SPATIAL LOCALISATION OF AUDIO SIGNALS

DSP MINI PROJECT

APRIL 2016

DONE BY:

ADARSH RAJESH

AKARSH PRABHAKAR

CHIRAG GOURAV G

UNDER THE GUIDANCE OF:

DR.PATHIPATI SRIHARI

DEPARTMENT OF ECE

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

Acknowledgement

The completion of this project on *'Spatial localisation of Audio Signals'* gives us much pleasure. We would like to show our sincere gratitude to Dr. Pathipati Srihari, course instructor, for providing us a good guideline for the project and his constant guidance through numerous consultations. We also thank Dr. A V Narasimhadhan, who introduced us to the fundamentals of Digital Signal Processing, and whose passion had a lasting effect on us. We also thank the Department of ECE, NITK for providing us with this wonderful opportunity.

Many of our friends, especially our classmates, have made valuable comments and suggestions on this proposal which gave us an inspiration to improve our project. We would like to extend our deepest gratitude to all those who have directly and indirectly supported us in completing this project.

Last but not the least; we thank all the authors of the references mentioned in the bibliography, for sharing their brilliant insight on the topic.

Adarsh Rajesh, 14EC201
Akarsh Prabhakar, 14EC104
Chirag Gourav G, 14EC212

Table of Contents

INTRODUCTION 4

ABSTRACT 6

HEADPHONE PLAYBACK 7

0.1 AZIMUTHAL VARIATION 7
0.2 ELEVATIONAL VARIATION 8
0.3 ROOM REVERBERATION AND RANGE 9

LOUDSPEAKER PLAYBACK 11

IMPLEMENTATION AND RESULTS 13

CONCLUSION AND FUTURE SCOPE 15

BIBLIOGRAPHY 16

Introduction

Audio recordings are generally mono or stereo recordings. When the mono recording is played back using headphones, the left and the right channels receive the same signals and thus the audio appears to have been generated within the head. On the other hand when a stereo recording is played back using headphones, the left and right channels receive signals that were recorded by the left and right microphones respectively, hence different. This form of playback provides a better sound localization than mono but nevertheless it doesn't provide the extra sense of reality and naturalness.

Spatial audio (sometimes called 3D audio) is a term which is used to provide playback on headphones or loudspeakers such that the listener is able feel the presence of the sound source, its movement, location, relative distance, altitude and so on. The listener feels as though he is actually standing in front of the sound source and listening. It is possible to recreate live concert, video conferencing, drama, gaming and many other environments using spatial audio.

The playback of audio signals on headphones with good spatial localisation of the source is called binaural playback. These audio signals can be directly obtained from binaural recordings or from mono recorded signals after some processing.

Binaural recordings use a dummy human head shaped model with two microphones, one in each ear. The sound to be recorded is played and the outputs of the microphones give binaural recording. It is called binaural because unlike stereo recording the left and the right channel microphones are placed inside a human ear. As a result of which the waves reaching the ear will not be coming directly from the source, but will face multiple obstacles like the frontal part of the head, the torso, the shoulder and the pinna. Thus the outputs from the microphones vary from person to person.

In order to use these characteristics of a person which affects the wave travelling from the source to the ear, we use the idea of Head Related Transfer Functions (HRTFs). This is how it works. For each person the microphones are placed in the ears and a sound is produced from all possible directions. The responses from the microphones are noted. The output is something like the impulse response of a system. Here the sound produced is impulse and the transmission medium, the

various face boundaries, torso boundaries and other edges form the system. Now for a given single channel audio input if we want to add spatial information, we can do this by convolving the audio input with the impulse response recorded from the experiments conducted on him. Hence the name, Head Related Transfer Functions.

But using HRTFs will involve setting up good anechoic chambers, using high quality microphones and requires huge data storage. Moreover HRTFs vary from person to person. In order to avoid this, we use the idea of doing some processing operations on a dry mono recording and giving two outputs, one for the left channel and another for the right channel. This not only saves a lot of operation costs but also makes it easily implementable on DSP hardware for real time applications.

