Surround and 3D-Audio Production on Two-Channel and 2D-Multichannel Loudspeaker Setups

Alexis Baskind[1], Thibaut Carpentier[2,3,4], Jean-Marc Lyzwa[5], Olivier Warusfel[2,3,4]

[1] *Independent sound engineer, Email: a@alexisbaskind.net*

[2] *CNRS, UMR 9912, Sciences et Technologies de la Musique et du Son, Paris, Email : Thibaut.Carpentier@ircam.fr*

[3] *IRCAM, UMR9912, Sciences et Technologies de la Musique et du Son, Paris, Email : warusfel@ircam.fr*

[4] *Sorbonne Universités, UPMC, UMR9912, Sciences et Technologies de la Musique et du Son, Paris*

[5] *Conservatoire de Paris (CNSMDP), 209 Avenue Jean-Jaurès 75019 Paris, Email: jmLyzwa@cnsmdp.fr*

Abstract

Most existing and arising 2D and 3D loudspeaker setups are primarily designed for cinema, and intend to render a stable frontal soundscape at the expense of precision in the reproduction of the non-frontal information. Another difficulty related to the introduction of new surround sound standards is the need for an efficient mixing approach to adapt 3D audio content to stereo and surround domestic systems.

The project presented here aims at providing tools to overcome those two limitations on two-channel and 2D-multichannel setups, i.e. to enable the reproduction of 2D- and 3D-surround content on pure 2D-systems, to ensure a stable reproduction of the lateral sound images and a homogeneous envelopment of the listener.

1. Introduction

One of the greatest challenges related to the introduction of 3D-audio standards for the home consumer is the backward compatibility with existing 2D-audio systems (2.0 stereo, 5.1 and 7.1 surround among others). In fact, although current data storage systems have enough capacity to store numerous versions of an audio master (and the standard bandwidth of internet connections will probably grow in the coming years so as to allow the same with streaming formats), the production time required to release each master may increase dramatically with the number of supported audio formats. There is thus an increasing need for tools that facilitate mixing for several standards while preserving as much spatial, timbral and balance characteristics of the original mix as possible.

This is the goal of the so-called *Transpan* suite that has already been presented among others in a previous paper [1]. Taking advantage of binaural processing and cross-talk cancellation (CTC) techniques, it provides:

- an extended multichannel panner, which allows for a stable lateral positioning, an enveloping sound image, and the ability to render sources outside the horizontal plane,
- a virtual 3D-surround panner on two-channel setups,
- a multichannel to two-channel downmixer that preserves much of the original spatial information,
- a two-channel enlarger of the stereo image,
- an innovative equalization technique, which considerably improves the preservation of the timbral quality.

The architecture of the multichannel panner, as well as the downmixer, were already presented in [1]. This paper explains in greater detail the principles behind the equalization method of the crosstalk-canceller, the two-channel panner as well as the two-channel enlarger.

2. Improved Crosstalk Cancellation Equalizer

The use of crosstalk cancellation (also called "CTC" or "transaural") in music production in order to extend the stereo image is not new. Suggested initially by Atal and Schroeder [2], it was further developed by other researchers such as Cooper and Bauck [3], Kirkeby and Nelson (stereo dipole) [4] and Ralph Glasgal (*Ambiophonics* technology) [5]. In the last fifteen years, several end-consumer products, such as "soundbars", aimed at using this technology to render surround sound material using only a small array of frontal loudspeakers.

One of the major criticisms of such solutions is the significant distortion of tone color. This distortion is inherent to the technique itself: a crosstalk canceller can in fact be seen as a cascade of a dual mono equalizer and a matrixing system (Figure 1 depicts for instance the "symmetric feedforward" architecture).

The equalizer itself (called "1/D" in Figure 1) is the inverse filter of a filter D, which is the determinant of the complex system transfer matrix to be inverted (see for instance [6] or [7] for mathematical details). Should the loudspeakers be positioned symmetrically with respect to the listener's median plane, the transfer function of D is defined by:

$$D(f) = H_i(f)^2 - H_c(f)^2 \quad (1)$$

where H_i and H_c are the filters modeling the ipsilateral path and the contralateral path, respectively.

