8th IEEE International Symposium on Industrial Embedded Systems (SIES 2013)

Embedded Soundscape Rendering for the Visually Impaired

Luca Rizzon and Roberto Passerone
Dipartimento di Ingegneria e Scienza dell'Informazione – Università degli Studi di Trento
Povo di Trento, Italy 38123

***Abstract*—The objective of this work is to improve the quality of life for the visually impaired by enhancing the ability of self navigating. Our system provides a 3D audio representation of the environment by synthesizing virtual sound sources corresponding to obstacles or as a guide for a safe path. The key characteristics of our system are low computational complexity and a simple user customization method. Low complexity makes our system suitable for a resource constrained embedded platform, such as a portable device, while assuring the real time reproduction of the auditive stimuli. In the paper, we discuss the basic perception model, its implementation, and experimental results that show the effectiveness of the approach.**

I. Introduction

Humans use five senses to analyze the world around them, make decision based on the stimuli they collect and react accordingly. In particular, two of these senses are involved in the tasks of orientation, localization and movement: sight and hearing. Able-bodied people commonly use, for the most part, only the sight to orientate when walking, even if this means capturing information lying only inside the field-of-view. Audio information is only partially used for orientation; for example, when a sound representing a dangerous situation is perceived as coming from out-of-sight. At the same time, visually impaired people are more trained to use audio signals to orientate themselves. Unfortunately, for those people who suffer from visually impairments, hearing is not sufficient to guarantee autonomous orientation since it cannot efficiently represent obstacles and safe direction.

The ability of humans to judge the position in space where a sound originates from is called *localization*. Our goal is to take advantage of this capability and synthesize an artificial audio representation of the environment, or *soundscape*, where virtual sounds appear originating from obstacles or guide a person through an established path. Localization is made possible because of the shape and the displacement of human ears, and because our hearing apparatus exploits some spectral cues of the sounds to determine the direction of arrival of sound waves. By modeling the phenomena involved in this process, it is possible to recreate artificial sounds that give the sensation of direction and distance using DSP algorithms. The audio generated by these algorithms is presented at the ears through headphones or earphones. In particular, the phenomena that humans take advantage of in order to localize the position in space where a sound originates are related to the relative position of the listener and the sound event, to the characteristics of the surrounding space and to individual characteristics such as anthropometric measures and personal ability or memories. The features related to human body measures and psychological aspects imply that each individual perceives position of sounds in his/her own way. This translate into the needs for such algorithms to adapt to the listener's needs.

978-1-4799-0658-1/13/$31.00 ©2013 IEEE

In this paper, we discuss the design and implementation of a sound spatialization algorithm which can be integrated into an assistive device for the visually impaired that has to be portable, possibly wearable. Given the limited computational power of a portable embedded platform, the algorithm must run efficiently on the limited resources available. The software must process positional audio combining *low complexity* with the possibility of *being adjustable on the go*. To do this, we developed a bottom-up physical model used to synthesize a simplified transfer function and play back audio signals over the headphones. The model permits the computational requirements to be reduced at the cost of lower accuracy of representation. Still the proposed system can meet the goal of describing spatial information to the listener. Moreover, since it is based on few tunable parameters, it is a promising solution for on-the-go individualization and embedded system implementation. In this paper we describe the algorithm, the implementation on an embedded platform and present the comparison between two different approaches for the rendering of spatial sound used as navigational aid.

II. Background and System Description

There exist different approaches for the synthesis of what is commonly referred to as 3D sound, binaural or positional audio. The most commonly used are based on the Head-Related Impulse Response (HRIR) [1]. The HRIR models the path between a point in space of the real world where the sound originates from and the ear of the listener. HRIRs (or their Fourier transform: HRTFs) are the result of the interaction between the sound waves and the listener's external ear, head and body. Thus, spatialized audio stimuli can be obtained by a convolution between a monoaural sound signal and the impulse response associated to a given space coordinate. Given the dependency of HRIR on the anthropometric measures of the subject, the sound localization is specific for each listener, thus audio representation algorithms must take care of this subjectivity. If audio stimuli are obtained using HRIRs of a subject different from the listener, the listener may commit an error in judging the position of the virtual sound source [2]. As a consequence, recognition performance depends on the degree by which the impulse response fits the listener's one.

