SPATIAL LISTENING AND ITS COMPUTER SIMULATION ON ELECTRONIC MUSIC

Prof. Oscar Pablo Di Liscia

Universidad Nacional de Quilmes

odiliscia@unq.edu.ar

SPATIAL LISTENING

When we are interested on the spatial quality of sound, our Auditory System inspects an incoming acoustic signal trying to answer two main questions:

1-Where am I? (information related to the room or environment)

2-Where is it? (information related to the location and / or movement of a sound source)

To do this, our Auditory System uses ***cues***.

I will use the word ***cues*** to name those special features of the sound which are regarded by our Auditory System as delivering relevant information.

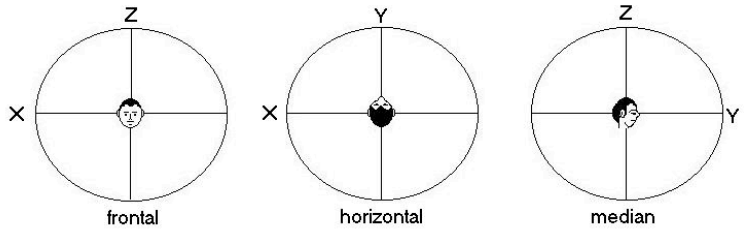
Because not all features of the incoming acoustic signal have useful information to answer the former questions, our Auditory System takes account only of those cues that seem to be most reasonable for each case.

Also, this information is combined with other kind of information coming from our other senses (mainly our view), and with our knowledge of the behavior of the sound source. That is, our perception of space is ***holistic***.

There are several ways to deal with a three-dimensional space. The most used is to think it as formed by three planes, each being different "slices" of a sphere.

The three planes are: frontal, horizontal and median.

Figure 1: The three spherical planes.



To refer to any point at the theoretical sphere, either polar or rectangular coordinates can be used. For a three dimensional space, three rectangular coordinates are needed: X (left-right), Y (front-rear) and Z (above-below). When using polar coordinates, for a three dimensional space, two angles (azimuth and elevation) and a magnitude vector (distance) are needed. When dealing with angles I will consider them being "0" degrees located at the right of the origin (counter clockwise, for the azimuth), and being "0" degrees located directly in front of the origin (for the elevation).

You can see at figure 1 that the listener is always facing the ***positive Y***. In some systems this is different (e.g., British recording engineers use to exchange Y and X…).

Cues used to judge spatial location

May be classified in three main groups:

1-Related to the horizontal plane (azimuth)

2-Related to the median plane.

3-Related to distance.

1-Horizontal plane:

1.a **ITD** (Interaural Time Difference): The difference in arrival time of the signal at our different ears due to the location of the sound source at an horizontal angle other than 90 degrees and 270 degrees. Seems to be effective at frequencies below 900 Hz.

1.b **ILD** (Interaural Level Difference): The difference in level of the signal at our different ears. Seems to be effective at frequencies above 500 Hz. Lower frequencies have a wavelength larger than our head, thus diffracting it.

It should be pointed out that we meet our best angle of discrimination on the horizontal plane, of course, when we face the sound source. We loose accuracy as this angle goes to both sides of our head (that is, from 90 degrees to 0 degrees –right- and from 90 degrees to 180 degrees –left-). That's why we turn our head when we think that the source is located at any of our head sides, attempting to get a better sound image of the sound source and checking whether the produced image is coincident with the first one.

2-Median plane:

HRTF (Head Related Transfer Functions): when the source is located on this plane, the cues rendered by the ITD and ILD are the same for both ears, so it seems that the complex effect of filtering due to the convolution of the signal with the shape of our upper torso, neck, head, and external ears is most responsible for rendering data for location. It should be noted that the information delivered is useful both to determine the elevation of the source and to know whether it is behind or in front of us. It is also true that this information is different when the horizontal angle changes, therefore reinforcing the ITD and ILD when appropriate. The measured spectral difference is called HRTF.

