Binaural sonification of disparity maps

Alfonso Alba, Carlos Zubieta, Edgar Arce-Santana

Facultad de Ciencias, Universidad Autónoma de San Luis Potosí.
E-mail: fac@fc.uaslp.mx, zurwolf@gmail.com, arce@fciencias.uaslp.mx

Abstract

A methodology for the realtime spatial sonification of disparity maps obtained from stereo image pairs is presented as part of the development of a scene sonification system for the visually impaired. The proposed system is based on a very efficient segmentation technique, which allows one to detect the most relevant objects in the scene, and a binaural sonification method that produces good results with low computational resources. A first implementation of these techniques is also presented and discussed.

1 Introduction

In computer stereo vision, two cameras, separated by a certain distance, take pictures of the same scene, just like our eyes do [9]. From the stereo pair of images, one can estimate the distance between the camera arrangement and the objects in the scene by finding the matching objects in each picture, and computing the difference between the position of the objects in one image and their positions in the other. The position shift is called *disparity*, and is related to the object's distance Z by the equation $Z = fT/d$, where f is the focal length, T is the distance between the cameras, and d is the disparity. To compute the distance from the cameras to all the objects in the scene, one must estimate the corresponding disparity map; that is, the disparity for each pixel in the images.

Data sonification, on the other hand, uses auditory events to represent information [7]. It has already been successfully used in various fields as an alternative or complement to visualization techniques, when the characteristics of the data may be naturally associated with the properties of sound (pitch, volume, timbre, and duration). A good example is the data obtained from human EEG recordings, where variations in the amplitude and frequency content of an EEG signal can be easily translated into volume and timbre/pitch variations in a certain sound [6]. Sonification is also used in the development of human-computer interfaces (HCI) [5], particularly for visually impaired users.

Humans are capable, to some extent, to determine the approximate location of a sound source. The process of sound localization involves a series of complex physical and psychological processes, which depend on many factors, including

the shape of the listener's ears, head, and shoulders, the acoustic properties of the room, and the sound source itself (e.g., high-frequency sounds are easier to localize than low-frequency ones) [1]. The combined effect of those physical processes on the source signal can be modeled as a head-related transfer function (HRTF) and its corresponding head-related impulse response (HRIR). Therefore, if one knows the HRTF or HRIR for each ear (left and right), corresponding to a sound source at a given position, it is possible to synthesize a stereo audio signal which, when played back through headphones, gives the illusion of a sound coming from that location. One can obtain the HRIR by placing a pair of small microphones on a person or a model's ears, and then reproducing an impulse-like signal (e.g., a hand clap, or a short burst of white noise) that can be captured by the microphones. By repeating this process with the sound source at different positions, one can generate a database of HRIR's which can then be interpolated to estimate the HRIR at any location, thus allowing the simulation of a continuously-moving sound source [3]. Another approach consists in modeling the HRTF by simulating the physical processes that affect the sound until it reaches the ears [2]. These models are usually a oversimplification of the true processes, and thus the quality of the results obtained from these models may not be as good as with pre-recorded HRIR's, but it may be adequate for some applications.

Here we present one such application, which consists in the realtime sonification of disparity maps obtained from a stereo pair of images. This is part of a more ambitious project whose goal is the development of a scene sonification system for the visually impaired.

2 Methodology

An overview of the proposed methodology is as follows: given a disparity map $D(x, y)$ (whose estimation is beyond the scope of this work), we first perform a segmentation to detect the location of the different objects in a scene. Each object is then assigned a different sound, whose corresponding audio signal is fed to a binaural spatialization model with coordinates determined by the object's location. The spatialized stereo signals from all the objects are mixed and output through headphones. The following sections describe each step in detail.