If these binaural signals are played on a loudspeaker the spatial effects are lost. This is because the binaural recordings are done at the entrance to the ear canal. When played back, the recordings should reach only that ear. Headphone playback is hence ideal. But if loudspeaker playback is done, the left channel signals reach left and right ears and so does the right channel signals. Therefore all spatial effects are lost. We need to ensure that the left channel signals reach only the left ear and the right channel signal reach only the right ear. We achieve this by building some sort of a wall between the two ears. Thus we are able to retain the spatial effects of binaural signals in loudspeaker playback as well.

Abstract

The objective of this project is to spatially localize the direction of the sound's source by just listening to a processed single channel mono audio recording being played on a pair of stereo headphones or stereo loudspeakers.

Sound sources are localized by our brain using different inter-aural cues. The brain interprets the difference in time and amplitude on the sound arriving to each ear. The time difference and resonances are due to the shape of the body, frontal head and pinna. Sound arriving at the ears is frequency altered or filtered according to the physiognomy of the listener.

Using available mathematical models for the head and various other body parts we build our filter which gives audio output with spatial effects. The models are chosen to account for azimuthal, elevation, range variations and room reverberations. For azimuthal variation we include the head shadow filter and inter aural time delay filters. In order to incorporate elevation changes we use pinna and shoulder filter. The room reverberation is achieved by just controlling the echo intensity and time delay filter. The output of all these filters put together drives the headphone output.

The output of all the above filters must be further processed before driving the loudspeakers. A technique called crosstalk cancellation is used in order to nullify the effect of the farther channel at each ear. A combination of head shadow filter, time delay filter and comb filter with a low pass filter in the feedback loop is used to implement this technique. On passing binaural signals through this combination we get signals which can drive the loudspeaker and ensure that the spatial effects remain intact.

With this project we aim to show that instead of using HRTFs which require high operational costs and storage costs, using the concept of modelling the human parts responsible for affecting waves from source to ears is more elegant and suitable for real time applications. Also this can be made to suit any person by just changing a few parameters in the model. With this approach we can create a non personalised spatial audio environment using headphones and loudspeakers.

Headphone playback

Modeling the structural properties of the system pinna-head-torso gives us the possibility of applying continuous variation to the positions of sound sources and to the morphology of the listener. Much of the physical/geometric properties can be understood by careful analysis of the HRTFs, plotted as surfaces, functions of the variables time and azimuth, or time and elevation. This is the approach taken by Brown and Duda who came up with a model which can be structurally divided into three parts:

- Head Shadow and ITD
- Shoulder Echo
- Pinna Reflections

0.1 Azimuthal variation

Starting from the approximation of the head as a rigid sphere that diffracts a plane wave, the shadowing effect can be effectively approximated by a first-order continuous-time system, that is, a pole-zero couple in the Laplace complex plane:

$$s_z = \frac{-2\omega_0}{\alpha(\theta)}$$
$$s_p = -2\omega_0$$

where, wo is related to the effective radius *a* of the head and the speed of sound c by

$$\omega_0 = \frac{c}{a}.$$

The position of the zero varies with the azimuth according to the function

$$\alpha(\theta) = 1.05 + 0.95\cos\left(\frac{\theta}{150^\circ}180^\circ\right)$$

The pole-zero couple can be directly translated into a stable IIR digital filter by bilinear transformation, and the resulting filter (with proper scaling) is

$$H_{\mathrm{hs}} = \frac{(\omega_0 + \alpha F_s) + (\omega_0 - \alpha F_s) z^{-1}}{(\omega_0 + F_s) + (\omega_0 - F_s) z^{-1}} .$$

The ITD can be obtained by means of a first-order all pass filter whose group delay in seconds is the following function of the azimuth angle

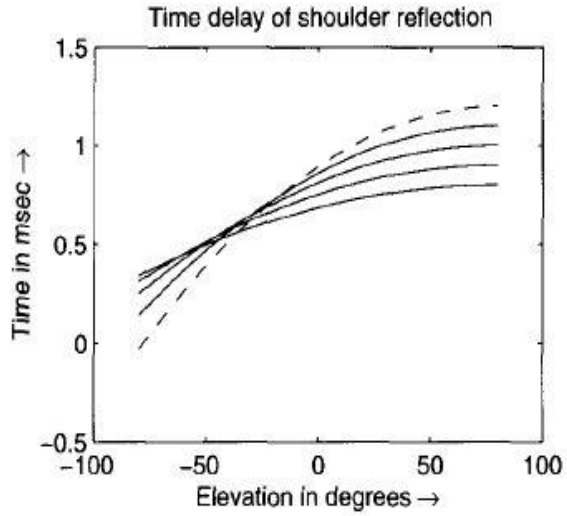
$$\tau_{\mathrm{h}}(\theta) = \begin{cases} -\frac{a}{c}\cos\theta & \text{if } 0 \le |\theta| < \frac{\pi}{2} \\ \frac{a}{c}\left(|\theta| - \frac{\pi}{2}\right) & \text{if } \frac{\pi}{2} \le |\theta| < \pi \end{cases}$$