HRIRs can be either measured or derived from a physical model. Measuring HRIRs is a long and expensive activity

that requires complex professional audio equipment. On the other hand, physical models require a personalization algorithm used to individualize a generic impulse responses [3]. These algorithms are used to compute a synthesized impulse response of the listener according to some anthropometric measures and sound source relative coordinates. In addition, in order to decrease the interpretation error, a navigational aid system has to account for the sound source and the listener's movements, including movements of the head. Indeed, it has been demonstrated that the localization task becomes easier during movements [4]. For instance, an inertial platform placed on top of the user's head allows the correct displacement of the sources with the listener orientation to be estimated [5]. When HRIRs are measured only for a finite set of locations, responses for an intermediate position must be interpolated, otherwise audible clicking noise are generated when sound position changes. On the contrary, a model-based response can be obtained easily for any point in space.

A complete navigational system must include environment interpretation, position tracking and sound synthesis. To avoid confusing the listener, the system has to process the information and reproduce sound coherent with the head orientation as quickly as possible, meeting the temporal resolution of hearing [4]. In particular, the system must achieve continuous real-time playback over all possible load conditions, otherwise temporal incoherence between reproduction and perception will make it unusable. In this context, we have to adopt design techniques that assure that the overall delay of the system is lower than the timing constraints that arise from the perceptual mechanisms [6]. To achieve this, we use a model-based approach that uses a simple, yet tunable, characterization of the auditory system that can be easily customized to the listener's features. In particular, the design is oriented towards a continuous personalization approach, where the users movements are taken as feedback to estimate and compensate for the localization interpretation errors.

A. Propagation Model

To provide an authentic spatial representation, sounds must be processed so that they give a sensation of direction and spaciousness of the scene. The sensation of direction depends on the relative position between the listener and the sound event, and on the listener's anthropometric features. The sensation of distance and spaciousness in the scene can be rendered by combining two methods. The primary cue of range comes from the amplitude of the incoming sound, which decreases linearly with the distance. However, this cue alone is not sufficient for accurate range judgments, and humans in daily life take advantages of reverberation cues in solving the task of range recognition. Reverberation depends on the room size, its geometry, the materials on the surfaces and the presence of other objects. Contrary to directional cues, this aspect is not subjective, so there is no need for individualization.

We focus our attention on virtual sound sources lying on the horizontal plane. In this case, a listener can solve the localization task thanks to binaural cues [2], that are due to ear displacement and the presence of the head. Effects introduced by the body and the external ear (monoaural cues) are not considered. The two main cues for localization on the horizontal plane are:

- **Interaural Time Difference (ITD):** the difference in arrival time of waves at the two ears due to the different path lengths they travel across;
- **Interaural Level Difference (ILD):** the amplitude difference of the sound wave at the two ears, which is frequency dependent.

These two cues are sufficient to characterize the sensation of left and right displacement of sound sources but cannot convey the sensation of elevation; however, the system can easily be extended to take variations of the elevation angle into consideration. To design the algorithm for the synthesis of approximated responses, we borrow the concept of simplified spherical head model from Brown and Duda [7]. This model can be adapted to the needs of the user by tuning few parameters.

We use a polar coordinate system as shown in Figure 1. According to the model, the synthesis of approximated responses for rendering the sensation of left and right displacement is performed in two steps. First, the signal on the contralateral ear channel is delayed to account for the interaural time difference ΔT, given by