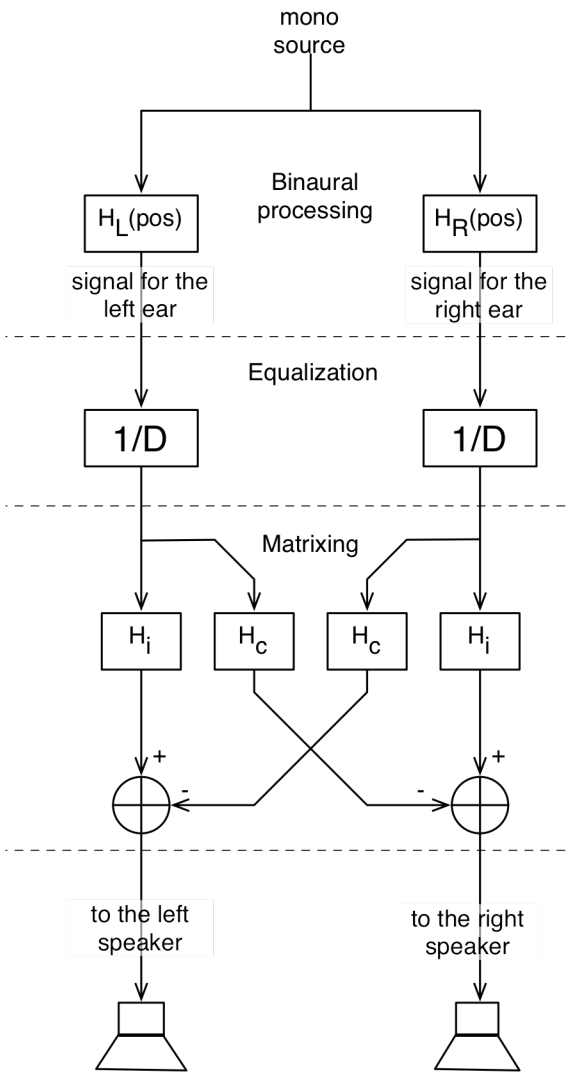

Figure 1: ***Binaural Processing and Symmetric Feedforward Crosstalk Cancellation****: H_L and H_R are the binaural filters for the selected position of the source. H_i and H_c are the binaural filters modeling the ipsilateral path and the contralateral path, respectively. 1/D is the equalizer, corresponding to the inverse of the complex matrix determinant D. The crosstalk canceller itself is the cascade of the equalization stage and the matrixing stage.*

Therefore, if H_i and H_c are close both in amplitude and phase, the determinant D tends to zero (the system is said to be underdetermined), and its inverse tends to infinity. This happens typically in three cases:

- if the loudspeakers are very close to the median plane. This case is however not relevant in the context of a standard two-channel or multichannel setup
- at low and low-mid frequencies (below 300Hz), where wavelengths are significantly larger than the head.
- around the so-called "pinna-notch" in the high frequency region (around 10 kHz): due to destructive interferences in the pinna, the ipsilateral path and the contralateral path tend then to have (small) similar amplitudes

In the two latter cases, the equalizer presents high peaks (up to several tenths of a dB). In theory, these peaks are compensated for by the corresponding notches in the actual acoustic paths between the loudspeakers and the listeners. In practice however this case is unlikely to occur, since it would mean that the actual acoustical paths correspond perfectly to the simulated paths H_i and H_c being used in the crosstalk canceller, which is impossible for at least three reasons:

- the filters H_i and H_c are typically derived under anechoic assumption, which does not hold in practical listening conditions,
- when using non-individualized HRTFs (in the binaural processing stage), the actual H_i and H_c paths differ from the simulated/computed ones,
- the listener is rarely in the sweet spot, looking strictly forward, and even if he/she tries to, he/she is unlikely to stay exactly in the sweet spot: as an order of magnitude, the wavelength corresponding to 10kHz is approximately 3.5 cm. This means that even a one-centimeter movement of the head destroys the coherence in this frequency region, thus yielding unwanted distortion.

