3-D AUDIO WITH DYNAMIC TRACKING FOR MULTIMEDIA ENVIRONTMENTS

José Javier López, Alberto González

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

ABSTRACT

This papers 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 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 for rendering 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) in order to allow free movements of the listener are continuously estimated. Cross-talk cancellation filters and virtual sources locations are updated depending on these head coordinates.

1. INTRODUCTION

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 associated peripherals prices.

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 mentioned elements.

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 presented. 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.

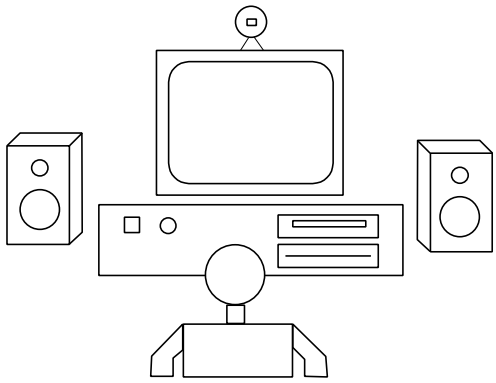

Figure 1. *Common multimedia environment*

2. BINAURAL SYNTHESIS

Humans and other animals have the ability to localize sounds in three dimensions, despite having only two ears. It has been said that "the purpose of the ears is to point the eyes." While the ability of the auditory system to localize sound sources is just one component of our perceptual systems, it has high survival value, and living organisms have found many ways to extract directional information from sound. Although perceptual mysteries remain, the major cues have been known for a long time. Consider a sound source to the right of a listener: sounds from the source arrive at the right ear first, and a short time later they reach the left ear. Also, the amplitude of the left ear sound will be attenuated, particularly at high frequencies, due to the head shadowing.

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. Moreover, its frequency response is directionally dependent.

To find the sound pressure that an arbitrary source x(t) produces at the ear drum, all that is needed is the impulse response h(t) from the source to the ear drum. 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.

A binaural spatializer synthesizes the auditory experience of one or more sound sources located around the listener. This means that, in principle, they can create the impression of a sound being at any desired 3-D location, right or left, up or down, near or far. In practice, because of person to person differences and computational limitations, it is much easier to control azimuth than elevation or range. However, HRTF-based systems are quickly becoming the standard for advanced 3-D audio interfaces.

For doing this operation a database consisting of a measured HRTF sampled at different angles of incidence is needed. It consists of a set of finite impulse responses of several milliseconds for each azimuth and elevation. Figure 2 shows the concept of binaural spacialization. The sound coming from the audio source is filtered through the impulse response of the HRTF for a given angle of incidence.

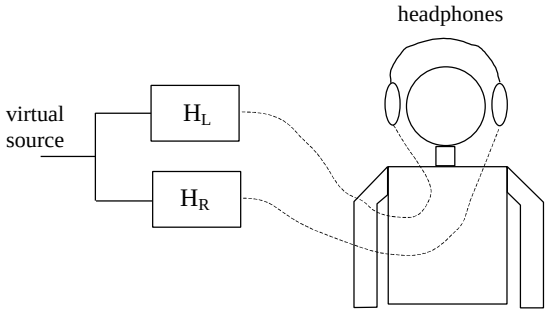

Figure 2. *Binaural spatialization for one source.*

3. CROSS-TALK CANCELLATION

Delivery of immersive binaural signals can be achieved through headphones or loudspeakers. In this project, we focus our work on loudspeaker methods for two reasons: there is a large installed base of desktop computers with two loudspeakers on either side of the monitor, and headphone delivery is uncomfortable for the user producing in-side-the-head. Whereas using headphones it is possible to deliver the appropriate sound field to each ear, loudspeakers use brings an undesirable cross-talk effect, which must be cancelled for real 3D immersion.

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 Schoeder and Atal [1] and later refined by Cooper and Bauck [2]. It is commonly called "transaural audio". Figure 3 shows a cross talk cancellation system.

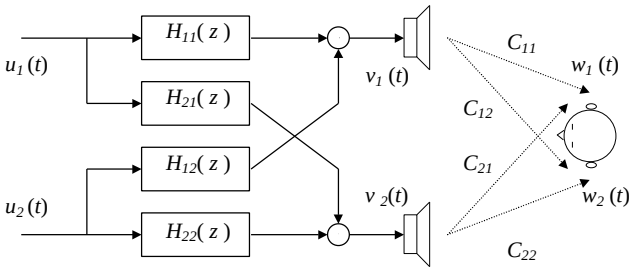

Figure 3. *Cross-talk cancellation scheme.*

3.2. Degradation of the cancellation

However, there exists a practical drawback in the 3-D sound systems through loudspeakers. 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 in 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 overall system performance.

