after a review of the binaural and transaural systems, the implementation of the system will be explained. The technology used in this system is not new, and the performance is far from ideal, but this system innovates in that it is implemented in real time over a personal computer.

Accepted: 5 June, 2001

Correspondence: Universidad Politécnica de Valencia, Departamento de Comunicaciones, Ctra. Nazaret-Oliva s/n, 46730 Grao de Gandia, Spain. E-mail: jjlopez@dcom.upv.es

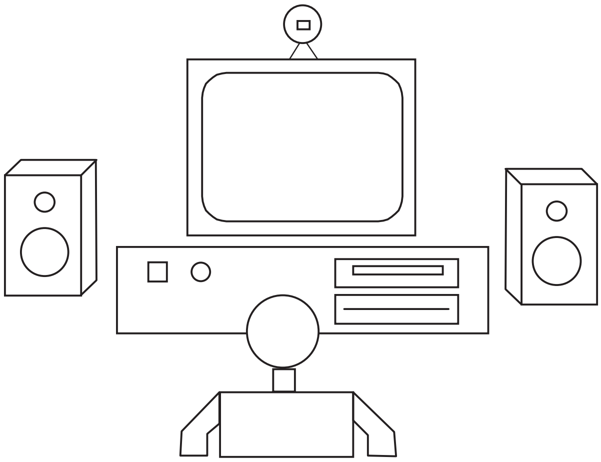


Fig. 1. Common multimedia environment.

2. Binaural synthesis

Humans have the ability to localize sounds in three dimensions. Although perceptual mysteries remain, the major 3D-perceptual cues are already known. We can consider a sound source located to the right side of a listener: sounds from the source will arrive first at the right ear, and they will reach subsequently the left ear. Moreover, the level of sound at the left ear will be attenuated, mainly at high frequencies, due to the head shadowing effect.

The Duplex Theory asserts that the first effect called ITD (Interaural Time Difference) and the second one called IID (Interaural Intensity Difference) are complementary. At low frequencies (below about 1.5 kHz), there is little IID information, but the ITD shifts the waveform a fraction of a cycle, which is easily detected. At high frequencies (above about 1.5 kHz), there is ambiguity in the ITD, since there are several cycles of shift, but the IID resolves this directional ambiguity.

In addition, the auditory system can also determine whether sounds are in front of or behind the listener, and can estimate the elevation of sound sources. This is possible because the incident waves interact with torso, head and pinna prior to arriving at the inner ear. These interactions produce reflections and diffractions that cause a spectral modification of the sound, which depends on the angle of the incidence. It can be guessed that our outer ear or pinna acts like an acoustic antenna. Its resonant cavities amplify some frequencies, and its geometry leads to interference effects that attenuate other frequencies. Thus, its frequency response is directionally dependent.

In order to find the sound pressure that an arbitrary source produces at the eardrum, the impulse response $h(t)$ from the source to the eardrum is needed. This is called the Head-Related Impulse Response (HRIR), and its Fourier transform $H(f)$ is called the Head Related Transfer Function (HRTF). The HRTF captures all of the physical cues to source localization. Once the HRTF for the left ear and the right ear is known, it is possible to synthesize accurate binaural signals from a monaural source.

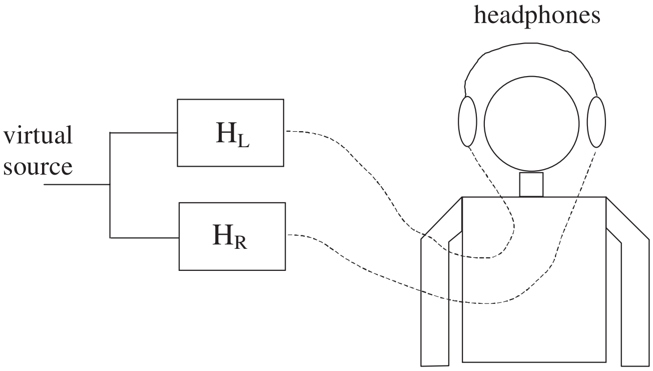


Fig. 2. Binaural spatialization for one source.

