

Acoustic Representation of Images for the Visually Impaired: The Human Bat

Mario Alejandro Rojas Quiñones

MSc VIBOT
Heriot Watt University
Universitat de Girona
Université de Bourgogne
marioarojasq@gmail.com

Abstract

Current approaches for exploiting the ability of the auditory system as a mean to transmit information address this issue only to the point of using it as an auxiliary channel or restrict it to environments based on computer graphical interfaces, which limit the complete capability for conveying information about the surroundings. The current project is aimed to tackle this from a different perspective. In order to do so, the relevant aspects of the auditory system with respect to perception are reviewed, the binaural interaural sound perception and the spatialized sound perception models are implemented, on the other hand the nature of sounds and the models for information mapping namely earcons and auditory icons, were also reviewed, guidelines were drawn from the studies that have been done on the subjects and sets of sounds were used. With this theoretical basis, a system is proposed where the combination of a basic video processing block that extracts lateral coordinates from visual markers in a scene and transforms them into azimuth angles with the model of the Head Related Transfer Function that convolves the monophonic sound source signal with the appropriate filters, generates a spatialized version of the created earcons. The system is successfully implemented and its different parts are tested in two separate stages with the purpose of assessing the proposed blocks, sounds, aparent motion schemes and delays and findings confirming the limitations of the HRTF and the challenges of the sound sources design as well as the formulation of human related tests were made.

1. Introduction

In the context of information coding the use of sound is not a new approach and developments regarding physical devices are not by any means a forgotten matter, conversely is the sound perception itself, its nature and the in-

formation transmitting capacity of sounds what is still to be studied, understood and developed. In this direction research and improvements have been made in the fields of Computer-Human Interfaces and in the virtual reality immersive worlds to help user to navigate through them, but little research has been made looking to help sight limited people to move in real environments exploring other sensory channels' abilities to convey information. The perception of sound from the physical point of view is referred to as spatial hearing which is what humans do (hear) every day provided that sounds come to us from all and any direction and distance, both of these parameters supply the sound source with its dimensionality i.e. the spatial location. Sound localization is —although an everyday situation, a complex human process that is a primary tool as far as providing cues about our environment is concerned. In order to artificially generate a three dimensional localization or spatialization of sound, the mechanisms involved in human auditory system must be understood. Different cues are used by humans to localize sounds, there are static and dynamic cues, the first set comprehend the interaural time difference, head shadow, pinna response and shoulder echo, where the second includes head motion, early echo response, reverberation and even vision and are called dynamic because they involve movement of the body affecting how the sound interacts with the ear.

1.1 Physical cues in sound localization

ITD: Interaural Time Difference

The ITD is the time delay between sounds arriving at left and right ears. It is the primary cue for detecting the lateral position of sounds; its value is maximum when sources are located to the sides of the head and minimum when sources

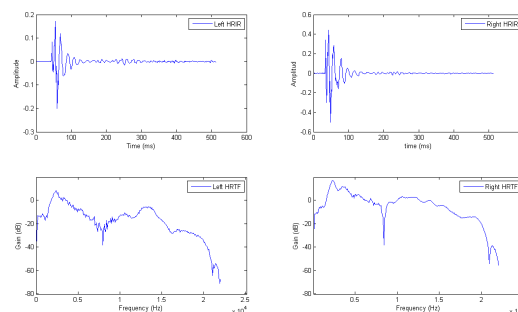
lay in the median plane¹; its computation is a function of the angular (azimuth) location of the source as formulated in [14].