Actually, the group delay provided by the all pass filter varies with frequency, but for these purposes such variability can be neglected. Instead, the head shadow filter gives an excess delay at DC that is about 50 percent. This increase in the group delay at DC is exactly what one observes for the real head.

0.2 Elevational variation

The shoulder echo and pinna reflections contribute to the elevation variation. In a rough approximation, the shoulder and torso effects are synthesized in a single echo. An approximate expression of the time delay can be deduced by the measurements

$$\tau_{\mathrm{sh}} = 1.2\frac{180^\circ - \theta}{180^\circ}\left(1 - 0.00004\left((\phi - 80^\circ)\frac{180^\circ}{180^\circ + \theta}\right)^2\right) \text{ in ms} ,$$



The pinna provides multiple reflections that can be obtained by means of a tapped delay line. In the frequency domain, these short echoes translate into notches whose position is elevation dependent and that are frequently considered as the main cue for the perception of elevation. A formula for the time delay of these echoes is given:

$$\tau_{\mathrm{p}n} = A_n \cos(\theta/2) \sin(D_n(90^\circ - \phi)) + B_n$$

n	$\rho_{\mathrm{p}n}$	A_n[samples]	B_n[samples]	D_n
2	0.5	1	2	$\cong 1$
3	-1	5	4	$\cong 0.5$
4	0.5	5	7	$\cong 0.5$
5	-0.25	5	11	$\cong 0.5$
6	0.25	5	13	$\cong 0.5$

The parameter Dn, allows the adjustment of the model to the individual characteristics of the pinna, thus providing an effective knob for optimizing the localization properties of the model.

The structural model of the pinna-head-torso system is depicted in the figure in the next page with all its three functional blocks, repeated twice for the two ears. Even though the directional properties are retained by the model, anechoic sounds filtered by the model do not externalize well.

0.3 Room reverberation and range

In order to get rid of the internalization we add the room model. It is well known that simulated room reverberation produces an externalization effect**.**

The monaural input is delayed by an amount TE and mixed with the outputs of the head model. The ratio of direct to reverberant energy is adjusted by the ratio of the channel gains KL and KR to the "echo gain" KE. Unless the source is close to being directly ahead or directly behind, it usually seems externalized when TE was around 15 ms and the echo gain was 15 dB below the channel gains.

The range model just accounts for the amplitude of the audio signal. The amplitude drops inversely with distance. This simulates the distance of the source.