3-Distance:

3.a **Global loudness of the sound**: despite of its apparent simplicity the loudness of sound is a weak cue to judge the distance between source and listener. Physically speaking, it is well known that the acoustic energy drops proportionally to the square of distance. But handling only amplitude will not render the fair

impression of loudness, since loudness is not linearly related to amplitude and is also strongly dependent on our knowledge of the behavior of the sound sources involved. As a matter of fact, we do not think that the sound produced by a person whispering close to us is louder than the sound produced by a person screaming far away, even when the signal of the former were larger in amplitude than the latter.

3.b **Ratio between reverberated and dry signal**. In closed rooms, the energy of dense reverberation will remain more or less constant (in average amplitude, and for a source delivering the same energy) while –if the distance changes- the energy of the direct (dry) signal will drop with distance. This seems to be the main responsible cue for our judgment on distance on reverberant environments.

3.c **Absorption of high frequencies** due to gasses in the air. This effect is similar to a lowpass filter, and is considered relevant only for distances larger than 50 meters.

Cues used to judge the features of a room

1-**Reveberation time**: also called T60, that is, the time it takes for the reverberant signal to drop by 60 dB (.001 of its amplitude) from the time at which the direct signal is over.

2-**Spectral balance**: rooms act as filters, modifying the spectrum of the acoustic signals. The spectrum of the impulse response of a "well behaved" room must be flat enough on the full audio band.

3-**Diffusion**: due to absorption of materials that cover the floor, ceiling and walls, high frequencies tend to drop more rapidly than low ones.

4-**Interaural Decorrelation**: how similar is the reverberation incoming to the different ears of the listeners. Though reverberation is fairly the same regarding amplitude and density for both ears, it is slightly not the same for a given moment, specially concerning its spectrum. In real rooms, listeners experiment a "spatial quality enhancement" when the reverberation is slightly decorrelated.

5-**Time between direct and first relevant echo**: this helps us to guess how big may the room is.

Other Cues influencing spatial listening

1-**Doppler effect**: it is a powerful cue for judging movement quality and relative velocity of source and listener. That is, by the Doppler shift we are able to know whether an object is moving, and if it is moving in our direction or away from us. It should be noted that our auditory system is not particularly able to track location of a moving sound source, specially when the source is moving very fast. Some experiments have shown that, at an angular velocity of 15 degrees/sec., we are able to discriminate steps of 5 degrees of difference, but at 90 degrees/sec. of angular velocity even steps of about 20 degrees may be lost.

2-**Precedence effect**: also called the Haas effect, or also "*the law of the first wavefront*". It states that the first wavefront that arrives at our ears will be considered the direct signal by our Auditory System. This helps us to avoid confusion between early echoes and direct. Of course, we may be fooled by loudspeakers…!!!

3-**Directional characteristics of the sound source**. A sound source radiating equal energy to all directions is considered an Omni-directional source. This is an ideal situation, because all natural sound sources do not behave *exactly* like that. Thus, a (non omni-directional) source pointing at the listener will not radiate the same amount of energy than when pointing at other location. This will affect both the direct sound and reverberation.

COMPUTER SIMULATION

When working in electronic music, ***spatial enhancement*** of sound must not be confused with ***location*** of a sound source. Composers sometimes want a strong effect of spatial environment (e.g. to give the impression that the source is in a big church…) while not being concerned with its precise location into the environment. On the other hand, some composers attempt to get a strong impression of location and/or movement of a sound source and are not mainly concerned with the kind of environment the source may be in. As a matter of fact, there are strong room cues that –under some conditions- may obscure a clear impression of sound location (e.g., too much dense reverberation). Furthermore, ***location*** is not the same as ***movement***, because we may guess (e.g., by the Doppler shift) that a source is moving, while we may not be able to know its precise location at any point of its path.

Not always the most accurate simulation of the real situation yields the most effective impression of location and/or sound source movement, because while some features of sound improve our sensation of location and/or movement, other may obscure it. So, there seems not to be a perfect solution (though, of course, there are some which are better and particularly elegant).