A binaural spatializer synthesizes the auditory experience of one or more sound sources located around the listener. This means that it can create theoretically a virtual image of a sound source located at any desired place. However, because of person-to-person differences and computational limitations, it is much easier to control azimuth than elevation or range in practice.

HRTF-based systems are quickly becoming the standard for advanced 3-D audio interfaces. These systems use a database consisting of a measured HRTF sampled at different angles of incidence. An acoustic mannequin, or dummy head, of anthropometrics characteristics with two flush mounted microphones at the ears is normally used in order to compose the database, (Gardner et al., 1995), (Lopez, 1999). The database consists of a set of finite impulse responses of several milliseconds for each azimuth and elevation. Figure 2 shows the concept of binaural spatialization.

The sound coming from the audio source is convolved through the impulse response of the HRTF for a given angle (azimuth and elevation) of incidence. In order to increase reality, the HRTF can be combined with acoustic characteristics of the synthesized room. This way the binaural room-impulse response that completely describes the transfer characteristics of the environment from a sound source to a listener is obtained.

3. Cross-talk cancellation

The delivery of immersive binaural signals can be carried out through headphones or loudspeakers. In this project, we focus our work on loudspeaker systems for two main reasons: there is a large installed base of desktop computers equipped with two loudspeakers, and headphone delivery is uncomfortable for the user producing in-side-the-head virtual sounds (Gardner, 1999). Whereas using headphones it is possible to deliver the appropriate sound field to each ear, loudspeaker systems bring an undesirable cross-talk effect, which must be alleviated for real 3-D immersion.

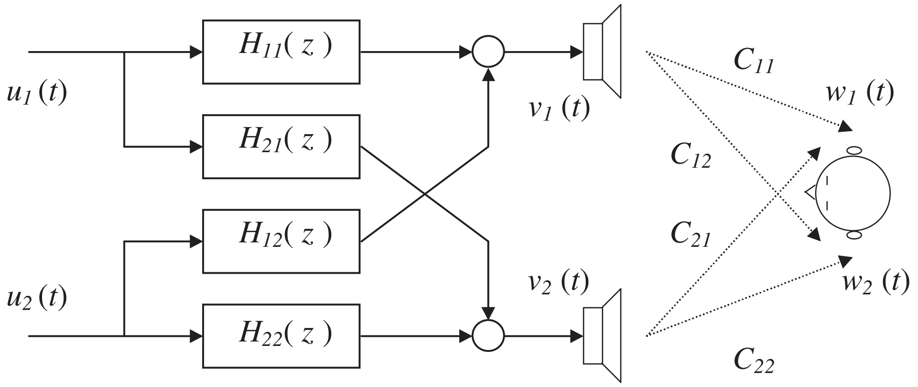


Fig. 3. Cross-talk cancellation scheme.

3.1. Principles of cross-talk cancellation

Cross-talk effect is produced when loudspeakers are used, that means the signal emitted from the left loudspeaker also reaches the right ear, and vice-versa. The common way of removing this distortion consists in placing a filter bank before reproduction through the loudspeakers. Usually, this bank performs the inverse of the Room Impulse Response (RIR) between sources and receivers. This technique was first put into practice by Schroeder and Atal (1963) and later refined by Bauck and Cooper (1996). It is commonly called "transaural audio". Figure 3 shows a cross talk cancellation system formed of four filters.

3.2. Degradation of the cancellation

Due to the sound propagation through air, time arrival delays in the signal between each speaker and each ear are produced, as well as delays of the wall reflected signals. When the listener moves away from the initial (or sweet) point, these delays are altered, modifying the acoustic channel impulse response and therefore, the cross-talk cancellation system performance.

This performance degradation tends to increase with the listener separation from the sweet point, (Nelson et al., 1995), (Lopez et al., 1999a), (Lopez et al., 1999b). Theoretical studies in order to evaluate the influence of the loudspeakers placement in the extension of the cross-talk cancellation zone have been already carried out, (Ward & Elko, 1998), (Asano et al., 1996). For instance, it has been observed that lateral movements larger than 5 cm completely destroy the spatial effect.