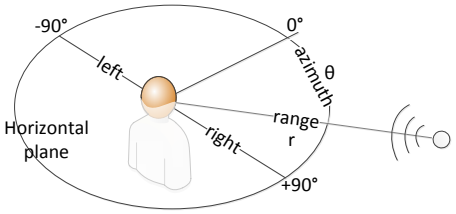

Fig. 1. The vertical polar coordinate system

$$\Delta T = \frac{a}{c}(\sin\theta + \theta), \tag{1}$$

where a represent the head radius of the listener, c the speed of sound in normal condition and θ is the azimuthal angle, which takes value 0 in front of the listener, negative values on the left hand side and positive values on the right. For an average human head radius of $8.75\ cm$, the maximum delay at $\theta = \pm 90°$ is $660\ \mu s$. Then, the interaural level difference is accounted by the shadowing effect introduced by the listener's head, computed using a low pass filter on the signal of the contralateral ear, with the transfer function:

$$H(\omega, \theta) = \frac{1 + j\frac{\alpha\omega}{2\omega_0}}{1 + j\frac{\omega}{2\omega_0}}. \tag{2}$$

The coefficient α, which depends on θ, controls the location of the zeros and is defined in the work by Brown and Duda [7]. To evaluate the effectiveness, we developed a MATLAB model according to eq. (2), and found that a realistic effect can be achieved using a FIR filter implementation with 35-taps. Our implementation computes, for each given value of θ, the correct coefficients that model the frequency dependent low pass phenomenon introduced by the listener's head. This model does not account for the reflection of sound on the shoulders, which are typically included in measured HRIRs, and which would require a larger number of taps. Measured HRIRs, however, only consider a listener facing forward, and are not accurate when the head is turned or tilted. Delayed

versions of the stimuli could easily be added to our model to obtain a more accurate representation. This extension is part of our future work.

To summarize, ΔT represents the ITD while the ILD is expressed by the transfer function$H(\omega, \theta)$. The first is the primary cue for localization of sounds below 1.5 kHz while the latter is more prominent for higher pitch tones. Figure 2

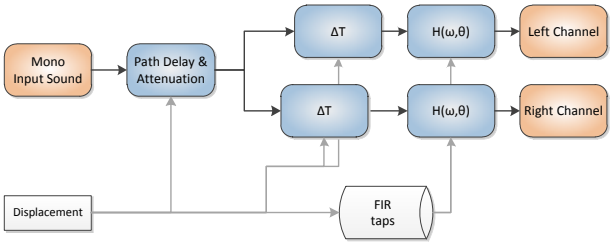

Fig. 2. Block diagram of the software for spatializing one sound source.

shows the block diagram of the propagation model sound algorithm. The monoaural input, which represents the obstacle, is attenuated by a factor r which represents the distance between the user and the relative object in the scene, and then the processing is split into two paths for the left and right channels. The signal for the contralateral ear is delayed and filtered according to the model.

B. Reverberation Model

HRIRs measured in an anechoic room, as well as synthesized HRIRs, do not capture environmental reflections that are present in the real world; this results in an unrealistic sensation, and the listener may feel the source as coming from inside the head. Adding a reverberation effect increases the sensation of spaciousness and helps to cope with the problem of lack of immersion. Reverberation can be seen as a sequence of delayed attenuated replicas of the original sound. Replicas resulting form a single bounce are called early reflections. Higher order reflections are waves that bounce off the surfaces many times before impinging the listener ears and, as a result, they arrive later and very attenuated. If reverb is too large it can be detrimental for position recognition, however, when configured properly, it adds the sensation of distance and spaciousness to the environment. The human brain captures only early reflections for distance judgment; therefore, high reverberation accuracy is not required. In our proposed approach, reverberation is implemented according to a simplified algorithm that derives from ray tracing techniques: the Image Source Method (ISM). The algorithm simulates the presence of a rectangular room whose size depends on the position of the farther sound source in the scene. A sound wave bouncing off a wall can be modeled as a virtual source on the mirror image of the original, behind the wall, as in Figure 3. The line connecting the virtual source to the listener represents the path equivalent to the free field contribution of the mirror image source. Depending on the scene, the algorithm computes the position of the four primary virtual sources that account for the early reflections, as stated above. For each image source, the algorithm computes the correct distance in order to derive the delay at which the sound arrives to the listener's ear with respect to the original direct path source. Each source is then spatialized according to its relative