2.1 Disparity map segmentation

In order to detect the various objects in a scene, we perform a segmentation of the disparity map $D(x, y)$ using a very efficient region-growing algorithm with automatic seed selection. Specifically, we compute a region label field $l(x, y)$, which indicates the region to which each pixel belongs. The value $l(x, y) = -1$ indicates an unlabeled pixel. The algorithm chooses a seed from the set of unlabeled pixels, based on a fitness measure $h_q(x, y)$, given by

$$h_q(x, y) = D(x, y)/(d_q(x, y) + 1), \tag{1}$$

where $d_q(x,y)$ is the average neighbor distance defined as

$$d_q(x,y) = \begin{cases} \frac{1}{|N(x,y)|} \sum_{(x',y') \in N(x,y)} \left[(x-x')^2 + (y-y')^2\right], & \text{if } q = 0 \\ \frac{1}{|N(x,y)|} \sum_{(x',y') \in N(x,y)} d_{q-1}(x',y'), & \text{if } q > 0 \end{cases}, \quad (2)$$

where $N(x,y)$ is the first-order neighborhood of (x,y), and q is a quality parameter which controls the size of the neighborhood used to compute the fitness. Increasing q makes the seed selection process more robust to noise (we use $q = 1$ in our tests). Note that the fitness function favors regions which are both homogeneous, and represent the nearest objects (higher disparity). Once a seed (x^*, y^*) is chosen, it is grown into a region by performing the following steps:

1. Assign a new region label to k. For example, if there are currently N_r regions, then let $k = N_r + 1$.
2. Let $p = 1$, $r_k = D(x^*, y^*)$, $l(x^*, y^*) = k$, and $S_p = (x^*, y^*)$. While $p > 0$, do the following:
   (a) Let $(x,y) = S_p$, and decrease p by one.
   (b) For each $(x',y') \in N(x,y)$, if $l(x',y') = -1$ and $|r_k - D(x',y')| < \epsilon$ (for a given threshold ϵ), then let $l(x',y') = k$, $p = p + 1$, and $S_p = (x',y')$.
   (c) If the field l has changed, recompute r_k as $E_{(x,y):l(x,y)=k}[D(x,y)]$, where $E[\cdot]$ denotes the expected value.

2.2 Synthesis of the sound source signals

Ideally, each object in the scene would be assigned a sound that describes it. However, this would involve an object-recognition stage which may be very complex. To simplify things, we use very simple sounds to represent the objects in the scene. These sounds are produced using frequency modulation (FM) synthesis [4], which, in its simplest form, consists of two sinusoidal oscillators: the carrier, and the modulator, with frequencies f_c and f_m, respectively. The source signal, for each object k, is obtained as

$$s_k(t) = E_k(t) \cos\left\{\left[2\pi f_c + i_k(t) \cos(2\pi f_m t)\right] t\right\}, \quad (3)$$

where i_k is the amount of frequency modulation applied to the carrier, and $E_k(t)$ is an envelope function which controls the overall amplitude of the sound. The spectrum of the resulting signal s_k is composed of an infinite number of partials whose frequencies are given by $f_c + n f_m$, $n \in \mathbb{Z}$, and whose amplitudes depend on $i_k(t)$. In particular, if f_m is a multiple of f_n, then the resulting sound will be harmonic, with a fundamental frequency equal to f_c. When $f_m = f_c$, and $i_k(t) = f_c$, the resulting waveform resembles a sawtooth wave, which has a relatively wide spectrum and is therefore a good candidate for binaural spatialization (see below).

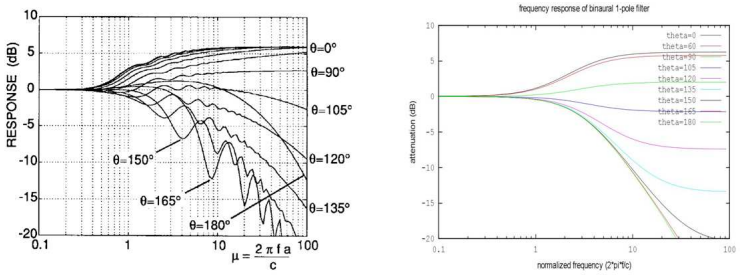

Figure 1: Azimuth-dependant frequency response curves for a rigid sphere model (left graph - taken from [2]), and our one-pole one-zero filter model (right graph). In both graphs, θ represents the angle with respect to the near ear axis.

The envelope function $E_k(t)$ is also a sawtooth wave with frequency between 0.5 and 3 Hz and values between 0 and 1; this results in repetitive ping-like sounds with a fast attack and slower decay. Also, the frequency modulation amount is given by $i_k(t) = f_c * E_k(t)$, so that the sound becomes duller (i.e., less harmonically rich) as it decays; this results in a more pleasant sound that does not interfere so much with sounds from external events. The carrier frequency f_c and envelope frequency are both proportional to the object's average disparity r_k, which makes the sound more alerting as the corresponding object becomes near.