In order to allow free movements of the listener, tracking of the listener head has been suggested in previous works, (Gardner, 1997). This tracking can be carried out using electromagnetic trackers, (Blauert et al., 2000). However these devices are very expensive and less common than multimedia devices. That is why digital image processing techniques are proposed in this work to deal with the head tracking task.

The position of the head can be obtained from a multimedia camera mounted just over the monitor, a common and cheap multimedia video camera present at most personal computers, can suffice. Starting from the knowledge of new head positions it is possible to update the bank of cross-talk cancellation filters. The acoustic paths are estimated on-line combining modeled delays for each listener position and a stored HRTF database. Then, the inverse filters are updated in real-time. Interpolation of the HRTF database and averaging between time windows have been used to obtain smooth transitions.

4. Head tracking

Several approaches for estimating head position and orientation have recently proposed, (Horprasert et al., 1997), (Tsukamoto et al., 1994), (Horprasert et al., 1996), (Jebara & Pentland, 1996) with good results. However their high computational cost prevent them from being implemented as a part of a real time system working on a personal computer. In (Birchfield, 1998) a simpler method than the aforementioned ones is proposed, but this method is not as exact as it would be desirable in practice. We have chosen a method that efficiently combines different techniques to give a low computational cost.

The head tracking algorithm implemented comprises three stages. The first stage carries out segmentation of the head from the background. The background can be quite complex, as is shown in Figure 4a). However, fortunately, in most situations it can be considered stationary. A second stage extrapolates the head position in three dimensional coordinates from the segmented head. Finally an eye tracking algorithm estimates yaw rotation of the head based on projected eyes distance.

4.1. Head segmentation

The simplest technique for separating moving objects from a stationary background requires examination of the difference between each new frame and an estimate of the stationary background. Segmentation of moving objects in an uncontrolled environment also requires that the background

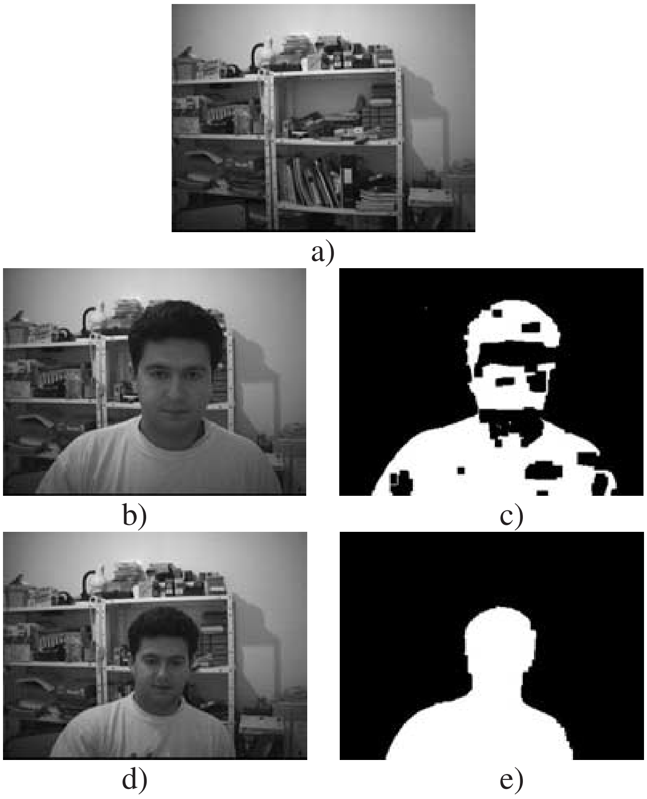


Fig. 4. Head segmentation. a) Background; b) Close face; c) Segmentation of b); d) Far face; e) Segmentation of d).

estimate evolves over time as lighting conditions change. Changes in stationary parts of the image must not be confused with changes due to moving objects.

