RUDIMENTS OF SPATIAL AUDIO SYNTHESIS

A. Girbau, C. Nadeu

Universitat Politècnica de Catalunya
Teoria del Senyal i Comunicacions
andreu.girbau@upc.edu - climent.nadeu@upc.edu

ABSTRACT

For many application areas the binaural synthesis has become a field of interest. In this paper, we present the basics of binaural synthesis for 2 channels -left and right- providing examples and figures, using different approaches: from just delaying both signals to the usage of **HRIR** (*Head-Related Impulse Response*) and **HRTF** (*Head-Related Transfer Function*), and a way to relate spatial information with temporal information of an audio signal.

Index Terms— Binaural synthesis, HRIR, HRTF, FIR, Spatialization.

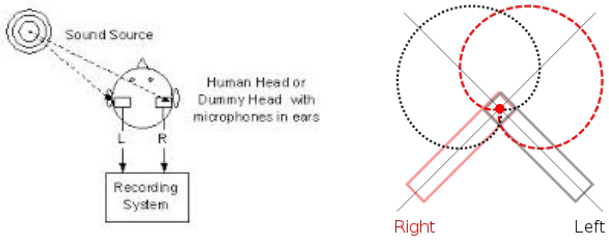
1. INTRODUCTION

Nowadays we can see many loudspeaker distributions used to achieve certain spatial sensation of the sound. From the simple stereo to the 22.2 distribution, we are used to see the need of having large -or not that large- amount of loudspeakers. The virtual sound spatialization can be used to improve the presence of virtual fonts in a virtual environment, embedding the information in a two channel standard headphone.

We are interested on obtaining HRIR (HRTF in frequency domain) responses from different loudspeaker positioning. This will allow us to place *virtually* a sound source [1], so that the listener will perceive the sound as it was placed somewhere in space, instead of the standard stereo that headphones can provide. The environment we use to produce the modified signals is Matlab.

2. BINAURAL SYSTEMS

Binaural definition is: *Relating to or involving both ears* or *Relating to sound recorded using two microphones and usually transmitted separately to the two ears of the listener*. The second definition could lead someone to think that binaural and stereo recordings are the same. The main difference between binaural and stereo is that binaural systems take in account the physical effects of the pinna and the head/chest filtering (or shadowing), and the stereo recording systems do not take in account these factors. Our application is interested on the binaural recording for the fact that it, theoretically, gives a better spatial audio sensation to the listener. How we are going to provide this sensation to the listener is described in section 3.



Binaural recording Stereo recording

Fig. 1. Example of binaural recording -left - and stereo recording -right-.

3. BINAURAL RECORDING AND RESPONSES DESCRIPTION

This section aims to explain the way to obtain the HRIR (HRTF in frequency domain) responses and how to use them for modifying a signal in order to give the spatial positioning sensation to the listener.

HRIR responses are obtained using a set of loudspeakers positioned in a certain azimuth and certain height. This means that, using binaural recording, the position and height information of the loudspeaker will be inside the HRIR response. A point to remark here is that, as a specific person with certain physical characteristics is the one used to record the head responses, some users may not have the same clear sensation than others, who are more like the person used to record the responses. HRTFs are obtained using the DFT of the HRIRs. This project uses the IRCAM *listen* HRTF database. `http://recherche.ircam.fr`.

We can distinguish 2 HRIR responses for each loudspeaker position: left and right response. We obtain the HRTF by simply computing the FFT of each HRIR response. The azimuth

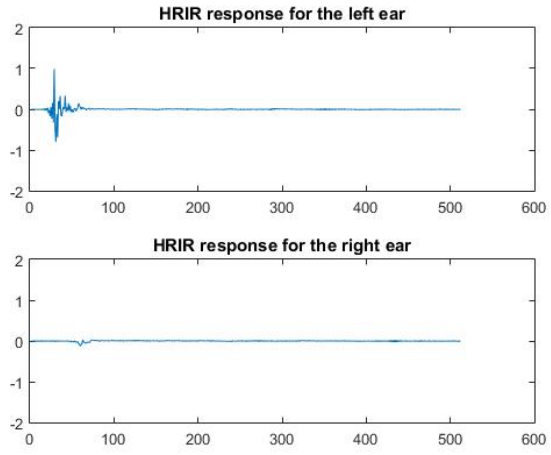


Fig. 2. HRIR responses corresponding to an azimuth of 105° and an elevation of 15°