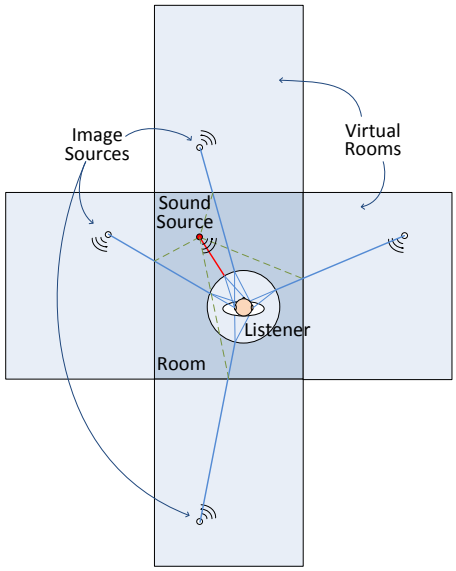

Fig. 3. Intuitive graphical representation of the Image Source Method implementation

direction of arrival. The contribution of the original sound source and the four contributions from the first order reflection are combined (summed up) for the playback.

Contrary to other reverberation algorithms, the ISM combined with our propagation model takes into account the listener and the image source relative position, and properly renders the early order reflections as a binaural sound. While this concept can be extended to higher order reflections, increasing the order may not necessarily be beneficial for the user. Figure 4 shows the complete diagram for our algorithm, including processing of the ISM. Each sound-processing block

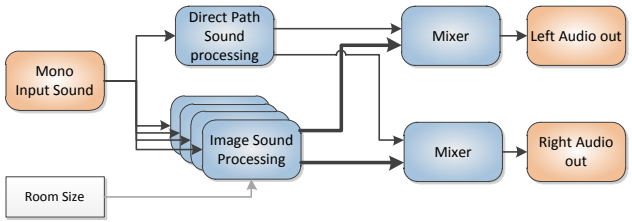

Fig. 4. Block diagram of the complete spatialization algorithm with ISM

is implemented as described in the diagram in Figure 2. The output corresponds to the superimposition of five voices: one for the direct path and four for the virtual image sources.

III. Experimental Results

The method described above has been tested using MATLAB, and was then ported into C in order to test its feasibility on an embedded platform. At this stage of development, the head tracking and environment sensing algorithms have not been implemented yet, so we only discuss the sound synthesis software. We assume that our algorithm takes as input the relative coordinate of the object it has to render and it produces sound. Each time the monitored area changes, the software computes new data it needs to render the scene, and it starts the playback of the relative audio signal as soon as possible. A sampled sound is associated to each source and spatialized

according to the relative azimuthal angle. The concept is to have a listener and a sound source in a rectangular virtual room whose dimensions are proportional to the 2D map the algorithm has to render.

We implemented different versions of the sound rendering algorithm in C. To interface with the sound card we chose the Advanced Linux Sound Architecture (ALSA). It is a framework that provides an API for sound card device drive. ALSA is part of the Linux Kernel and it allows the sound card to be configured, and handles the capture and playback of audio signals. The reason for using ALSA is that it sits just above the hardware, without any intervening layer, minimizing the playback latency. The described implementation can be executed in different Linux based machines without changes. In this work, we compare the time required to process sound according to two algorithms. The first is based on measured HRIRs, available from the CIPIC database [8], downsampled to 100-taps. The other is based on synthesized responses we computed as stated early. For each of the approaches we report the time required to synthesize a single voice output and a five-voice output according to the ISM. For the sake of comparing the execution time in different devices, we run the software on a laptop and on an embedded system platform. The laptop is a 2.5 GHz Intel Core i5 machine with 4 GB 1600 MHz DDR3 RAM. We choose as embedded system the TI Beagleboard, with a 1 GHz ARM Cortex A8 and 512 MB of RAM. Both systems run Ubuntu Linux 12.04 LTS with kernel version 3.1. The algorithm generates audio output at CD quality, 44.1 kHz sample rate, stereo, 16-bit interleaved PCM.