A modified version of the moving object segmentation method suggested by Karmann and Brandt (1990) is being used. This method is based on an adaptive background model. Background model is updated following a Kalman filtering strategy, thus allowing for dynamics in the model as lighting conditions change. The background is updated each frame via the following equation.

$$B_{t+1} = B_t + (\alpha_1(1 - M_t) + \alpha_2 M_t)D_t \tag{1}$$

Where B_t represents the background model at time t, D_t is the difference between the present frame and background model, and M_t is the binary moving objects hypothesis mask. The gains α_1 and α_2 are based on a estimate of the rate of change of the background. In a complete Kalman filter implementation, these values would be estimated along with the background since they correspond to elements of the error covariance. From our experience, we found that small constant values of $\alpha_1 = 0.1$ and $\alpha_2 = 0.01$ produced good results.

After computing the difference Equation (1), the resulting image is thresholded yielding the silhouette of the listener, as is shown in Figures 4c) and 4e). The head is separated from the rest of the body exploiting the concavity produced by the neck.

4.2. Coordinate extrapolation

The implemented algorithm is capable of obtaining three dimensional coordinates of the head (x,y,z). The x and y coordinates (which define a plane parallel to the screen) are obtained straightway from the mean point of the segmented head.

The z coordinate (distance from the screen) is estimated from the width of the segmented head. The width of the projected head in conic perspective depends on several factors (focal distance of the camera, head width, . . .) that must be also taken into account. Figure 4 shows a head at two different distances. A hyperbolic function relates the head width in pixels in the projected image to the distance from the camera in cm. Figure 5 illustrates the error between the theoretical curve and experimental measurements.

4.3. Yaw estimation

The yaw (head rotation in the vertical axis, ϕ) is determined by tracking the listener eyes. When head rotates, eyes are observed closer compared to the frontal position. A set-up phase is needed before the start of the head tracking. In this set-up, the software asks the user to move the head in order to place the eyes inside two rectangles drawn in the image, Fig. 6a). At the following phase the algorithm stores these sub-images from the eyes in order to track them. A logarithmic search based on correlation of sub-images of the eyes is used to track the movements of the head, Fig. 6b). The relation between the estimated projected eyes distance and head yaw is explained in details in (Horprasert et al., 1997). Distance from the head to the camera must be also taken into account in order to correct the projected image of the eyes as a function of distance. A typical 65 mm eyes distance is used by the algorithm, but can be modified by the user.

5. Implementation

5.1. Binaural synthesis

The implementation of the binaural spatializer is fairly straightforward. Two different selectable databases of HRTF have been used in the prototype. The first one was obtained from a KEMAR acoustic mannequin. The database contents were measured at MIT Media Lab (Gardner et al., 1995), and they are available at its web site. The HRTF data were measured in 10 degrees elevation increments and 5 degree azimuth increments, employing a sampling rate of 44.1 kHz.

The second one was measured by the authors using a B&K 4100 acoustic mannequin (Lopez, 1999). The use of an automatic rotating platform allowed us to measure the HRTF at each degree of azimuth for the horizontal plane and at every few degrees for other angles of elevation. Maximum-Length Sequences of order 16 were used for the measurements that took place in an anechoic chamber in order to