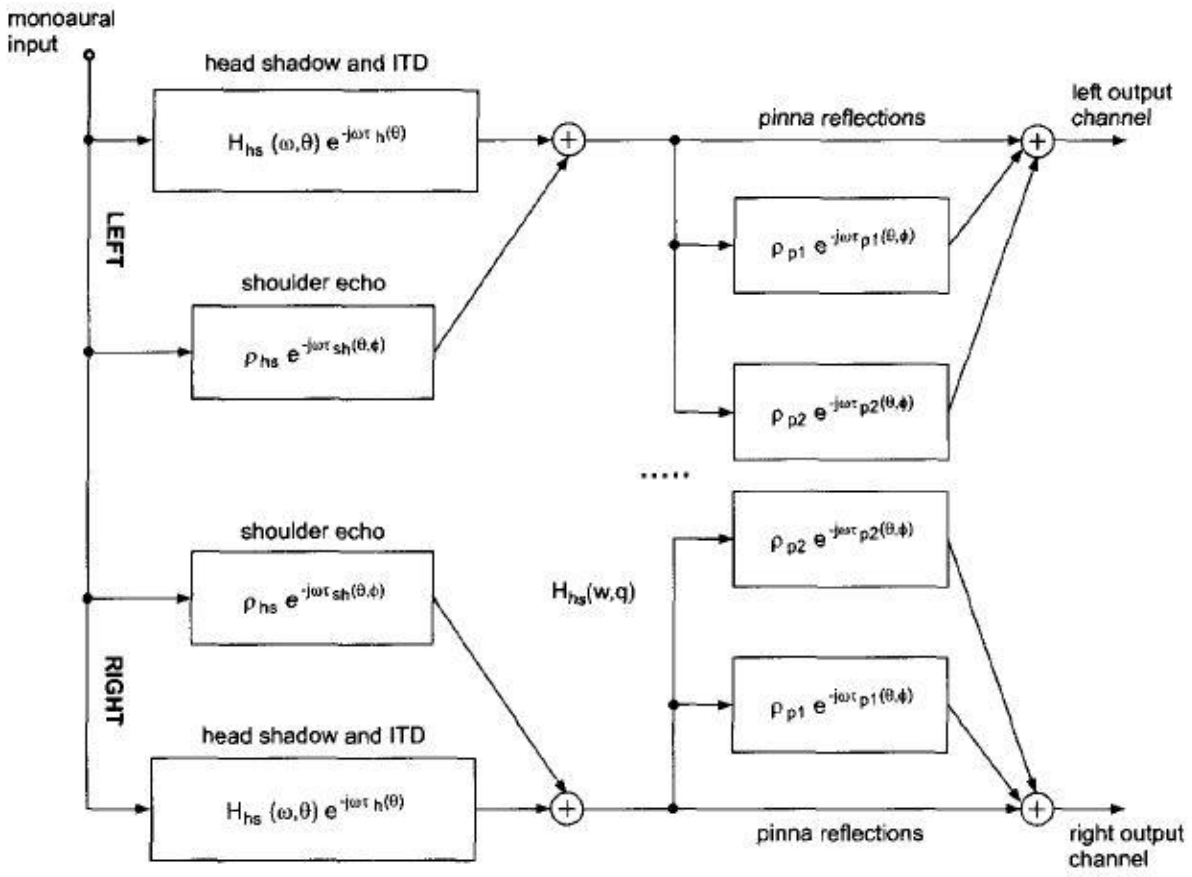


Figure 1: Structural model of the pinna-head-torso system.

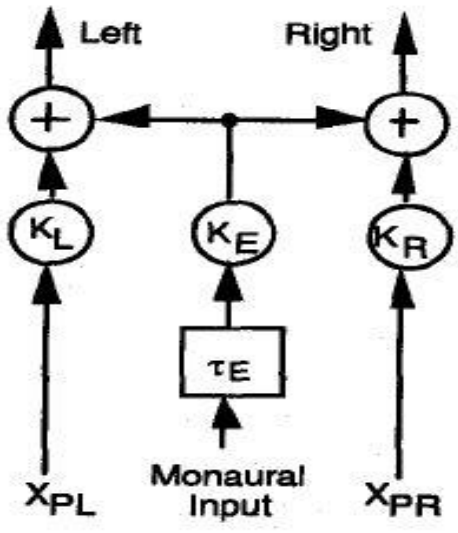


Figure 2 : The room model

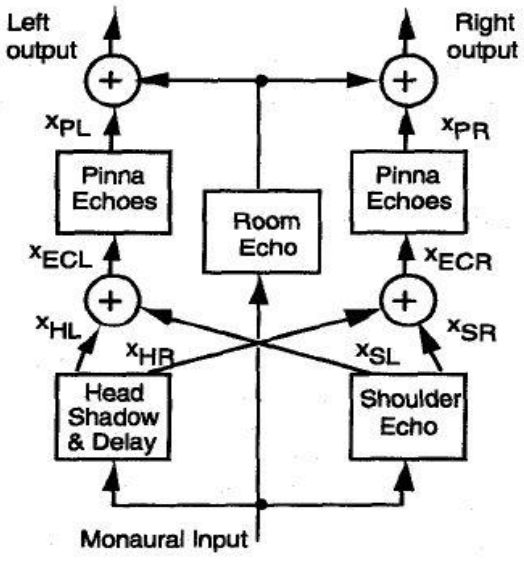


Figure 3 : Components of the entire model.