Head Shadow

Is represented in the intensity attenuation suffered by the sound wave that travels through or around the head. This phenomena is termed in the literature [1, 14, 9] as the Interaural Intensity (or Level) Difference (IID or ILD) and is the consequence of some of the energy of the incoming wave being prevented from reaching the ear that is turned away from the direction of the sound. The head shadow also has a filtering effect on the incoming waves that can lead to problems in the perception of direction of the sound sources (specially as far as front/back differentiation is referred). These two cues (ITD and IID) are the most important for localizing a sound's angular position; they work in complementary frequency ranges in a free field location. For signal frequencies' for which the head is larger than the sound wavelength e.g. 1.5 kHz, the IID effect is more pronounced due to the reflection of the incoming sound wave [14]. On the other hand although the ITD effect is present at all frequencies, only sounds for which the maximum ITD is less than half of the period of the wave form, can be interpreted unambiguously thus, this effect is perceived greatly for stimulus components below 1.5 kHz. The manipulation of the values of the IID and ITD parameters determine a special case of localization of the sound, such special case is called lateralization as mentioned in [1]:1. The spatial percept is heard inside the head, mostly along the interaural axis between the ears and 2. the means of producing the percept involves manipulation of the interaural time and intensity differences over headphones. From measurements obtained in different studies, an assessment of the efficacy of these cues regarding the mapping of the displacement of sources has been done and spaces such as the *cone of confusion*[9] have been proposed. It is a virtual surface in which both parameters remain constant, is a revolution surface around the interaural axis of an ideal spherical head where no change in the mentioned parameters would result if a sound source is moved in it. Although changes do present in this region (due to the non spherical shape of the head and the asymmetries present in the outer ears), the magnitude is so small that localization confusions — specially front back, result as a consequence. Other cues for sound source localization include but are not limited to, shoulder echoes, reverberations and pinna response. This last one has an important filtering effect in sounds specially of those with high frequency content. The individual features of each person impose specific spectral features on

¹Imaginary plane that divides the head in right and left halves in a equidistant line from the two ears

Figure 1: Top row: Head Related Impulse Responses. Bottom row: Head Related Transfer Function.



the signals incoming to the eardrums.

Head Related Transfer Function

To produce a real localization of sound rather than a lateralization of it, systematic and controlled measurements have been carried out: using probe microphones in the ear the transfer function from the sound source to the eardrum has been measured in different environments such as anechoic chambers. The result as shown in figure 1² is a complex response function that describes the modifications suffered by the wavefront until reaching the ear canal and is referred to as the *Head Related Transfer Function* or HRTF and its time domain analog the *Head Related Impulse Response* or HRIR which contain all the information about the sound source's location —direction and distance from the listener [9]. The HRTF is defined as the ratio between the sound pressure at the eardrum and, the sound pressure at a point at the center of the head when the listener is absent; given that it requires both ear responses to form a full virtual acoustic source, the HRTF is a binaural function for which the amplitude and time delay differences are the result of the physical shape of the (principally but not only) pinnae which causes specific delays, resonances and diffractions depending on the position of the source and its characteristics hence, unique spectral modifications are imposed on an incoming signal for a given position. The HRTF can then be used for synthesizing stimuli in order to generate virtual 3D acoustic environments through headphones. The use of this function is the most accurate way of producing spatial sound yet an important number of front-back confusion in the localization has been observed and documented when non individualized HRTF are used.

²The plots are obtained from the data recorded by Gardner and Martin[5]

1.2 Sound coding scheme

Alternative means of transmitting information to the user other than the visual system have been explored. Sound is the next (natural) option given that for instance it can be perceived from any direction avoiding the need of having to focus (exclusively) on the display, it can be used to capture the attention of the user as some unexpected or irregular event happens or can be combined with visual information either to present information that would be visually occluded or to complement and increase the amount of it relieving the load of the visual system. Two major approaches have been developed in the area of CHI in order to answer the question, which sound should be used as a tool to communicate information to the user?.

Earcons

Blattner *et.al.* [6] define earcons as “non-verbal audio messages used in the user-computer interface to provide information to the user about some computer object, operation or interaction” and state the structured nature of the sound or combinations of them. Different parameters can be put together to generate the earcons and conform their structure:

Rythm: the duration of the notes.

Pitch: the frequency of the sound element (e.g. middle C is 261.63 Hz).

Timbre: the voice of an instrument.

Register: the location along the different octaves.

Dynamics: the intensity of the sound.

Mechanisms such as **combination** or concatenation of the sound elements, **inheritance** in tree like structures and **transformation** or retrogression of the sound elements or inversion of the pitch, can be used to create compound earcons.

Auditory icons

Auditory icons are defined according to Buxton [2] by two specific features, one is the use of environmental sound controlled along parameters of the events which cause them and the other one is the use of intuitive mappings between the sounds and the events they indicate, highlighting the importance of this latter point. From this definition certain issues arise related to the pertinence and use of auditory icons. Mapping events to produce auditory icons allows the conveyance of multidimensional information, enables a systematic way of relating it and facilitates the generation of a coherent model that increases the transparency of the interface, but in order to comply with these premises researchers

[2, 8] have identified some factors that may affect usability of auditory icons hence should be examined:

Function: the type of information the auditory icons can provide e.g. background, landmarks or interactions.