Therefore, the mathematical model behind the classical equalization principle does not fit a real situation, which leads not only to a disruption of the spatial effect (in the high frequencies notably), but also to a strong coloration in various regions of the spectrum. This is especially true for the high frequencies (around 10kHz), but the sources of biases mentioned above also apply to less critical regions: typically, the mid-range (around 1.5kHz) tone color suffers from a discrepancy between the "ideal" and the actual situations.

There is unfortunately no perfect solution to this dilemma, since the possible listening situations are numerous and unpredictable. One solution, presented in previous publications [1][8], consists in detecting and minimizing the biggest dips in the determinant which would lead to exaggerated spectral peaks in the equalizer.

An alternative approach presented here, consists in designing an equalizer that minimizes those unwanted spectral rises for as many positions of the virtual source as possible (and not only for the case where the virtual source coincides with a loudspeaker's position).

This is achieved by a modification of the equalization paradigm: the filter to be inverted (D(f) in the original case) is not anymore considered as the determinant of the transfer matrix of the rendering system, but as the filter that whitens the whole processing (binaural processing and CTC). "Whitening the whole processing" means here that the total power spectrum of both signals emitted by the loudspeakers has to be estimated. The simplest way to perform this is to sum the two left and right power spectra. This assumption is not true, strictly speaking, since the two power spectra are not statistically independent. We use this method nevertheless, since informal subjective tests showed that it leads to a significant improvement of the tone color quality.

The estimated total power spectrum of the total processing (binaural + crosstalk cancellation) without equalization for

one arbitrary position *pos* of the virtual source (defining thus the spectrum to be inverted) is defined by:

$$TotalPower(pos, f) = Power_L(f)^2 + Power_R(f)^2 \quad (2)$$

According to the terms defined in Figure 1, this writes:

$$\begin{aligned} TotalPower(pos, f) = {} & |H_L(pos, f).H_i(f) \\ & - H_R(pos, f).H_c(f)|^2 \\ & + |H_R(pos, f).H_i(f) \\ & - H_L(pos, f).H_c(f)|^2 \end{aligned} \quad (3)$$

The power spectrum of the equalizer is then defined as the inverse of this total power spectrum. The calculation of frequency response requires an additional assumption, since the phases are missing in the power spectrum. The minimum-phase assumption, which is very common for equalizing filters, has been chosen here, since it minimizes the time effects of strong amplitude variations. In particular, minimum-phase filters do not present pre-echoes like linear-phase filter may. The calculation of the frequency response and impulse response of the equalizer are then straightforward thanks to the Hilbert transform [9].

If the virtual source is set at the position of one of the speakers, it leads to:

$$TotalPower(speakerPos, f) = |H_i(f)^2 - H_c(f)^2|^2$$

which is also the power of the filter to be inverted in the traditional method as defined in Equation (1).

This new definition of the filter to be inverted is thus formally coherent with the classical one, at least in terms of power spectrum.

This new equalization method can then be implemented in two different ways:

1. Static Equalization: the equalizer does not depend on the virtual source's position. An average equalizer has to be defined: the method consists in averaging the total power as defined in Equation (3) for several positions of the binaural source. This averaging can be weighted, for example to put the emphasis on the sides, where the use of crosstalk cancellation is the most relevant. The equalization filter is then calculated by deriving the phase from the square root of the average power spectrum under the minimum-phase assumption

2. Dynamic Equalization: the equalizer depends on the position of the virtual source, and is implemented for instance as a mono filter before the HRTF filtering. Of course, if the virtual sources are moving, all three filters (the equalizer and the two HRTFs) become time-variant.

Informal listening tests revealed that these methods provide a significant improvement in the preservation of the timbres compared to the classical crosstalk canceller equalization. The other problem mentioned above, namely instabilities in the low frequencies, is yet to be solved at this stage: since the total power theoretically tends to zero in the low frequencies irrespective of the position of the virtual source, it cannot be directly inverted. As there is no real solution to this fundamental limitation of crosstalk cancellation, it was decided to high-pass the equalization filter, so that the lowest frequency range is not processed at all by the crosstalk canceller. In order to preserve the integrity of the spectrum in the whole audible range, a crossover filter has to be used (see section 3).

