Binaural and transaural spatialization techniques in multichannel 5.1 production

(Anwendung binauraler und transauraler Wiedergabetechnik in der 5.1 Musikproduktion)

Alexis Baskind, Thibaut Carpentier**, Markus Noisternig**, Olivier Warusfel**, Jean-Marc Lyzwa****

* Hochschule der populären Künste FH, Otto-Suhr-Allee 24, 10585 Berlin, a.baskind@hdpk.de
** UMR STMS IRCAM-CNRS-UPMC, 1 place Igor-Stravinsky, 75004 Paris,
Thibaut.Carpentier@ircam.fr, warusfel@ircam.fr, Markus.Noisternig@ircam.fr
*** Conservatoire de Paris (CNSMDP), 209 avenue Jean-Jaurès 75019 Paris, jmLyzwa@cnsmdp.fr

Abstract

The work presented here aims at overcoming some of the weaknesses of the 5.1 standard in terms of stability and precision of lateral sources. In addition to traditional surround panning (which relies on constant-power panning and/or multichannel microphone recordings) an extra spatialization layer based on binaural/transaural processing is introduced. This article discusses the audio signal processing and panning methods, and shows different applications scenarios and use cases.

1. Introduction

One of the main challenges in the process of mixing music is to provide the listener with a soundscape of the greatest possible clarity. In this sense, the ITU 5.1 standard represents a considerable enhancement over 2-channel stereophony. However, its main drawback is that it privileges the frontal region and blurs the side and rear regions of the sound scene.

The work presented here aims at overcoming this problem by providing an additional sound spatialization layer to the surround mix (i.e. to surround sound recording techniques using main and spot microphones). This approach is fully compatible with the ITU-R BS 775 standard for 5.1 surround sound playback. In the proposed approach, the spatialization of a single sound source relies on the parallel use of three techniques, the final of which is the result of the study described in the following sections:

- Multichannel microphone arrays, if available, create a first layer with a coherent spatial image directly at the recording.
- Constant-power panning utilizing spot microphones forms a second layer that plays a major role in balancing the timbral, spatial and amplitude features of the mix. The simultaneous use of multichannel microphone arrays and spot microphones distributed with constant-power has been used for many years for 5.1 music production, in particular for classical music.

- A third layer, based on binaural/transaural processing of the same spot microphones, using two loudspeaker pairs (L/R, and Ls/Rs), provides the spatial precision that lacks for the lateral images.

Therefore, this approach aims at combining the advantages of standardized surround panning techniques with those of binaural/transaural processing, and at compensating for their respective drawbacks. Moreover, this processor can be used to considerably widen the stereophonic space when down-mixing from 5.1 to 2.0.

This technique is currently implemented in two forms: (1) as a standalone application called "Transpan" that aims at being used as an external insert in the production workflow (a snapshot of the GUI is shown in Fig. 1), and (2) as a module in the real-time spatial audio processing library Spat~ (*Spatialisateur*) [1] developed by IRCAM for the Max/MSP [2] environment.