Loudspeaker playback

If binaural audio material is played through a stereo loudspeaker layout then almost any spatial effect is likely to disappear. Assume that we want to recreate, by means of a couple of loudspeakers, the signals as they arrive to the ears from a couple of headphones. Delta dash (symbol read as) is the distance between the ears in spatial samples, i.e., the distance in meters multiplied by fs /c**,** where c is the speed of sound and fs is the sample rate, D is the distance in samples between a loudspeaker and the nearest ear, Theta (symbol read as) is the angle subtended by a loudspeaker with the median plane. Under the assumption of point like loudspeakers, the excess distance from a loudspeaker to the contra-lateral ear produces an attenuation that can be approximated by

$$g \cong \frac{D}{\bar{\delta} \sin\theta + D}$$

The head of the listener introduces a shadowing effect which can be expressed by a low pass transfer function H (z). These considerations lead to the following matrix relationship between the signals at the ears (e1 (z), e2 (z)) and the signals at the loudspeakers (L (z), R (z))

$$\begin{bmatrix} e_1(z) \\ e_2(z) \end{bmatrix} = \begin{bmatrix} 1 & gH(z)z^{-d} \\ gH(z)z^{-d} & 1 \end{bmatrix} \begin{bmatrix} L(z) \\ R(z) \end{bmatrix} \triangleq \mathbf{A}(z) \begin{bmatrix} L(z) \\ R(z) \end{bmatrix}$$

where d is the arrival time difference in samples of the signals at the ears. We should consider d a function of frequency. Generally,
d = 1.5 * delta dash * sin(theta). The symmetric matrix A can be easily inverted, thus giving

$$\mathbf{A}^{-1}(z) = \frac{1}{1 - g^2 H^2(z) z^{-2d}} \begin{bmatrix} 1 & -gH(z)z^{-d} \\ -gH(z)z^{-d} & 1 \end{bmatrix}.$$

The conversion from binaural to loudspeaker output is realized by the 2-input 2-output system with two functional blocks: T (z) is a lattice section, and C(z) is a comb filter with a low pass in the feedback loop.