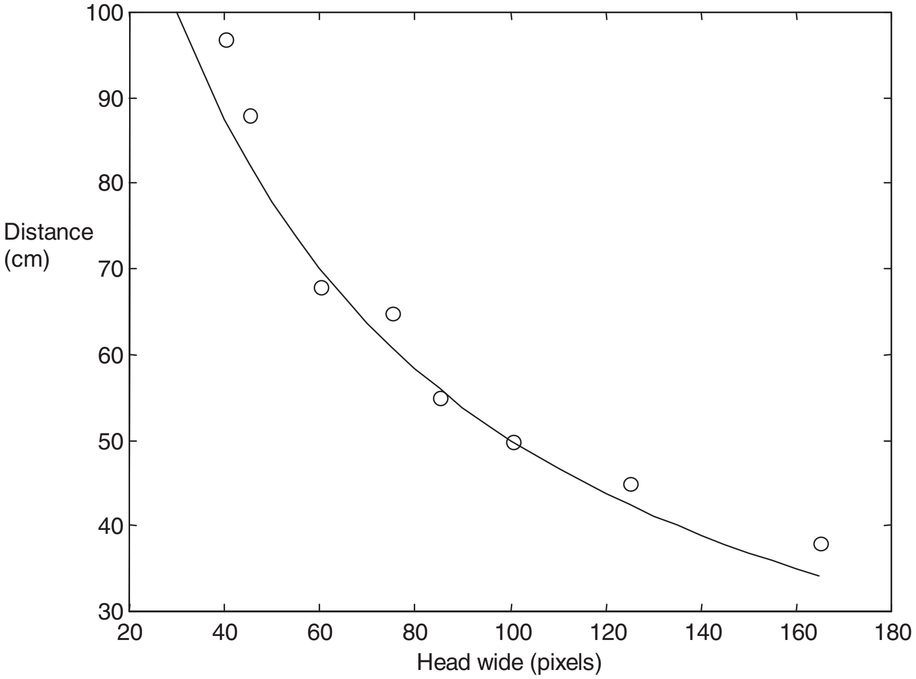


Fig. 5. Head wide versus distance from the screen and error obtained in the measure.

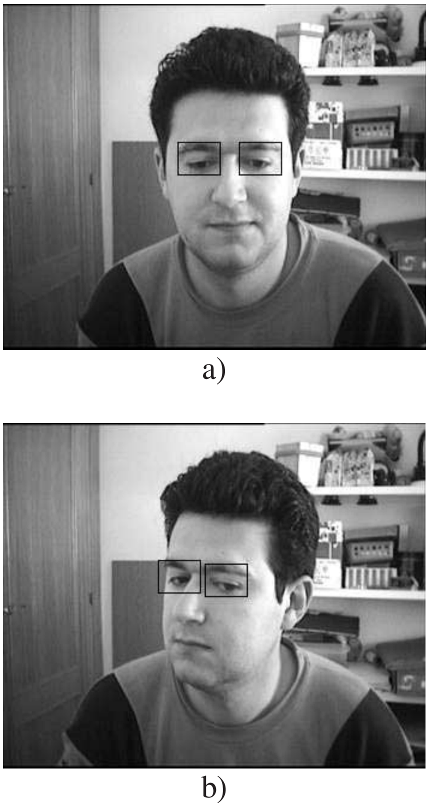


Fig. 6. Eyes Tracking: a) Front view; b) Yaw rotation.

avoid undesirable reflections. Data is available to public the domain at the web site.

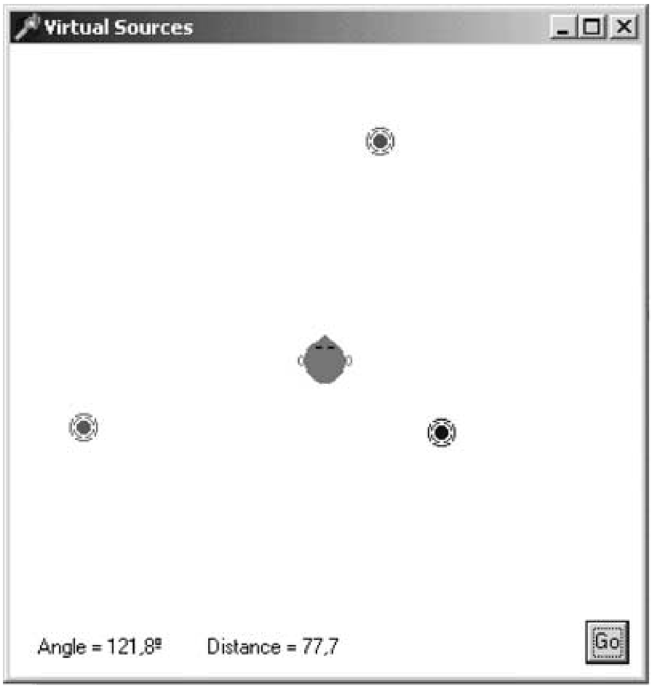
The database of head responses is used for two main purposes in the software system. Firstly, the production of the binaural signal from monaural sources in order to place them at a virtual point as was explained in section 2. When the source is not exactly placed where HRTFs were measured, a bilinear interpolation from the four closest points is made. This interpolation is calculated from the impulse responses in the time domain. The source can be moved around the listener in real time by means of a predefined trajectory or using a pointing device (mouse, joystick or tablet), Figure 7.