The main resource to simulate location of a sound source is to try to fool our auditory system by creating *phantom sources*. Those are sounds that do not come from where the virtual source seems to be (in fact, all the sound MUST come from the loudspeakers…), but gives our auditory system the impression that the sound source is located at some -real- point in the space. There may be phantom sources simulating the direct sound as well as its reflections on the surface of a virtual room (i.e., the echoes). The DSP procedures involved on their simulation deal with energy, time, and spectrum information.

1-Energy: **Multichannel Gain scaling**. it is possible to calculate the amount of energy delivered by the source for a given spatial location, listener location, number of channels, location of the loudspeakers and directional characteristics of the source, to scale the gain and the phase of a signal at each output channel. The procedure involved is referred to as *intensity panning*. Though widely used, intensity panning is at present being strongly criticized, because it is effective only for a small group of listeners located at the center of the listening room and also because the signals coming from two loudspeakers will reach both ears of the listener, meanwhile rendering redundant information. The latter is often referred to as the *crosstalk* of the loudspeakers. Other problem with intensity panning is that, once a mix is done for a particular array of loudspeakers, it is not possible to reproduce it (at least it SHOULD not be possible…) using a different one.

Some British recording engineers -mainly Michael Gerzon- created a technique called *Ambisonic* (which is a registered Trademark of Nimbus Communications International), widely used at present. Ambisonic attempts to overcome the limitations above mentioned by encoding the signal on the same way that a special microphone would record it (as a matter of fact such microphones exist, and one of them is the *Calrec Soundfield* microphone). This encoding keeps the information of the energy delivered by a sound source located on a three dimension field using four signals (there are other encoding formats using more than four signals, but we will not deal with these now) as if these were recorded by an array of three figure of eight microphones (each one pointing to the three axes) plus an omni-directional microphone. The decoding procedure attempts to recreate the wavefront that the microphone "had listened" for a given array of loudspeakers (Ambisonic's specialist refers to that array as the *rig*). In Ambisonic, all the loudspeakers works together "pulling and pushing", and that is why it is not advisable to mix signals not encoded using Ambisonic with an encoded one. The advantage of Ambisonic, however, is that if we have a signal properly encoded we may further decode it for the rig of loudspeakers we are to use in a particular situation.

2-Time information: Delay Lines. to simulate all the temporal information delivered by a sound source it is possible to calculate the time it takes to the signal to reach some particular point. There are several ways to do this, whether to calculate the arrival time to a global point where the listener is supposed to be located, to calculate the two arrival times to the different ears of the listener (ITD simulation), or to calculate the arrival time to the different loudspeakers. For the simulation of any of these, *delay lines* are used and, if the

sound source is moving, interpolation processes are used to get the intermediate values. The latter gives rise also to the simulation of the Doppler shift. It should be pointed out that the temporal information simulated may be totally distorted if the listener is not located at the ideal point.

3-Spectral information: HRTF filtering. To simulate this, a set of recorded Impulse Responses using either a dummy or a real head with binaural microphones is often used. The data gathered with the binaural recordings is used to filter the signal, either by convolution (FIR filters) or designing dynamic IIR filters. Since space must be sampled on a discrete way, again very complex processes to get intermediate values for the filters must be performed. It is still a matter of discussion whether this simulation actually renders an effect of accurate location. It seems that there are subtle differences on the shapes of our external ears that are significant enough to make one particular set of HRTFs not useful for all people. Even if a set of "universal HRTFs" could be developed, on a reproduction over loudspeakers situation the signal will be processed twice: the first one when filtered by the system, and the second one when filtered by our body. The simulation is also strongly dependent on the location of the listener and the position of his head and so, not very reliable for loudspeaker reproduction. However, it seems that, under headphones, it is possible to obtain some fairly good results.

Simulation of room characteristics

There are two procedures widely used to simulate room features: fast convolution and networks of IIR filters.