Mapping: possibilities of expressing the model world object or event in terms of auditory icons i.e. symbolic, metaphorical or iconic depending on the degree of resemblance.

Vocabulary: intuitive mapping, sound effects or clichés.

Annoyance: the unobtrusiveness of sounds.

When considering these topics the bandwidth of the sounds has to be taken into account, the identifiability and ease of remembrance have to be kept in mind as well as different conceptual mappings and problems such as masking have to be evaluated. The main objective of this project is to propose, implement and evaluate a system that explores the possibility of supplying auditory information that resembles the one provided by the visual system providing a user with navigational cues that facilitate motion in a given environment. This is to be achieved 1. by proposing a coding scheme between the visual part and the auditory part, 2. by developing an algorithm to analyse and extract features from a video sequence in real time, 3. by proposing and implementing an algorithm to generate stereo auditory signals (specifically 3D simulated), 4. by deriving a simulation scenario in which the previous scheme can be experimentally evaluated, 5. building a physical set up that enables the capture of live video and the generation of real time sound to a moving subject and 6. experimentally testing in a physical setup the system proposed.

Reseachers have evaluated the different audio coding schemes and have obtained ranges for the different parameters of the earcons as well as provide guidelines for their design [12]. Both coding schemes have been found to be useful in navigational hierarchical menus [13] and modifications on the different parameters are suggested as well [7]. In general the assessment and validation of the two methods for representing audio information present earcons as a useful and versatile mean of communicating information when these are properly parameterized (using the pitch, rythm, timbre, register attributes). On the other hand auditory icons show a useful aid when object recognition tasks are involved (in comparison with earcons) provided that they are congruently related to the object. The application of the (specially) auditory icons in different systems include complex process simulations where background awareness had to be assessed altogether with cooperative work; virtual reality environments that increase the immersiveness of the projection and office applications where the schemes are

used to maintain updated schedules of the workplace. Regarding the use of spatialized or 3D audio, in [4] Lorho *et al* describe a study made in order to determine the ability to localize spatially separated sources of non-speech sounds, also approaches using head tracking to highlight the intensity of the sound being looked at have been done in [3, 11]. None of these approaches take advantage of the capabilities of the auditory system as a main source of information, which serves as a cue to continue this approach.

2. System Formulation

The proposed system captures live video and extracts from it the coordinates (vertical and horizontal) of markers that have been placed in the scene. This set of (horizontal) coordinates maps to angles in the azimuth plane which will determine the HRTF array that will be used to filter the sound to be played to the user via headphones. The project has been divided in two parts, the first comprehends the formulation of the system's functional blocks, the definition of the sound coding scheme and the determination of the most appropriate sound within the defined scheme on the sound component of the system. For the 'visual' component a preliminary analysis of the scenes to process was necessary in order to determine and formulate the algorithm to extract the coordinates, once the algorithm was implemented it had to be merged with the audio block and the entire system was to be tested with a synthetic video sequence that had to be created. The second part uses the findings of the previous one and adapts the video component of the system to work with a video camera. All the algorithms were implemented in MatLab.

2.1 Sound Generation

Models

The two models (ITD&IID and HRTF model) were implemented, first for the interaural axis model to vary the relative delay between the two channels a variable-length zero-padding operation was realized at both ends of the array that represents the monophonic signal. Depending on the relative difference in the padding, the location of the sound along the interaural axis will change; following the literature the maximum delay is 0.66ms that translates into 15 samples when a 22kHz sampling frequency is used. For the HRTF model Gardner and Martin in [5] made the measurements in the MIT's anechoic chamber. They used a KEMAR dummy head equipped with two different pinna models (one for each ear), mounted on a turntable. The speaker was mounted on a boom-stand that enabled vertical positioning to reach accurately any elevation with respect to the KEMAR. The two objects were separated 1.4m

from each other and the measurements were made one elevation at a time starting at an azimuth angle of zero degrees (right in front of the head). Elevations were sampled from -40° (below the horizontal plane) to 90° (directly above the head), and azimuths covered the entire 360° span maintaining 5° steps (where possible). The authors decided to crop the data³ leaving each HRTF 512 samples long and stored in 16 bit integers. Additionally three different attenuation manipulation functions were proposed (gaussian, cosine and constant) to simulate depth perception.