The second purpose of the database is related to the estimation of the acoustic paths between the loudspeakers and the listener ears. This path could be measured using two microphones situated at the listener ears. However, this method can be quite uncomfortable for the listener. Instead of that, we estimate the acoustic path combining the HRTFs of the angle formed between each ear of the listener and each loudspeaker, and an estimation of the distance between them. By means of the distance between the loudspeakers and the listener ears, the HRTF impulse response is conveniently delayed. In order to obtain these angles and distances, the position of the head must be known as explained in section 4.

5.2. System architecture

A software tool has been programmed to carry out tests and measurements of the explained system. The software runs over Windows 98/NT/2000 on a standard personal computer. Only the computer main processor is used for audio and video processing, avoiding the use of external processing devices. Therefore the compatibility of the system remains and the hardware is simplified.

The system has been developed employing standard operating system sound drivers allowing to use any compatible



Angle = 121,8º Distance = 77,7

Fig. 7. Software window showing sources position with respect to the listener.

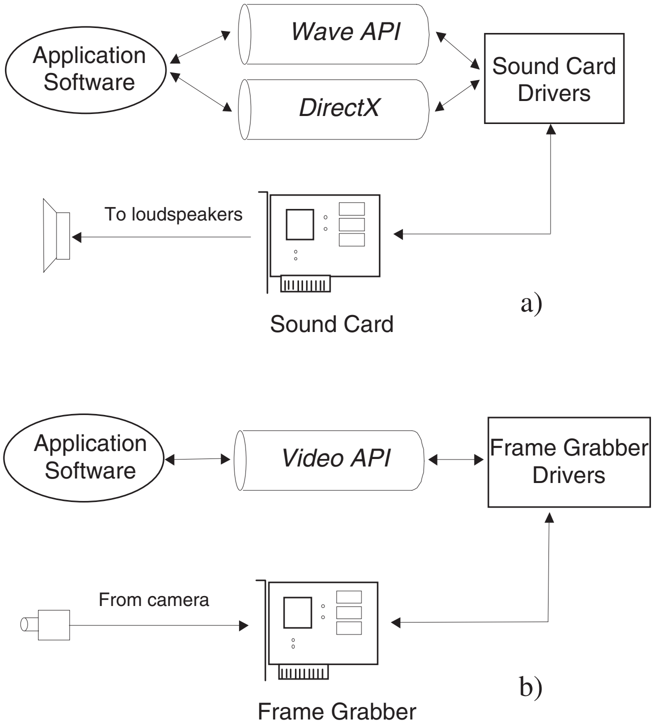


Fig. 8. a) Sound subsystem architecture; b) Video subsystem architecture.

sound card, however a high-quality sound card is recommended for scientific or professional applications. For image acquisition purposes, video capture drivers that work with any compatible frame grabber have been selected. We have used a common and cheap video capture card.

Figure 8 illustrates how the different parts of the system interact. The application software drives the sound card through the standard operating system and the sound card drivers. A similar scheme is used for the video capture subsystem. In order to improve performance while accessing sound card and to reduce the latency of the system presented in (Lopez & Gonzalez, 1999), the new DirectX driver technology of Windows has recently been incorporated in the software. This driver technology in junction with Windows 2000, provides a response time of only 20 ms.

6. Conclusions and future work

A 3-D audio system has been developed for desktop multimedia environments. The system has the ability to place virtual sources at arbitrary azimuths and elevations like other 3-D audio systems, but it presents the innovation of listener tracking. This feature is not implemented in most commercial systems. Dynamic tracking of the listener is an emerging technique that greatly improves audio immersion as it has been tested in this work.