2.3 Binaural spatialization of sound sources

Binaural sonification is achieved by running each of the source signals $s_k(t)$ through a series of processes which provide the cues that our brain uses to localize the sounds. To do this, one must first establish an adequate coordinate system, which in this case consists of spherical coordinates with the center of the head as the origin, the X axis running from the left to the right ear, the Y axis from the base to the top of the head, and the Z axis from the back of the head to the nose. Given the source's location, we define the *azimuth* as the angle between the YZ plane and the source, the *elevation* as the angle from the XZ plane, and the *range* as the distance from the origin.

From the segmentation process, one can easily to obtain, for each region k, the average disparity r_k, geometric center (x_k, y_k), and region size n_k. Regions larger than a given size threshold ϵ_s are sonified by computing the range R_k, azimuth θ_k, and elevation ϕ_k, for each region as

$$R_k = 1/(r_k + 1), \tag{4}$$

$$\theta_k = \sin^{-1}\left[(x_k - W) * \Theta/W\right], \tag{5}$$

$$\phi_k = \sin^{-1}\left[(H - y_k) * \Phi/H\right], \tag{6}$$

where W and H are the half-width and half-height of the disparity image,

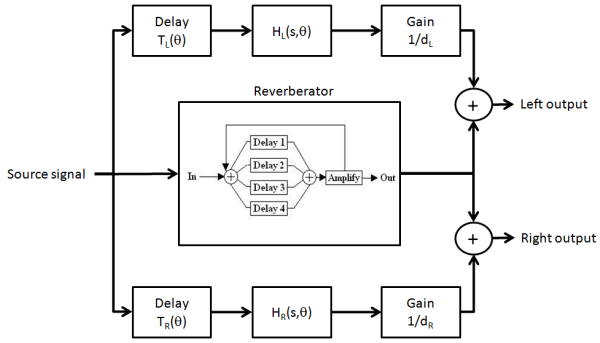

Figure 2: Diagram of the binaural spatialization system.

and Θ and Φ specify the aperture of the sonification space; in other words, $-\Theta \leq \theta_k \leq \Theta$, and $-\Phi \leq \phi_k \leq \Phi$.

In this work we are only modeling azimuth and range cues. The most important azimuth cues are the inter-aural time difference (ITD), and the inter-aural level difference (ILD). The range cues are provided by attenuation and reverberation. See [1] for a detailed description of these and other cues.

ITD cues are synthesized with the model proposed in [2], in which the source signal is delayed by a different amount for each ear. The delay times are given by

$$T_n = a - a\sin(\theta), \qquad \text{and} \qquad T_f = a + a\theta, \tag{7}$$

where T_n and T_f are the delay times for the near ear and far ear, respectively, and a is the head radius (approximately, $a = 0.0875$ cm).

The ILD, also called *head shadow*, is the effect of the sound having to pass through or around the head to reach the far ear. It consists on a frequency-dependent attenuation which varies with both the direction and the distance of the sound source. In particular, we model the head-shadow using a one-pole one-zero filter. The transference function $H(z)$ of the filter is given by

$$H(z) = \left(\frac{1-p}{1-c}\right)\left(\frac{z-c}{z-p}\right), \tag{8}$$

where p and c are the (real) pole and zero, respectively. To obtain the desired response, we fix the pole at $p = 0.85$, and the let the zero vary with respect to the azimuth as follows:

$$c = \left(\frac{\theta'}{\pi}\right)^6 (c_{\min} - c_{\max}) + c_{\max}, \tag{9}$$

where θ' is the angle with respect to the near ear axis, $c_{\min} = -0.175$, and $c_{\max} = 0.93$. These particular values faithfully reproduce the response of a rigid sphere model [2] (see Figure 1).