TABLE I. COMPUTATION TIME REQUIRED FOR RENDERING ONE SECOND OF SPATIALIZED SOUND. TIME IS EXPRESSED IN MILLISECONDS

Algorithm	PC	ES
Measured HRIRs	14.6	229.2
Synthesized HRIRs	5.4	82.8
Measured HRIRs and ISM	94.6	1549.6
Synthesized HRIRs and ISM	49.0	793.5

Computation times in Table I are obtained by averaging the time required for 100 executions for each algorithm configuration. For each execution, the system processes 1 second of monoaural input sound and spatializes it for a random coordinate in space. The results show that the spatialization software with measured HRIRs and IMS reverberation is unfeasible on the embedded platform: the computational time is higher than the playback time. Therefore, it cannot provide continuous sound reproduction in real time. On the contrary, the described algorithm can generate sounds within 80% of the playback time. This suggests that the described algorithm is a promising candidate for the development of an assistive device over embedded platforms.

A. Personalization

The use of a physical model makes it easy to personalize the response by changing the values of the parameters. Complex physical models, however, may involve several parameters, which affect the results in hard to predict ways. Our approach simplifies this task by providing a single adjustable parameter: the listener's head radius. As we increase the head radius, the time difference at the two ears increases, as does the cutoff frequency of the filter that models the low pass characteristic of the head. Likewise, the binaural effect is also accentuated when the azimuthal angle increases. Thus, the head radius allows us to vary the "aperture" of the sensation.

If a user has a tendency to overestimate the angle, we may compensate the error with a smaller head radius. Conversely, if a user appears to perceive the space as narrow, we can compensate by increasing the head radius parameter. This information can be acquired implicitly by looking at how the user follows the directions provided by the navigational aid, thus inferring the user real perception. The step size and rate of the radius modification must be propertly set to avoid possible instabilities due to the delay between perception and action, which may induce a positive feedback control loop. Experimenting with this feature is part of our future work.

IV. CONCLUSION AND FUTURE WORK

We have addressed the problem of designing an auditive navigational aid for embedded system applications combined with a customization process. The proposed approach combines techniques that have shown to be both effective in terms of perception, yet sufficiently simple for an efficient embedded system implementation. Experimental results support the feasibility of the approach. The current work is focused on the integration of an inertial platform to detect the user's head orientation. This work is developed in the context of the DALi European project, whose aim is to develop a smart walker for the cognitive impaired. In the project, a vision system is under development to determine the position of obstacles relative to the user, and to establish a safe path in a structured environment. From a methodological point of view, our current work is also dedicated to studying proper resource allocation and scheduling algorithms to be used when partitioning the tasks and sizing the implementation platform, in order to meet the real time constraints.

ACKNOWLEDGMENT

This work was supported by the EU project DALi, grant number ICT-2011-288917.

REFERENCES

[1] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia*. NASA Technical Memorandum, 2000.

[2] J. Blauert, *Spatial Hearing-Revised Edition: The Psychophysics of Human Sound Localization*. MIT press, 1996.

[3] F. Rund and F. Saturka, "Alternatives to hrtf measurement," in *Telecommunications and Signal Processing (TSP), 2012 35th International Conference on*, July, pp. 648–652.

[4] G. M. Murch, *Visual and auditory perception.* Bobbs-Merrill, 1973.

[5] D. Begault, E. Wenzel, A. Lee, and M. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized hrtfs on the spatial perception of a virtual speech source," *J. Audio Eng. Soc*, vol. 49, no. 10, pp. 904–916, 2001.

[6] A. Pinto, A. Bonivento, A. L. Sangiovanni-Vincentelli, R. Passerone, and M. Sgroi, "System level design paradigms: Platform-based design and communication synthesis," *ACM Transactions on Design Automation of Electronic Systems*, vol. 11, no. 3, pp. 537–563, July 2006.

[7] C. Brown and R. Duda, "A structural model for binaural sound synthesis," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 5, pp. 476–488, Sep 1998.

[8] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The cipic hrtf database," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, 2001, pp. 99–102.