The main drawback of the system is the time delay between a change in the physical environment and the corresponding system response, which is commonly known as latency. Although the latency has decreased in the current software version, it takes too much time to present a fast movement of the listener's head to the listener ears. Moreover a sustained frame rate of 25 Hz cannot be processed by the actual personal computers. The system losses tracking for fast movements of the subject, due to the size limitations in the search area. The use of two computation devices, one dedicated to the head tracking and the other to the 3D audio, seems to be a good alternative in order to considerably improve the performance of the system. The increase of personal computer computational power will allow the use of more refined tracking algorithms and more complex transauralization schemes.

Acknowledgements

This work has been partially supported by Generalitat Valenciana grant no. GV00-094-14 and CICYT grant no. TIC2000-1683-C03.

References

Asano, F., Suzuki, Y., & Sone, T. (1996). Sound equalization using derivative constraints. *Acta Acustica*, *82*, 311–320.

Bauck, J. & Cooper, D.H. (1996). Generalized transaural stereo and applications. *Journal of the Audio Engineering Society*, *44*, 683–705.

Birchfield, S. (1998). Elliptical Head Tracking Using Intensity Gradients and Color Histograms. *Proceeding of IEEE Con-*

ference on Computer Vision and Pattern Recognition, Santa Barbara, California.

Blauert, J., Lehnert, H., Sahrhage, J., & Strauss, H. (2000). An interactive virtual-environment generator for psychoacustic research. I: Architecture and implementation. ACUSTICA – *Acta Acustica*, *86*, 94–102.

Gardner, W.G. & Martin, K.D. (1995). HRTF measurements of a KEMAR. *Journal of the Acoustic Society of America*, *97*, 3907–3908.

Gardner, W.G. (1997). 3-D Audio using loudspeakers, PhD Thesis, MIT.

Horprasert, T., Yacoob, Y., & Davis, L.S. (1996). Computing 3D head orientation from a monocular image sequence, *Proceedings of International Conference on Automatic Face and Gesture Recognition*, 242–247.

Horprasert, T., Yacoob, Y., & Davis, L.S. (1997). An anthropometric shape model for estimating head orientation. 3rd International Workshop on Visual Form, Capri, Italy.

Jebara, T.S. & Pentland, A. (1996). Parametrized structure from motion for 3D adaptive feedback tracking of faces. MIT Media Laboratory Technical Report #401.

Karmann, K.P. & Brandt, A. (1990). Moving object recognition using and adaptive background memory. *Time-Varying Image Processing and Moving Object Recognition*, *2*. Amsterdam: Elsevier.

Lopez, J.J. (1999). HRTF measurements of the acoustic mannequin B&K 4100, Communications Department UPV Internal Report, http://www.dcomg.upv.es/jjlopez/.

Lopez, J.J. & Gonzalez, A. (1999). 3-D audio with dynamic tracking for multimedia environments, Proceedings of DAFx99, Trondheim, Norway.

López, J.J., González, A., & Orduña, F. (1999a). Measurement of cross-talk cancellation and equalization zones in 3-D sound reproduction under real listening conditions. AES 16th Int. Conference on Spatial Sound. Reproduction. Rovaniemi (Finland).

Lopez, J.J., Orduña, F., & Gonzalez, A. (1999b). Equalization zones for cross-talk cancellation as a function of loudspeaker position and room acoustics, *Proceedings of ACTIVE'99*.

Nelson, P.A., Orduña-Bustamante, F., & Hamada, H. (1995). Inverse filter design and equalization zones in multichannel sound reproduction. *IEEE Transactions on Speech and Audio*, *3*, 185–192.

Schroeder, M.R. & Atal, B.S. (1963). Computer simulation of sound transmission in rooms. *IEEE Conv. Record*, *7*. pp. 150–155.

Tsukamoto, A., Lee, C., & Tsuji, S. (1994). Detection and pose estimation of human face with synthesized image models. *ICVV '94*, 754–757.

Ward, D.B. & Elko, G.W. (1998). Optimum Loudspeaker Spacing for Robust Crosstalk Cancellation. In *Proceedings of ICASSP 98*, pp. 3541–3545.