Multichannel Mobile Listening: A State Of The Art

Jean Cruypenynck

May 2014

Abstract

Nowadays, the multiplication of audiovisual contents media such as computers, tablets and smartphones has created a major change in our behavior of content consumption. Broadcasting companies have to adapt their contents to fit our habits.

Binaural and transaural listening are parts of this conversion to mobility.

Introduction

Even though binaural recording goes back to 1881 and Clément Adler's *Théâtrophone*, it is currently expanding. Major broadcasting companies are all thinking about binaural reproduction for a mobile audience.

In fact, we are approaching hyperrealism in increasing the numbers of audio channels and engaging from 2D to 3D in video. These improvements are set up to immerse the audience in the action.

After becoming stereophonic, the sound became multichannel for the first time in 1937 with *Fantasia* from Walt Disney in quadriphony. Since this moment, playback has passed by 5.1, 7.1, finally coming to *Dolby Atmos* and *Auro 3D*. If 5.1 has now become a standard, there is no doubt about the interest of public in obtaining more and more sound channels. But, having 64 channels as in *Dolby Atmos* in our room will not became the new standard. Here is the role played by binaural and transaural listening.

1 Prerequisites

This section presents some of the key concepts in mobile multichannel listening.

1.1 Convolution

As will be explained in next section, convolution is at the heart of both binaural and transaural listening.

Dating from 18^{th} century, convolution is now a widely-used operation in digital signal processing. If you have never been enlightened as to the joys of convolution, here you can find a short recap of this operation.

Definition. *Let us assume that f and g are two functions of a time variable t. So, the convolution product of f by g is defined by:*

$$f(t) * g(t) = \int_{-\infty}^{+\infty} f(t)\, g(x-t)\, \mathrm{d}t$$

The impulse response of a linear, continuous-time, time-invariant system (afterwards called LTI systems) is defined by the output of the system when it is submitted to a Dirac distribution.

The Dirac distribution is defined as a distribution equal to zero for all t except zero, equal to $+\infty$ for $t = 0$ and whose integral is equal to 1.

Knowing the impulse response of a LTI system, we can can convolve it with an input signal to reproduce the behavior of the system. We can also predict the output of this system for a specific input.

An interesting fact in convolution is that the convolution product of two temporal functions is equal to the inverse Fourier transform of the simple product of the two frequency functions

or, in other means the Fourier transform of our two temporal functions.

It leads to the following equation:

$$\mathcal{F}\left[f\left(t\right) * g\left(t\right)\right] = F\left(\nu\right) \times G\left(\nu\right)$$

considering respectively F and G as the Fourier transform of f and g.

Without being a main fact of the theorical part, this tip is important in the implementation of convolution, especially in terms of the CPU load.

1.2 The Listening Process

The location of sound sources depends on three components:

- interaural time difference;
- interaural level difference;
- spectral indices, which are the characteristics of our head (skin, hair, beard...).

In fact the acoustic pressure of *localized* sound source, meaning by this a sound source not placed in the median plan, arrives in a first ear at a certain instant t with a level A and in the second ear at the instant $t + \delta t$ at an amplitude of $A - \delta A$. This phenomenon is better shown in figure 1.

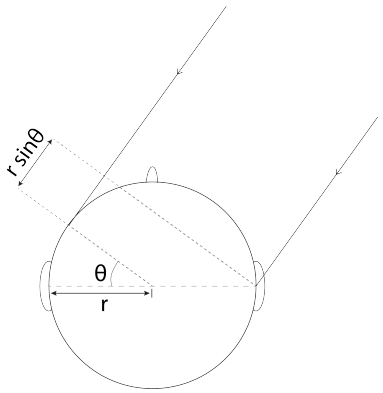

Figure 1: Interaural Time Difference

According to the figure, the interaural time difference is equal to $r\left(\sin\theta + \theta\right)$. The interaural level difference would be equal to the δA mentionned before.

2 About Binaural

Binaural sound has the ambition to reproduce a three-dimensionnal sound field with two in-ear audio channels only. This technology, although seeming innovative is not new and is based on simple physical principles.

Acquisition of the BRIR The principle is to record the result of a known test signal reproduced by a spatialized sound source in each ear. Little electret condenser microphones placed in ears are often used to achieve these measurements. After being recorded, the signal is deconvoluted in an impulse response.
The group of impulse responses obtained by repeating the operation for each channel of a multichannel system is known as *Brain Related Impulse Response* (BRIR). From BRIR, we can easily extract the *Head Related Transfer Function* (HRTF), in fact in going in frequency domain by Fourier transform.

Use of the BRIR When BRIR is obtained, each channel of a signal is convoluted with its corresponding left and right impulse responses, as shown on figure 2. The output signal is therefore a stereophonic signal reproducing a multichannel system.