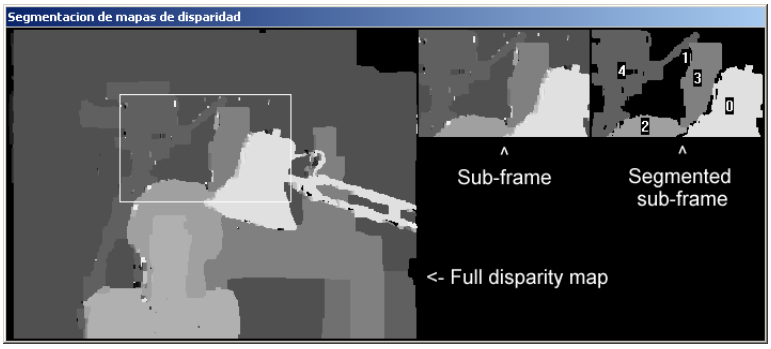

Figure 3: Segmentation of disparity maps. From left to right: full disparity map showing the position of the moving sub-frame (left), 160×100 sub-frame (middle), and segmented sub-frame showing 5 regions and their centers (right).

After applying the ITD and ILD models to the source signal, the result is attenuated by an amount proportional to the square of the distance to the source, according to the inverse quadratic law. Finally, reverberation is added to the attenuated signal. Reverberation is the result of a large number of echoes which result from the reflection of the original sound waves on flat surfaces such as walls [8]. Within an enclosed space, the level of reverberation is roughly constant and independent of the location of the source; thus, the amplitude ratio between the attenuated dry sound and the room's reverberation, is an important cue for range determination. Specifically, we use a very simple reverberation model which consists of four parallel delay lines (with delay times of 3, 7, 13, and 23 ms); the output of the delay lines is mixed, attenuated, and fed back to the inputs. Figure 2 shows the full binaural spatialization system.

3 Preliminary results and conclusions

To test the techniques described above, we simulate a moving scene by taking a 160×100 sub-frame from a pre-computed disparity map and moving it within the larger map (see Figure 3). On a dual-core 2.4 GHz Intel workstation, we achieve segmentation times (with 10 seeds) of 5 ms per frame, and are able to fully process over 100 frames per second. However, these numbers do not take into account the estimation of the disparity maps from the stereo image pair, a process which is not easy to implement in realtime. On the other hand, these methods are intended to be implemented in a portable embedded device, which is much more limited than a PC workstation. Still, the results are encouraging and suggest that the project is certainly viable.

The spatialization technique also works well at placing the objects in their respective locations in the listener's space. It is very easy to determine the direction of each object; the distances, however, can only be determined in a

relative way, when the objects move in relation to the listener's position, or when there are two or more (relatively distant) objects in the scene. Thus, an important shortcoming of our test application is that the objects cannot move towards or away from the listener. In any case, with these techniques it is possible to sonify from 3 to 5 different objects before the representation becomes too acoustically cluttered.

Acknowledgements. This work was supported in part by grants PROMEP/103.5/-07/2416 and C07-FAI-04-19.21.

References

[1] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1997.

[2] P. C. Brown and R. O. Duda. A Structural Model for Binaural Sound Synthesis. *IEEE Transactions on Speech and Audio Processing*, 6(5):476–488, 1998.

[3] D. A. Burgess. Techniques for low cost spatial audio. In *UIST '92: Proceedings of the 5th annual ACM symposium on User interface software and technology*, pages 53–59. ACM, 1992.

[4] J. Chowning. The synthesis of complex audio spectra by means of frequency modulation. *Journal of the Audio Engineering Society*, 21:526–534, 1973.

[5] M. Fernström, E. Brazil, and L. Bannon. HCI Design and Interactive Sonification for Fingers and Ears. *IEEE Multimedia*, 12(2):36–44, 2005.

[6] T. Hinterberger and G. Baier. Parametric Orchestral Sonification of EEG in Real Time. *IEEE Multimedia*, 12(2):70–79, 2005.

[7] R. Minghim and A. Forrest. An illustrated analysis of sonification for scientific visualisation. In *Proceedings of the 6th IEEE Visualization Conference (VIS '95)*, pages 110–117. IEEE Computer Society, 1995.

[8] M. Puckette. *The Theory and Technique of Electronic Music*. World Scientific Publishing, 2007.

[9] E. Trucco and A. Verri. *Introductory Techniques for 3D Computer Vision*. Prentice-Hall, 1998.