Sound Coding Schemes

A mid point between auditory icons and earcons was taken, i.e. sounds that had a basic type of relation with the 'surrounding environment' and arbitrary sounds that were easily recognizable but had no relation with it. Four sounds were recorded and adapted as MatLab's **.mat** files: 1. person pronouncing letter 'a' 2. person pronouncing letter 'b', 3. person clapping once and 4. a solid object striking a wooden surface.

2.2 Video Processing

A binary type of image was established as the most appropriate scene to capture given the nature of the project. An algorithm based on binarization and morphological operators was proposed. Given that the processing is to be done on (mostly) moving objects in a video sequence, two consecutive frames were grabbed, each individually thresholded and the difference taken. Subsequently the 'opening' morphological operation was applied in order to remove small clusters of objects resulting from the binarization-difference process. The coordinates of the objects of interest were extracted from the complement of the opened (image) matrix and were returned as outputs to the user.

Clip Generation

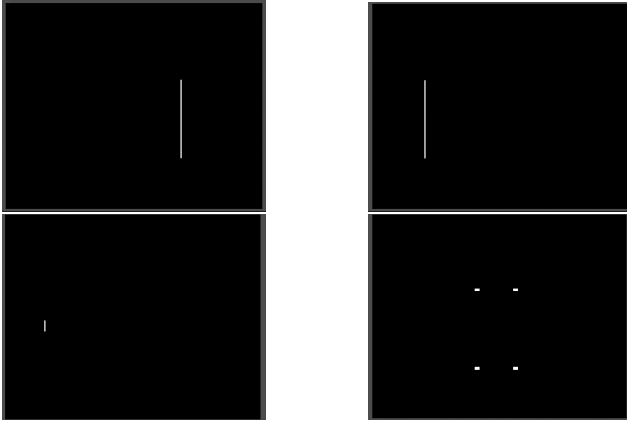
In order to test the system on ideal conditions, first four clips were created (as shown in figure 2) where objects were presented in different motion schemes altogether with the sounds matching the apparent motions being shown.

2.3 Components Combination

The merging of the sound and video components was made in a script that received as parameters the path for the video clip, the monophonic sound signal and the attenuation function identifier. From the video sequence the number of frames and the image dimensions were extracted. Subsequently the (two every five) frames were read and sent to the

³ Available at <http://sound.media.mit.edu/KEMAR.html>

Figure 2: Screenshots of the four clips.



function that obtains the coordinates. These coordinates are mapped to azimuth angles following the measurement sampling: if the horizontal plane (0° of elevation) is considered the sampling of the space around the dummy head is done in steps of 5° covering the full circumference, then the number of measures between 0° and 90° is 19. Taking into account that the coordinates are extracted from a video sequence of a camera that is pointing towards the front of the person, the semicircle covering the interval between -90° and 90° will be the maximum visible field in the scene (depending on the optics used). The horizontal coordinate-angle transformation takes the center of the image as a starting point and divides each half in the number of steps mentioned before, creating 19 bins per half of the image. When a coordinate is obtained, half of the image width is subtracted to offset it to the appropriate section, subsequently the result is divided by the bin index yielding the number of steps that represent the angular location which can be finally obtained by multiplying the last result by the angular width proposed in the beginning: the mentioned 5° . This is the angle passed to the HRTF retrieving function as azimuth parameter. Once the appropriate HRTF is loaded, each channel is convolved with the monophonic source signal and multiplied by the attenuation function.

Simulation

The test was devised in order to evaluate the different elements mentioned so far, the formulation is largely based on the tests realised by Lorho [4] and take elements of the ones carried out in [12, 7, 10]. It was divided in three main parts; the first is aimed to gather information about the sound sources; also to determine if one of the two ear responses performs better given that those functions are non-individualized and to evaluate the attenuation functions pro-

posed. The four sounds sampled were presented in a simulated motion scheme with nine positions and two different (alternated) directions. The wooden and the clap sounds were played moving front to back along the right side of the head conversely to the speech sounds that moved back to front along the left. The first two sequences played different ear responses and the third an inside of the head (ITD&IID model) sound; every three sequences the attenuation function was changed. The second part is aimed to determine the ability to discern simultaneous sounds with different sources that can be positively located and to confirm whether previous results are useful. Subjects were presented the four sounds one at a time simultaneously in two different spatial locations in a front to back apparent motion scheme. The process was repeated for the two ear responses. The third part displayed the four clips each with the spatialized sounds that resemble the shown motion and it was intended to determine the performance of the combined components.