This performance degradation generally increases with the listener separation from the sweet point [3][4]. Theoretical studies in order to evaluate the influence of the loudspeakers placement in the extension of the equalization zone have been already carried out [5][6]. As it was pointed above, the main limitation of these systems resides in the fact that listener movements can degrade the effect of the cross-talk cancellation filters. For instance, it has been observed that lateral movements larger than 5cm 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 [7]. The listener head tracking can be carried out using electromagnetic trackers, but these devices are very expensive. That is why digital image processing techniques are proposed in this work to deal with the head tracking task.

Position of the head would 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 knowing of new head positions is possible to update the bank of cross-talk cancellation filters in order to deal with new acoustic paths between loudspeakers and listener ears.

Acoustic paths must be estimated on-line combining modeled delays for each listener position and a stored HRTF database. Then inverse filters are calculated in real-time.

Interpolation of the HRTF database and average between time windows have been used for obtaining clean transitions.

4. HEAD TRACKING

The head tracking algorithm that was selected comprises two stages. First stage carries out segmentation of the head from the background. Background can be quite complex, as is shown in figure 4a). However, fortunately, in most situations it can be considered stationary. Second stage extrapolates the head position in three dimension coordinates from the segmented head.

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 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 [10] is being use. This method bases 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. Background is updated each frame via the following equation.

$$B_{t+1} = B_t + (a_1(1-M_t) + a_2 M_t) D_t \qquad ()$$

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 simulations, 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 obtaining the silhouette of the listener, as is shown in figures 4c) and 4e). Head is divided from the rest of the body exploiting the concavity produced by the neck.

4.2. Coordinate extrapolation

Implemented algorithm is capable of obtaining three dimensions coordinates of the head (x,y,z) and also head rotation in the vertical axis (ϕ). The x and y coordinates (that define a plane parallel to the screen) are obtained directly from the mean point of the segmented head.

The z coordinate (distance from the screen) is obtained from the wide of the segmented head. The wide of the projected head in conic perspective depends of some factors (focal distance of the camera, head wide,...) that must be also taken into account. Figure 4 shows a head at two different distances. An hyperbola function relates head wide 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.

Finally the head rotation angle (ϕ) is determined by tracking the listener eyes. When head rotates, eyes are observed closer compared to the frontal position. This algorithm is explained in details in [9].

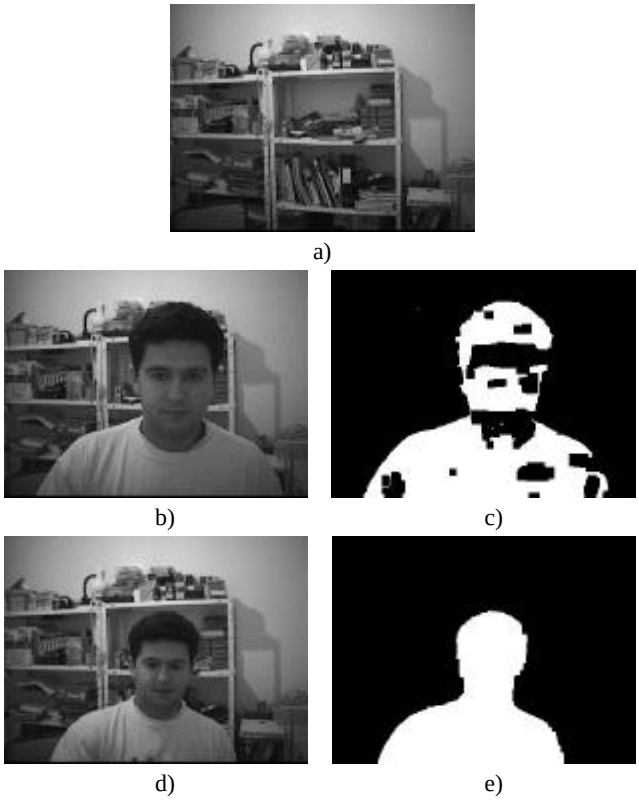

Figure 4. *Head segmentation. a) Background. b) Close face. c) Segmentation of b). d) Far face. e) Segmentation of d)*

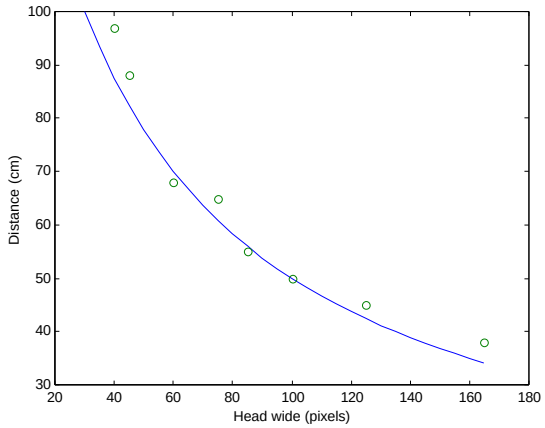

Figure 5. Head wide versus distance from the screen and error obtained in the measure.