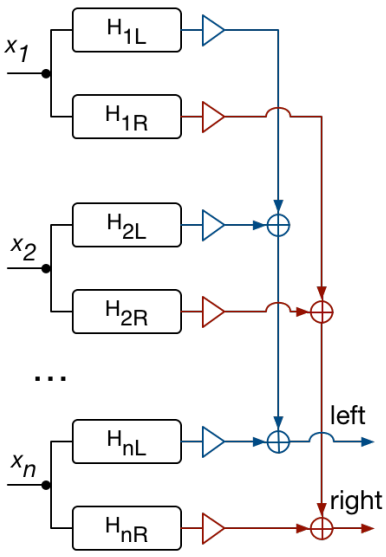

Figure 2: Multiple source binaural spatializer

Main issue The main issue in binaural listening is the unique nature of each BRIR. In fact, the BRIR depends on characteristics of our head. Sound sources localization can be very altered in a binaural signal made with a generic BRIR.

Test signals Test signals used to measure BRIR are essentially the same as in reverb impulse response measurements. Nowadays, it appears that the exponential sine sweep signal still seems the best in terms of fidelity and signal-to-noise ratio. This signal is defined by:

$$x(t) = \sin\left[\frac{\omega_1 T}{\ln\left(\frac{\omega_1}{\omega_2}\right)}\left(\mathrm{e}^{\frac{t}{T}\ln\left(\frac{\omega_2}{\omega_1}\right)} - 1\right)\right]$$

3 About Transaural

The founding principles of transaural are the same as binaural. The difference between both is in the playback means. In the case of transaural, we use external speakers while in binaural we use headphones.

In concrete terms, transaural listening reproduces multichannel listening in a specific room with two speakers, using the transfer function of the room.

This is achieved with cross-talk cancelling. Traditionnally, when listening with speakers, the left signal goes to the left ear and a little bit to the right ear and conversely. In transaural, we filter left and right signals to properly separate the signals. After obtaining the cross-talk cancellation function, it only remains to put binaural signals in input and the signal at the ears will be spatialized.

In transaural, the main problem is the unique nature of the transfer functions of the rooms.

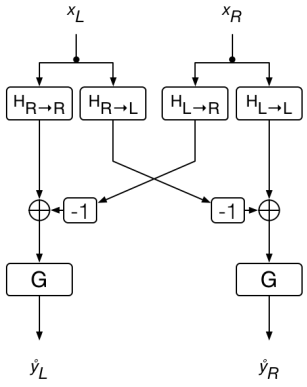

Figure 3: General transaural filter, where $G = \frac{1}{H_{LL}H_{RR} - HLRH_{RL}}$

4 Possible Future Evolutions

Without mating wild guesses, we already can think about believable audio innovations to revolutionise the current situation.

Recognition enhancements Firstly, recognition of spectral indices and rooms. It would put aside the uniqueness of:

- each BRIR;
- rooms in which transaural content would be listened to.

For the BRIR, two ways seem to emerge:

- Using different morphotypes to categorize most of the population.
  About ten morphotypes appear to be enough; the user would be asked to describe their face and the software would build a specific IR adapted to them.
- Using pictures to create a 3D map of the subject's face and design their BRIR. *Thirdlove* company has already launched a mobile application for women to find their real bra size with only two pictures. With maybe, let's say, five pictures of our face, it would be possible to get an accurate 3D map.

Google is currently developping a technology called *Tango*: it makes it possible to create a 3D map of our environment from our smartphone or tablet. Combined to the accelerometer and gyroscope which have already been present for a couple of years, we could get a tablet which adapts the sound output according to the position in space to create a transaural result. Of course, there would be other problems in this case like lack of acoustical power and informations about room materials.

Access enhancements We could think about a cloud platform like *Dropbox* containing our personal BRIR. When taking the tablet of a friend, we would just have to log into our account and access our BRIR to listen to multichannel content on their device.

In terms of production and broadcast, it would be possible to send the user an MPEG Surround stream which would be decoded into binaural content if headphones are plugged in, into transaural content if using the speakers of a mobile device, or into classical multichannel content matching the actual configuration of the user, *ie.* 7.1 to 5.1.

Algorithm enhancements Finally, said that the impulse response of a LTI system makes it possible to oversee the output of this system knowing the input. But our listening process can not really be assimilated to a linear system. So we can also imagine an enhancement of the binaural/transaural transcoding in this way: nonlinear convolution and Volterra series could be the basis of this tweak. These operations will not be seen in this paper but I invite the reader to find out more about them.

Conclusion

Technology allows us to achieve more and more complex processes and we have to take advantage of that. Now that our computers are able to process real-time convolutions, there just remains the issue of personalized BRIR and rooms.

For the first, it would be possible to create enough categories to fit the vast majority of users or to determine each BRIR via 3D modelisation based on pictures of our face.

For the second, we have to use waves to recreate the architecture of the room and use motion sensors and near field information to know the exact place of the user.

Although both binaural and transaural listening seem far from us, I think the two methods are a future logical implementation. Binaural personalisation does not seem too hard to create whereas the transaural modelisation of a room would take more years to develop.

References