3. Two-Channel 3D-Panning

The two-channel 3D-Panning method presented here, and designed for a standard two-channel stereo setup, works in a very similar way as the 5.1 3D-Panner presented in a previous paper [1]. It is inspired from a classical binaural/CTC 3D panner, but differs in three main aspects:

1. The classic CTC equalizer is replaced by the equalizer presented in section 2
2. A crossover filter separates the low-frequency range from the rest of the spectrum, and recombines it with the CTC output at the output stage. The crossover design has to be properly tuned in coherence with the high-pass filter used in the

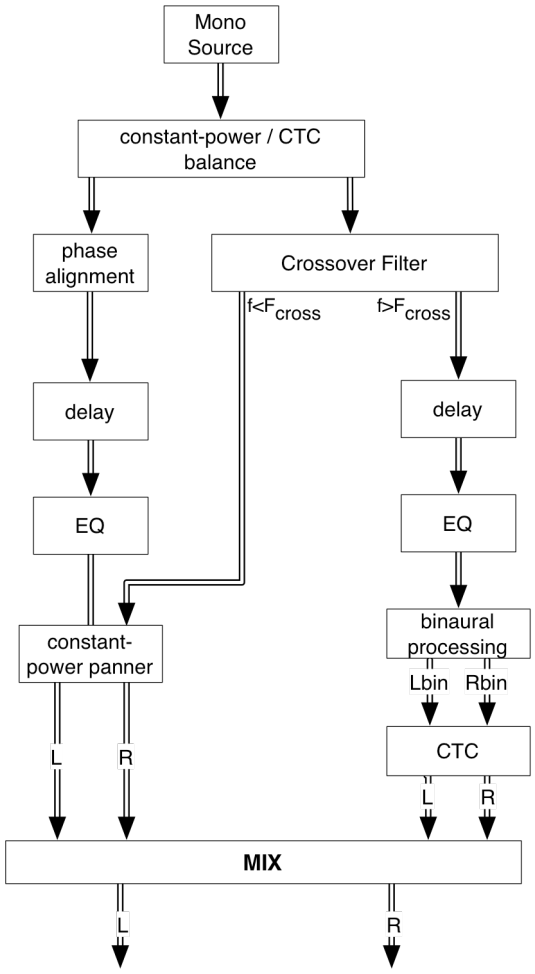

*Figure 2: **Architecture of a two-channel 3D-Panner**: the crossover is typically set to a frequency F_{cross} < 100 Hz)*

equalizer to ensure an interference-free reconstruction.

3. The crosstalk cancellation only operates if the virtual source is not positioned in the horizontal plane between the two loudspeakers. In the latter case, a standard constant-power panning algorithm is used. The balance between both techniques depends on the source position.

The architecture is shown as a diagram in Figure 2. The low frequency part of the signal is sent to the constant-power panner, whatever the overall balance between constant-power panning and CTC is. As crossover-filters imply a phase rotation [10], an allpass phase-alignment stage, working in accordance with the crossover settings, has to be inserted prior to the input of the constant-power panner.

As for the 5.1 3D-Panner, two delay blocks are inserted in the constant-power and crosstalk-canceller signal paths, in order to allow a fine time adjustment of both layers.

Although this was not explicitly mentioned in the previous paper [1], the same crossover principle was implemented in the 5.1 3D-Panner as well as the 5.1-to-2.0 downmixer that were presented there.