5. IMPLEMENTATION

5.1. Binaural synthesis

The implementation of the binaural spatializer results fairly straightforward. A database of HRTFs obtained using a KEMAR acoustic mannequin, also called dummy-head, have been employed. Database contents were measured at MIT Media Lab [8], and they are available at the web site. The HRTF data was measured in 10 degree elevation increments and 5 degree azimuth increments, employing a sampling rate of 44.1 kHz.

The database of head responses is used for two main purposes. Firstly, the production of the binaural signal from monaural sources in order to place them at a virtual point as it 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 directly calculated from the impulse responses in time domain.

The second purpose is related with the estimation of the acoustic paths between loudspeakers and listener ears. This path can be measured using two microphones at the listener ears and a common method of measurement like maximum length sequences (MLS). However, this method can result quite uncomfortable for the listener. Instead of that, we estimate the acoustic path combining the HRTFs of the angle between each ear and each loudspeaker, and a estimation of the distance between them. By means of the distance between the loudspeakers and listener ears, the HRTF impulse response is convenient delayed. In order to obtain this angles and distances, the position of the head must be know 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 on a standard personal computer.

The system has been developed employing standard operating system sound drivers allowing to use any compatible sound card, however a high-quality sound card is recommended for scientific or professional applications. For image acquisition purposes, video capture drivers that allow to employ any compatible frame grabber has been selected. We have used a common and cheap video capture card.

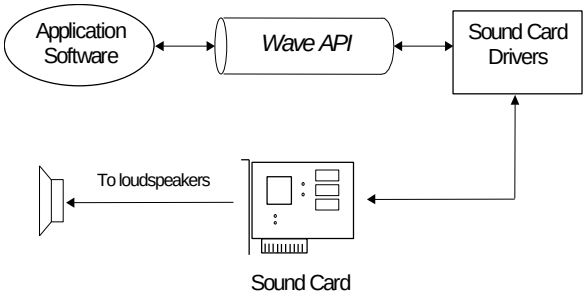

Figure 6. *Sound subsystem architecture*

Figure 6 illustrates how the different parts of the system interact. The application software drives the sound card through the standard OS and sound card drivers. Similar scheme is used in the video capture subsystem.

6. CONCLUSIONS

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 others 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.

Despite technology used in this system is not new, and the performance is far from ideal, real time implementation over a personal computer using common multimedia equipment means an interesting experience.

Main drawback is the latency of the system. It takes too time to present a fast movement of the listener head to the listener ears. Moreover a sustained frame rate of 25 Hz is not possible with the actual personal computers, but it is not too far away.

7. REFERENCES

[1] Schroeder M.R. and Atal B.S. Computer simulation of sound transmission in rooms. IEEE Conv. Record, 7:150-155, 1963

[2] Bauck J. and Cooper, Generalized Transaural Stereo and Applications, Journal of the Audio Engineering Society, 44 (1996) 683-705

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

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

[5] Ward D.B. and Elko G.W., Optimum Loudspeaker Spacing for Robust Crosstalk Cancellation, In Proceedings of ICASSP 98, (1998) 3541-3545

[6] Asano F., Suzuki Y. and Sone T., Sound Equalization Using Derivative Constraints, Acta Acustica, 82 (1996) 311-320

[7] Gardner W.G., 3-D Audio using loudspeakers, PhD Thesis, 1997, MIT

[8] Gardner W.G. and Martin K.D., HRTF measurements of a KEMAR, Journal of the Acoustic Society of America, 97 (6):3907-3908, 1995

[9] Horprasert T., Yacoob Y. And Davis L.S. An Anthropometric Shape Model for Estimating Head Orientation, 3rd International Workshop on Visual Form, Capri, Italy, 1997

[10] Karmann K.P and Brandt A., Moving object recognition using and adaptive background memory, Time-Varying Image Processing and Moving Object Recognition, 2, Elsevier, Amsterdam, 1990