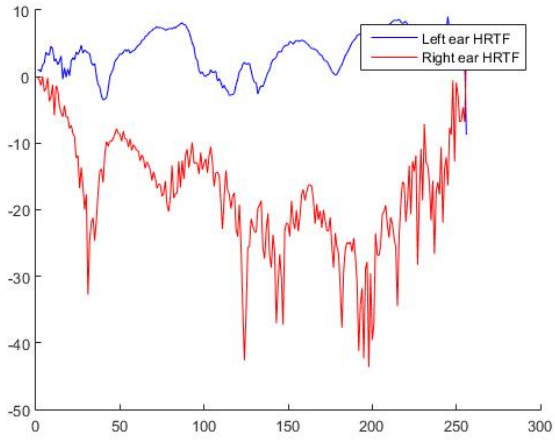


Fig. 3. HRTF responses corresponding to an azimuth of 105° and an elevation of 15° in log scale

is defined as ϕ =0° in the direction where the listener facing and increases counter-clock wise. The height is defined as θ =0° where both, ears and loudspeaker, are at 0° respect to each other. In our case, ϕ is defined from 0° to 345° in steps of 15° and θ is defined from $-45°$ to 90° in steps of 15°.

4. FIRST APPROACH: DELAY AND AMPLITUDE

To understand what we aim to achieve, we have to take in account the physical characteristics of a human head. For us, the 2 ears will be the input where the output of our characterization will have to go. This means that we will have to characterize a channel response to produce the spatial effect we desire for any input signal.

4.1. Delay between signals

We realize that the signal that arrives to the left ear and the right ear carry the same information but it is not the same signal produced by the audio source. If we think on a signal produced by this source positioned on the left side respect to our head, it is straightforward to think that this signal will arrive first to our left ear and, after a few microseconds -as it travels at the speed of sound-, to the right ear.

Considering this, we only will have to convolute the source signal for each *left* and *right* channel response. In this case, the channel responses will be $\delta(t - t_L)$ for the left channel and $\delta(t - t_R)$ for the right one.

4.2. Signals amplitude

To model the attenuation of the signal between ears we can simply multiply a constant for each channel response. In this case, $\delta(t - t_L) \cdot A_L$ for the left channel response and $\delta(t - t_R) \cdot A_R$ for the right one.

So, a first approach of what a our spatial synthesizer is, can be seen as:

$$z_L(t) = s_L(t) * \delta(t - t_L) \cdot A_L$$
$$z_R(t) = s_R(t) * \delta(t - t_R) \cdot A_R$$

where $s_L(t)$ is the left source channel and $s_R(t)$ is the right one, and $z_R(t)$ is the left output and $z_R(t)$ is the right one.

In section 5 we will have both -delay and amplitude- information -among other possible characterization of the channel- embedded into each HRIR.

5. SECOND APPROACH: USING HRIR AND HRTF

First of all we define the input signal to treat. As we work with stereo signals, we'll have 2 different channels. We'll define them as $s_L(t)$ and $s_R(t)$ as the left and right original input channels. The HRIR responses will be, also, for left and right channel, and they will be defined as $h_L(t)$ and $h_R(t)$.

The most simple way to combine the information of HRIR responses and the signal is the convolution of each channel signal with each channel response. The resulting signal will be denoted as $z_L(t)$ for the left channel and $z_R(t)$ for the right one.

$$z_L(t) = s_L(t) * h_L(t) \quad (1)$$
$$z_R(t) = s_R(t) * h_R(t) \quad (2)$$

As we work in time samples, the notation from now on will use n instead of t. *Example:* $s_L[n]$ *instead of* $s_L(t)$. We

use $fs = 44100Hz$ for the sampling frequency to be consequent with the sampling frequency of the HRIR database. Also, HRIR database uses a constant window of 512 samples for every response.

In (1) and (2), the whole $s[n]$ is convolved by the same HRIR [2], this means that only one virtual loudspeaker is being used. To use more than one during a sequence we use HRTF and FFT -as the convolution in time is a multiplication in frequency -.For simplicity we'll use the left signal $s_L[n]$ from now on, but the same technique is applied to $s_R[n]$ as well. The main idea is to do some windowing (3) of the original signal and, for each window -assuming L windows-, compute the FFT of both windowed signal (4) and HRIR (5), and multiply them in the frequency domain (6). After this, use the IFFT to get back to the time space (7) and sum up all the windows to recompose the signal (8). It is important to take in account that the number of samples of the FFT has to be greater than the number of samples of the $s[n]$ window plus the number of samples of $h[n]$ minus 1 in order to not distort the signal.

$$s_L^{(W)}[n] = s_L[n] \cdot window[n] \tag{3}$$

$$S_L^{(W)}[k] = FFT\{s_L^{(W)}[n]\} \tag{4}$$

$$H_L[k] = FFT\{h_L[n]\} \tag{5}$$

$$Z_L^{(W')}[k] = S_L^{(W)}[k] \cdot H_L[k] \tag{6}$$

$$z_L^{(W')}[n] = IFFT\{Z_L^{(W')}[k]\} \tag{7}$$

$$z_L[n] = \sum_{w'=1}^{L} z_L^{(w')}[n] \tag{8}$$

By using this, we are able to compose an output signal $z_L[n]$ that contains more than 1 virtual loudspeaker information.