It is possible to "capture" the impulse response of any room by recording an impulse (actually, a sound the more similar to an impulse as possible...as a fire-cracker or balloon explosion, or a pistol shot...) inside this room. This will give us all the information of the reverberation of the room about source and listener location. The recorded impulse response is further used to convolve any sound, meanwhile translating to it the features of the room. It is also possible to modify the impulse response, if desired, on any useful way before the accomplishment of the procedure of convolution. The procedure of convolution is, generally speaking, more accurate to the cost of higher computational expenses.

The other approach is to simulate reverberation through a network of Comb, Allpass, and Lowpass filters connected in parallel and in series. This procedure renders a somewhat artificial result, but is more controllable and less computationally expensive than the former.

A GENERAL VIEW OF A SPATIALISATION SYSTEM AND SOME FINAL CAVEATS.

As we have seen, in order to produce an effective simulation, a system must perform not only the proper DSP procedures, but also take account of the position of the listeners, and the reproducing system that will be further used.

Thus, a system must have information about:

1-Source location and/or movement.

2-Source directional characteristics (omnidirectional, cardioid, etc.)

3-Source orientation (if radiation of the source is different than omnidirectional)

4-Room properties: size, geometry, diffusion, spectral balance.

5-Listener location.

In order to have good results on spatial simulation, the input sound signal to be processed should meet the following conditions:

1-It should be recorded or generated without any other spatial cue that can distort the further process (i.e., reverberation, echoes, etc.)

2-It should have significant energy on the frequency band that is relevant for the used cues. To spatialize a sine wave at 200 Hz is not a good idea…unless you are using ITD.

3-If it is desired to simulate a closed location, the input sound must have significant energy.

4-When simulating moving sources, it should have enough average energy over all its length.

5-When simulating moving sources, it should have some spectral changes during its evolution.

6-Avoid any undesirable abrupt change in the signal (i.e., clicks). The reverberation will not fix it, it will make it more wider and evident.

Also, with respect to the movements indicated:

1-Avoid very fast movements.

2-Stop movement at some point, and remain at this point for a little time, to give the auditory system opportunity to track its location.

3-Avoid "impossible movements". Natural moving sources do not reach maximal velocity instantly, neither stop instantly.

4-Do not place or move a sound source exactly the same way as another at the same time. Our auditory system will tend to believe that there is ONLY one sound source (the law of "common fate" is applied, and besides this, it is impossible to have more than one object on the same space at the same time).

With respect to the reproduction system:

1-The loudspeakers used must have a flat response over the audio frequency band, at least over an angle of 60 degrees.

2-The loudspeakers used must have their phase response carefully aligned.

3-If the signal was processed considering distance and angles for the loudspeakers, the setup of the loudspeakers must meet the proper distances and angles. There are further possibilities to correct this slightly, but to the cost of a reduction of the listening space.

High quality software for spatial treatment of sound has been already developed by several DSP specialists. The differences between several programs are due to the fact that -since there seems not to be a perfect solution for the problem- their authors must state some limitations and make strategical choices. These are usually based on tradeoff between perceptual results and computational expenses as well as on the features of the chosen environment (hardware, Operating System, etc.). However, since the goal of this programs is to provide tools for artistic production, it is interesting to realize that the limits of each implementation may be transformed by the composers on a beautiful way to characterize their own musical structures. Here follows a list of -non commercial- programs that are -to my knowledge- very good and useful:

1-The unit generator ***space***, for the program ***cmusic*** (both by F. R. Moore)

2- The unit generator ***dlocsig*** (by Fernando López Lezcano.), for the program ***Common Lisp Music*** (by Bill Schottstaedt).

3- The unit generators ***locsig*** **and** ***space*** **(by** Richard Karpen**)**, for the program ***Csound*** (by Barry Vercoe).

4-The programs ***Spatialisation Tools*** and ***Vspace*** (by Richard Furse).

5-The program wdspa (by Oscar Pablo Di Liscia).

ACKNOWLEDGEMENTS

Many thanks to Vanessa Del Barco for her revision of the english version of this lecture.