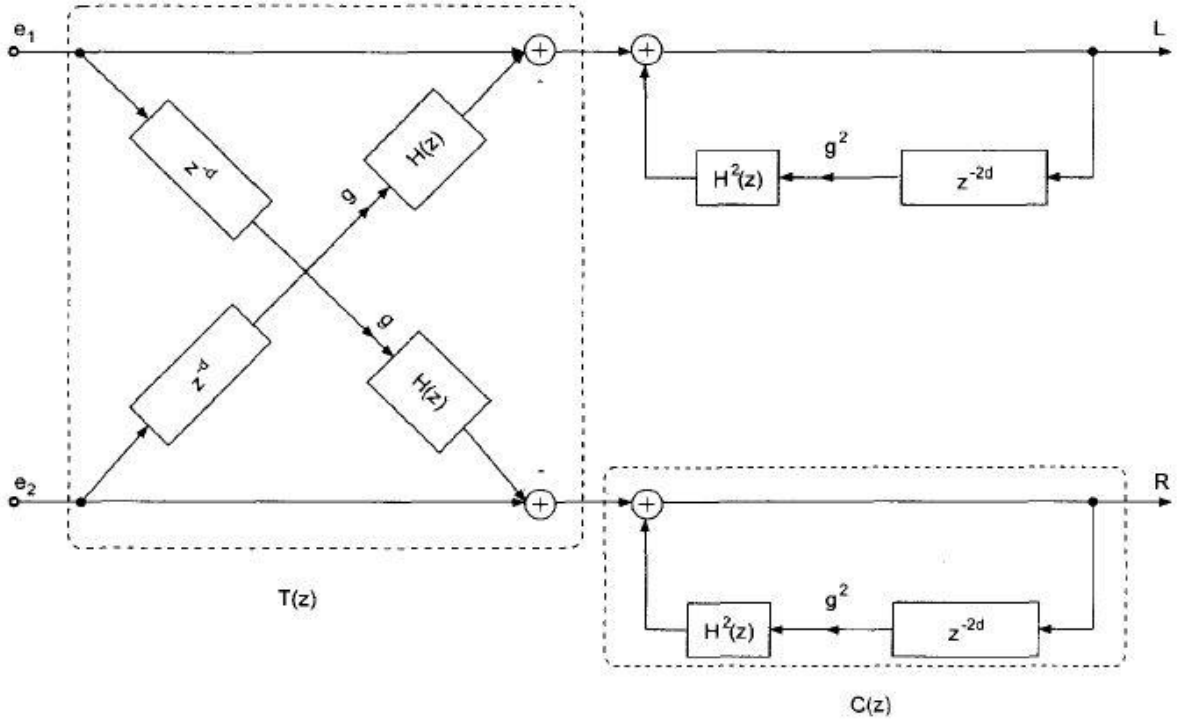


Figure 1 : T(z) – Lattice section, C(z) – Comb filter

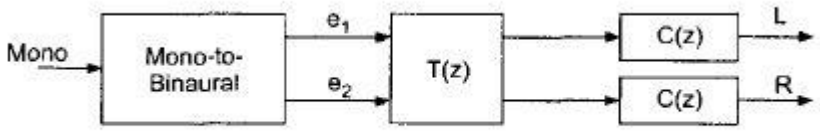
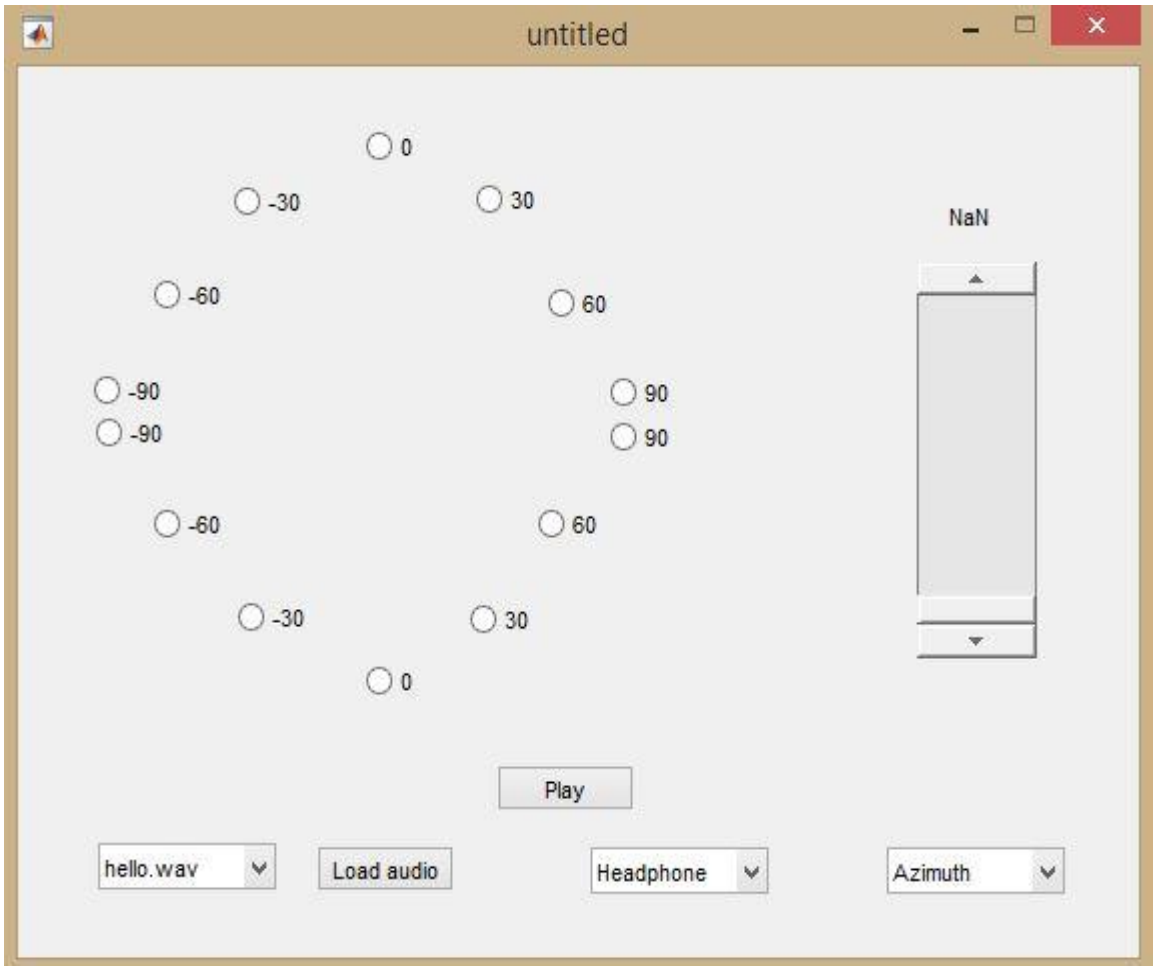