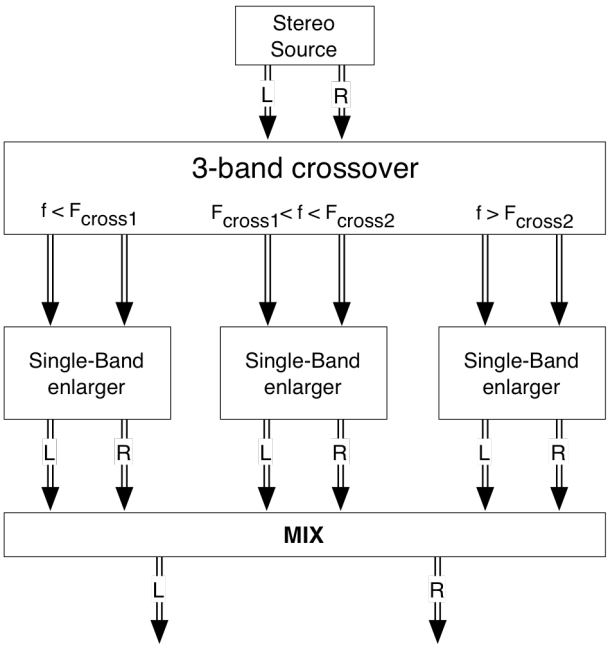

Figure 3: ***Two-Channel 3-Band stereo enlarger*** *- Global Diagram. See Figure 4 for the detail of each block entitled "Single-Band Enlarger"*

4. Two-Channel Stereo Enlargement

Three methods are commonly used in music production to enlarge the stereophonic image of a two-track recording:

- Delay Processing: this method consists in matrixing the stereo track and introducing short delays that differ slightly for each of the four paths (L->L, L->R, R->L, R->R), in imitation to early reflections in a room.
- Complementary filters: here the source material is also matrixed, but instead of delays, complementary filters (leading to a total power spectrum of unity) are used. Therefore the signals are split in the frequency domain, and then recombined with a different stereophonic balance.
- M-S Processing: the principle, introduced by the founding father of stereophony Alan Blumlein [11][12], consists in modifying the balance between the "M" and the "S" component of the stereo track to the benefit of the "S" component.

A closer look at Figure 1 shows that a crosstalk canceller is conceptually quite close to a M-S processor: basically, if the binaural filters H_i and H_c were simple gains, a crosstalk canceller would be an M-S processor. The "shuffler" implementation of a crosstalk canceller [3] makes this parallel even more obvious.

However, using a crosstalk canceller directly to enlarge the stereo image is not satisfactory, as it leads to a strong tone color distortion (even with an optimized equalization) and to a significant degradation of the spatial image due to the exaggerated amount of out-of-phase information.

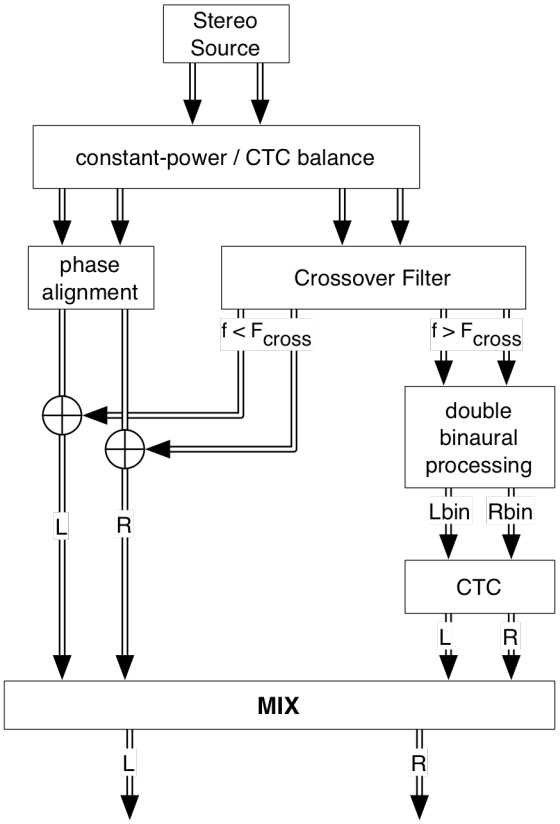

Figure 4: ***Two-Channel Single-band stereophonic enlarger*** *- Detail for any of the three bands from* ***Error! Reference source not found.****. The crossover is typically set to a frequency F_{cross} < 100 Hz). Note: this crossover frequency F_{cross}, corresponding to the bass management of the crosstalk canceller, should not be confused with the two frequencies F_{cross1} and F_{cross2} defined in Figure 3, which are freely adjustable by the user.*