Results and Discussion

Unfortunately only five subjects participated in the test, two females and three males, Due to this fact any statistical analysis is precluded, thus only the main observations and trends are commented. The test were carried out in quiet locations using Sony MDR-E818LP headphones and a text-based interface was presented. From the first part the left ear response was taken to provide a slightly better perception; results confirmed the fact of the front-back confusions in the non-individualized HRTF model and the interaural-inhead perception of the ITD & IID model. Regarding the speech sounds these were favoured yet recalling the fact that no significant delays were introduced for any of the sounds, causing a degeneration of the perception. Contradictions about the attenuation functions manipulating the intensity were observed furthermore, the last clip provided some evidence that it may not be useful to modulate it if not real depth (distance to the source) information is acquired.

3. System Tuning

Taking into account the findings of the simulation stage, the system was taken a step further to the physical setup. From the author's perception and the previously mentioned results, the LEFT ear response was to be used during the remaining of the project. Concerning the sounds, given that no clear preference was observed due to the questionable (contradictory) results between the first and second trials, the four sounds were still going to be analysed in this stage; finally given the results for the attenuation functions proposed, the idea was discarded at this stage. To obtain the binary image (which was decided to be the best alternative to the objectives of the project), infrared LEDs (central

frequency equivalent to 950nm) were used in the scene as markers and were connected in a series-parallel circuit and placed in front of a Retiga 1300R, 12 bit monochrome camera that that was equipped with a Nikkor 28mm NIKON lens and, which is interfaced via a FireWire IEEE 1394 link and driven using the MatLab Image Acquisition ToolBox. Only three modifications were made in the combined algorithm, first the video grabbing function was adapted by generating a video input object and the frame rate was reduced in order to decrease the delay, this influences the binarization process which is done on only one every fifth frame rather than in the difference. The second modification is in the sound set, the speech sounds were dropped given that they suffer a masking effect when presented simultaneously or with short delays, where as the other two sounds do not suffer this hence, a second sound set was designed following the guidelines given in the literature: a piano and a guitar sound were synthesized using a free software. The last modification is in the filtering stage, where a dynamic zero padding step is added in order to generate a delay between the sounds.

3.1 Test

Due to the fact that the main obstacle remained in the type of sounds to be used rather than in the combined sound-video system, a second sound based test was devised in order to determine the number of sounds perceived, in/out of the head perception, motion scheme evaluation, quantitative evaluation of the difficulty of localization, appropriate time delays and order of reproduction. A script was written that served to test the two sets of sounds. An array of 20 positions (where every 4 represented the same azimuth angle), was used to retrieve the HRTF that would be convolved with either two or three zero-padded sounds. The sound set, number of sounds, the amount of delay and the order of reproduction were passed as parameters. The test was divided in two blocks. The first presented two spatialized sounds with a motion scheme starting in front of the head and moving towards the ears; the first two trials changed the type of sound—first the wood/clap and second the synthetic instruments, maintaining the order of reproduction constant; the next two changed to a different order of reproduction and changed the types of sounds for the third and fourth trials. For this part two delay values (0.1 and 0.2 s) between sounds were presented yielding 4 trials. The second part presented three sounds, the two lateral kept the same motion scheme and the third was placed either in front or behind the head. First variations over the ‘front/back’ object were made keeping constant the sound type and the reproduction order; subsequently the sound set was changed and the trial repeated in the same conditions as before, finally the reproduction order was changed and all the previous tri-

als were performed.

3.2 Results and Discussion

The number of sounds as well as the out of the head localization (for the lateral sounds) was perceived consistently among subjects in all the trials. Front/back confusions were observed once more, furthermore the sound source located in the median plane was perceived on top of the head repeatedly. A delay of 0.2s was preferred and no order of reproduction showed a significantly better performance. For the motion scheme a poor perception of localization limited to a static position was observed and finally the original sound set was perceived as easier to localize and as less obtrusive.