6. SPATIAL AND TEMPORAL SAMPLING

As we aim to work on a virtual audio space, we would like to relate this virtual space with the information we have -$s(t)$ and $h(t)$ returning to *time notation*-. The signal $s(t)$ is the input to our system, so we will not modify it or its notation. Instead, we are going to focus on relating the HRIR $h(t)$ to every allowed position in space. This position is, as mentioned in section 3, related with θ and ϕ. Then, every spatial position will have a HRIR associated. These positions will be denoted as *space samples*. With this idea in mind, we are able to rewrite the head responses as $h_L(t, [\theta, \phi])$ and $h_R(t, [\theta, \phi])$ for the left and right channel respectively. This notation means that the HRIR are time dependant and space dependant.

As we already stated in section 3, we sample the space in steps of $15°$, so the spatial sampling becomes a natural analogy of it. Each space sample will have the time values of the correspondent related HRIR $h(t, [\theta, \phi])$. With this information, we are able to relate a virtual object position with time, and this can be extended to object movement.

When we talk about object movement, we are referring to the fact that some part of the signal $s(t)$ is convolved with an $h(t, [\theta, \phi])$ -it is *located* in our virtual environment- and the response $h(t, [\theta, \phi])$ is changed to $h(t, [\theta', \phi'])$. Then we assume that the source has moved from $[\theta, \phi]$ to $[\theta', \phi]'$.

In our experiments, one of the sound sources was an audio file of an helicopter. Using this example, we could hear the sound of the helicopter being modified for every $h(t, [\theta, \phi])$. The example consisted on having a constant height and an increasing azimuth for every specified time interval Δt. The sensation was to have an actual helicopter spinning around our head.

To put this example on a mathematical way, we define $w(t) = \left(\frac{\partial \phi}{\partial t}\right)$ as the angular velocity. As we have to do some windowing to the signal, Nw will be the number of windows for that signal portion $s_p(t)$. We will assume constant speed for every signal portion p. Remember that every spatial window is of $15°$. Nw can be seen as *how many spatial samples does it take to move from the first position to the ending one*. *E.g. to move from $\phi = 0$ to $\phi = \pi/12$ it would take Nw = 1 sample, to move from $\phi = 0$ to $\phi = 2\pi$ it would take Nw = 24 samples*.

$$\begin{aligned} w(t) &= \left(\frac{\Delta \phi}{\Delta t}\right) \\ &= \left(\frac{\pi/12 \cdot Nw}{t_{signalPortion}}\right) \end{aligned}$$

In our case, we experimented with a $t_{signalPortion}$ of 10 seconds. Nw can be controlled by means of $Nw = \frac{t_{signalPortion}}{t_{window}/2}$ This means that what is left is the length of the windows at the temporal windowing of the signal portion is proportional to Nw in spatial domain. So, to conclude this experiment, we designed 3 cases:

1. From $\phi = 0$ to $\phi = \pi/12$: $w(t) = 0.026rad/s$
2. From $\phi = 0$ to $\phi = 2\pi$: $w(t) = 0.628rad/s = 0.1rps$
3. From $\phi = 0$ to $\phi = 10 \cdot 2\pi$: $w(t) = 6.28rad/s = 1rps$

7. CONCLUSIONS AND FURTHER RESEARCH

This article is made to give some insights of the binaural spatial synthesis. The conclusions of this work are that there are ways to position an audio source in a virtual environment and they are not difficult to implement and their computational cost is not very high. If the modified signals are embedded into a 2 channel standard stereo headphones, the results are pretty good (depending on your physical characteristics too). The main drawback of this technique is this dependence to the listener physical shape, as we tested with several people we observed this behavior.

The further research will consist on four main things:

- Usage of more than one HRIR. Techniques to use more than one virtual sound source and the effects on the output signal.
- Real time signal modification.
- Smoothing the transitions between two virtual sound sources that are far away (e.g. from $\phi = 0$ to $\phi = pi$ without passing through all the others).
- HRIR synthesis. Is there a way to produce a synthetic HRIR relating some physical parameters of the listener?

8. REFERENCES

[1] J. Herre, Kuntz A. Hilpert, J., and J. Plogsties, "Mpegh audio - the new standard for universal spatial/3d audio coding," .

[2] V.N Ganesh, "Implementation of 3d audio effects using head related transfer function (hrtf) for real time application using blackfin processor," in *Recent Trends in Signal Processing, Image Processing and VLSI, ICrtSIV*. MIT, 2014, pp. 361–267.