The theoretical goal of a cross-talk canceller is to produce virtual headphones around the listener's ears, assuming that he/she sits at the sweet spot. Therefore, the resulting sound image of a cross-talk canceller applied to a non-binaural stereo recording sounds very close and "phasy", similar to what an exaggerated "S" component through an M-S processing would produce. This effect can of course be reduced through a proper balance between the unprocessed and the processed signals (similarly to increasing the "M" component in an M-S processor), but the approach developed in this project aims at producing a more subtle and externalized effect.

This can be achieved again through the simultaneous use of binaural processing and cross-talk cancellation: in the technique developed by the authors, the stereo recording is sent to binaural virtual sources symmetrically positioned beyond the loudspeakers (typically between +/- 40° and +/- 110°) and the binaural output is processed by the crosstalk canceller. In order to allow a more flexible and fine treatment, this processing is performed in three independent frequency bands whose crossover frequencies are fully adjustable. These frequency bands are recombined at the very end to form the final, full-band signal. For each band, the position of the virtual sources as well as the dry/wet balance can be set independently. A global diagram of whole architecture of the enlarger is shown in Figure 3 while a detailed diagram of any of the three frequency bands is shown in Figure 4.

5. Conclusion

The ensemble of techniques presented in this paper aims at producing and manipulating the stereophonic image in two-channel and two-dimensional surround setups in order to extend the spatial limitations of traditional panning methods. They all rely on the joint use of these traditional panning methods and binaural-based processing techniques with crosstalk cancellation in order to make them compatible with loudspeakers. The proposed design significantly improves the quality and the precision of the rendering in comparison to usual binaural/CTC implementations, allowing it to be used much more easily in a practical music production workflow.

References

[1] Baskind A., Carpentier T., Noisternig M., Warusfel O., Lyzwa J.-M.: “Binaural and transaural spatialization techniques in multichannel 5.1 production”. *Proceedings of the Tonmeistertagung 2012, Köln, November 2012*

[2] Schroeder M. R. and Atal. B. S., “Computer simulation of sound transmission in rooms”. *IEEE Conv. Record, 7:150-155, 1963*

[3] Cooper D. H. and Bauck J. L.: “Prospects for transaural recording”. *Journal of the Audio Engineering Society, vol. 37, no. 1–2, 1989*

[4] Kirkeby O. and Nelson P. A, "The stereo dipole - a virtual source imaging system using two closely spaced loudspeakers", *JAES, Vol. 46, No. 5, May 1998*

[5] Glasgal R.: "Ambiophonics: The Synthesis of Concert-Hall Sound Fields in the Home". *Presented at the 99th AES Convention October 6-9, 1995 New York*

[6] Gardner W.: “3-D Audio Using Loudspeakers”. *PhD dissertation, Massachusetts Institute of Technology, 1997*

[7] Kaiser F.: “Transaural Audio - The reproduction of binaural signals over loudspeakers”. *Diploma Thesis, Universität für Musik und darstellende Kunst Graz / Institut für Elektronische Musik und Akustik / IRCAM, March 2011*

[8] Cornuau C.: “Étude et optimisation de la synthèse transaurale à deux canaux”. *Diploma Thesis (in French), Formation supérieure aux métiers du son, Conservatoire National Supérieur de Musique et de Danse de Paris, March 2011*

[9] Oppenheim A. V., Schafer R. W., and Buck J. R., "Discrete-Time Signal Processing" (3rd Ed.). *Prentice Hall, New Jersey, 1999.*

[10] Linkwitz S., "Active Crossover Networks for Noncoincident Drivers", *JAES, Vol. 24, No. 1, January/February 1976. Reprinted in Loudspeaker Anthology, Vol.1, AES 1978*

[11] Blumlein A.: "Sound Transmission, Sound Recording and Sound Reproducting System", *US Patent No 2,962,275, 1936*

[12] Blumlein A.: "Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems", *GB Patent No GB394325, 1933*