4 Conclusions

A preview of binaural hearing theory and sound coding scheme was made providing the conceptual basis, also the survey of studies and systems that assess, validate and implement these two topic, provided a framework for the project. A system that from a video capture extracts lateral coordinates and generates spatialized audio was proposed and successfully implemented. The HRTF model and its main feature the “out of the head perception of localization” was confirmed as well as the front/back confusion consequence of the non-individualized function, was observed. Earcons were found to be better as a coding scheme for spatialized sounds, yet a more thorough study of the characteristics and design guidelines has to be made in order to take advantage of their features and to avoid effects such as masking. The application of small delays between sounds when several sources are presented improves significantly the ability to localize individual sources. The formulation and realization of tests involving human factors such as opinions on difficulties is a complex process that requires a more careful approach and which has to consider all the statistical premises about sample size for validation of the experiments. Overall the problem tackled seems to have a good or at least interesting possibilities of developing new concepts for aiding tools for visually impaired people.

5 Future Work

The most important point is to improve the sound coding scheme taking into account the psychoacoustics of human perception. On the other hand, obtaining individualized HRTF to avoid front/back confusions would reduce the amount of confusions even if it means reducing the amount of subjects it can be applied to. Performing the

capture of images with a stereo pair to gain full 3D information of the source, would work in the same direction as the previous point limiting the variables to examine. Finally a variation of the presentation of the location of the sound sources could provide different insights to the perception of the sounds i.e. not only playing the sound at its current position only, but reproduce the sound from a default location until it reaches the actual position of the source.

References

- [1] Durand R. Begault. *3-D Sound for Virtual Reality and Multimedia*. NASA, April 2000.
- [2] Gaver & Bly Buxton. *Auditory Icons*, chapter 6 Auditory Icons, page 17. Buxton, Gaver & Bly, 2002.
- [3] Alan Kan Dennis Lin Andr van Schaik Keir Smith Craig Jin, Teewon Tan and Mathew McGinity. Real-time head-tracked 3d audio with unlimited simultaneous sounds. In *ICAD 05 Eleventh Meeting of the International Conference on Auditory Display*, July 2005.
- [4] Juhu Marila Gaetan Lorho and Jarno Hiipakka. Feasibility of multiple non-speech sound presentation using headphones. *International Conference on Auditory Display*, page 6, 2001.
- [5] Bill Gardner and Keith Martin. Hrtf measurements of a kemar dummy-head microphone. Technical report, MIT, 1994.
- [6] Denise A. Sumikawa Meera M. Blattner and Robert M. Greenberg. Earcons and icons: Their structure and design principles. In Lawrence Erlbaum Associates Inc, editor, *Human-Computer Interaction*, volume 4, page 34. 1989.
- [7] Elizabeth D. Mynatt. Auditory presentation of graphical user interfaces.
- [8] Elizabeth D. Mynatt. Designing with auditory icons: How well do we identify auditory cues? In *Conference on Human Factors in Computer Systems*, 1994.
- [9] Rod Jard Paholio. 3d sound synthesis for headphones and spatial audio. DSP Final Research Paper, 2005.
- [10] Myra P. Bussemakers Paul M. Lemmens and Abraham de Haan. Effects of auditory icons and earcons on visual categorization: The bigger picture. *International Conference on Auditory Display*, page 9, 2001.
- [11] Chirs Schmandt and Atty Mullins. Audiostreamer: Exploiting simultaneity for listening. In *CHI 95*. Addison Wesley, 1995.
- [12] Peter C. Wriyth Stephen A. Brewster and Alistair D. Edwards. An evaluation of earcons for use in auditory human-computer interfaces. In *Conference on Human Factors in Computing Systems*, 1993.
- [13] Veli-Pekka Raty Stephen Brewster and Atte Kortekangs. Earcons as a method of providing navigational cues in a menu hierarchy. In Cunningham R.J. Sasse M. A. and Winder R., editors, *11th British Computer Society HCI Conference*, pages 167 – 183, August 1996.
- [14] DeLiang Wang. *Computational Auditory Scene Analysis*, chapter 5. Binaural Sound Localization, page 34. John Wiley & Sons, Inc., 2005.