Figure 2 : A block diagram for mono to binaural to transaural

Implementation and Results

The platform chosen for implementation was Matlab. A Matlab GUI was created and was interfaced with various other functions built using Matlab DSP toolbox. The GUI looks like this



hello.wav is a mono channel 44.1kHz recording made on Audacity. This is the test signal which we have used on all the subjects. There are three options for playback – Headphone, Loudspeaker with cross talk cancellation, Loudspeaker without cross talk cancellation. The available test signals are Azimuth (angle can be chosen from the circular layout), Azimuthal sweep, Elevation (angle can be chosen from the slider on the right side), Elevation sweep and Range sweep. The headphones and loudspeakers used for output were stereo based.

The functions on which the GUI was built were-

- Head shadow and time delay filter function
- Shoulder delay, pinna reflections, room model
- Crosstalk cancellation
- Comb filter with low pass filter in the feedback loop
- Headphone playback
- Loudspeaker playback

These functions do the same task as explained earlier.

When subject to listening tests,

- Azimuthal variation was very accurate in headphone playback ranging from -90 to 90. The acoustic image is always formed behind and not in front. Variation of +- 10 degree was also noticeable.

- Elevation variation wasn't very good in headphone playback. When the listener was told that the sound was coming from a particular altitude, he answered with affirmative and agreed that the sound was indeed coming from that altitude. But when asked to figure out the elevation on his own, there was confusion. Also the Dn parameters in the pinna reflections weren't tuned from person to person. It was kept constant for all subjects.

- Upon adding the room model for the headphone playback, the sound source appeared external to the head. The Azimuthal sweep was very smooth and the experience was better than the same test signal without room model.

- The range sweep did create an impression on the listener that the source was moving away or coming close. But some subjects felt that it sounded like the volume/amplitude was decreased, that is, it sounded too artificial. Also the range's effect on room model was not included which might have made the sweep sound more realistic.

- Azimuthal variation on loudspeaker without cross talk cancellation did not provide accurate results compared to the headphone playback. With crosstalk cancellation provided results very close to headphone playback and the acoustic image was formed in front of the person only.

Conclusion and Future Scope

- On compilation of the results of the listening tests it was observed that for loudspeaker playback if the two speakers are very close to each other then crosstalk cancellation wasn't essential, although the output with cancellation produced better results.
- Elevation variation cannot be easily achieved by just modelling pinna and shoulder.

Azimuthal variation was perfect. Elevation variation was ambiguous.
On the whole, this solution of modelling human body parts and using these as filters is cheap and easier compared to HRTFs. This can also be easily implementable on FPGA or generic DSP processor.

The future scope of this could be

- For gaming purposes. 3D audio can enhance gaming standards0
- For humanitarian purposes. 3D audio can be used to relay messages to a blind person about general things around him so that it becomes easier for him to access.
- For concert purposes. 3D audio is used to indicate that the instruments on one side of the concert stage are heard on one channel and hence we can locate the instrument in the acoustic image field easily.
- For video conferencing purposes. 3D audio can be used to connect two table conferences over video. The people's position around the table can be distinguished by the people on the other table by just listening to the spatial audio alone.

Bibliography

1. DAFX – Digital Audio Effects, *John Wiley and Sons,LTD*
2. *A Structural Model for Binaural Sound Synthesis*, C. Phillip Brown and Richard O. Duda, IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 6, NO. 5, SEPTEMBER 1998
3. *Modeling Head Related Transfer Functions*, Richard 0. Duda, Preprint for the Twenty-Seventh Asilomar Conference on Signals, Systems & Computers (Asilomar, CA, October 31-November 3, 1993)
4. *An Efficient HRTF Model for 3D Sound*, C. Phillip Brown and Richard O. Duda