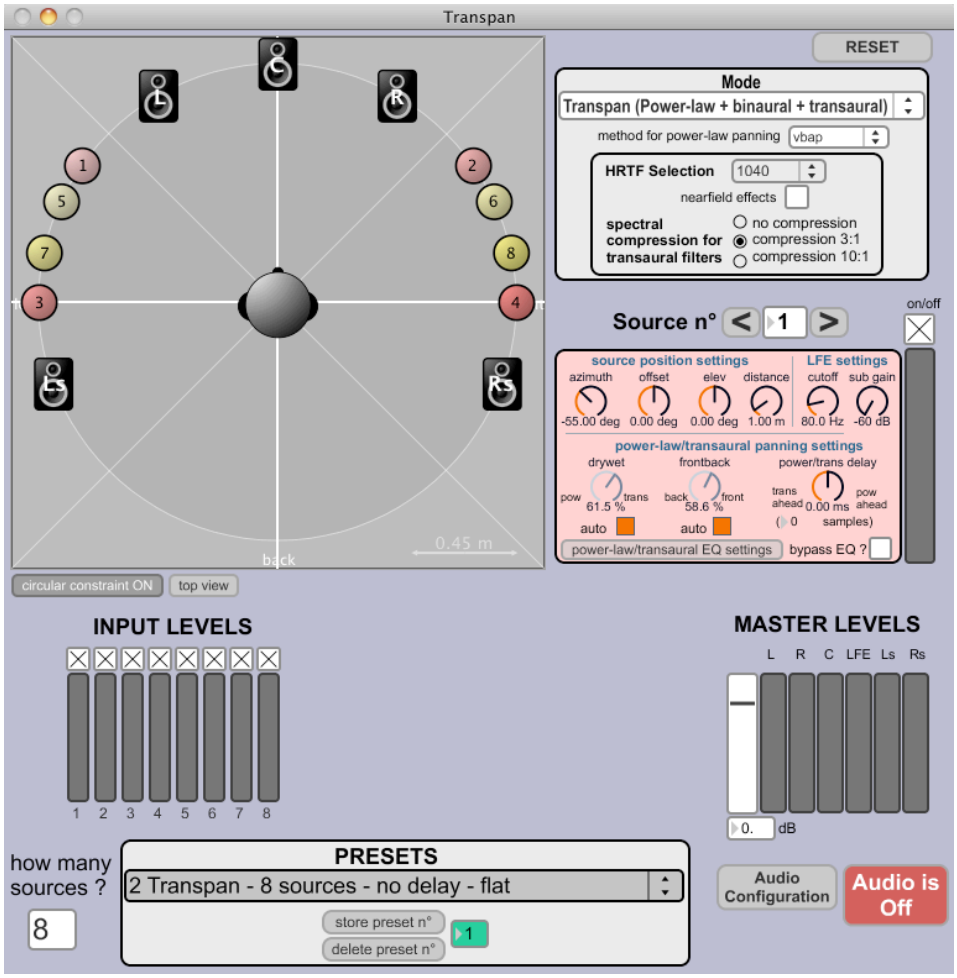

Figure 1: GUI of the "Transpan" application (alpha-version)

[1] http://forumnet.ircam.fr

[2] http://cycling74.com

2. Why use binaural and transaural techniques in 5.1 production?

A traditional way of creating a stereophonic space in a 5.1 recording, especially in classical music production, relies on the simultaneous use of two layers:

1. Multichannel microphone arrays extend the idea of main microphone pairs in 2-channel stereophony to 5.0 surround recordings. The surround sound main microphone technique aims at providing a faithful and coherent reproduction of the scene being recorded. It gives a significant amount of information about the room, as for instance depth and source positions. However, the image provided by this layer often lacks precision and stability, and moreover, the balance between the instruments cannot be modified.
2. A monophonic spot microphone can be spatialized coherently in reference to the image provided by the microphone array using constant-power panning techniques. If no divergence is involved, constant-power panning relies only on the two adjacent loudspeakers (pairwise panning). It aims at correcting balance issues, as well as providing the timbral and spatial precision that may be lacking in the microphone array layer. Individual equalization of both layers gives an extra degree-of-freedom, in regards to spectral and spatial rendering. Moreover, depending on the context, this layer can be time-aligned or not with the multichannel array in order to fine-tune the spatial image and correct possible comb-filtering problems.

This two-step method (generally supplemented by surround/ambiance microphones and/or artificial reverberation) has been massively used for decades on all 2-channel and multichannel systems relying on stereophony. It offers a simple and reliable way of creating a soundscape. However, its application on 5.1 suffers from several limitations based on the specification of the standard: namely, by the fact that sources spatialized in the lateral and rear sections are often instable and imprecise. This problem is due to the large angle between the front and rear loudspeakers (80°), as well as between the two rear loudspeakers (140°)[1], which act to weaken summing localization. In addition, human's ability to localize sound is also poorer on the sides (the so-called "localization blur" is more pronounced) because the interaural differences have a smaller variation as a function of the source position than in the median plane [1].

This instability of lateral sources is especially obvious when the listener moves in the front/back axis, thus reducing in this dimension the size of the sweet spot compared to the case of 2-channel stereophony.

The solution proposed here, based on an approach developed for several years by the Conservatoire de Paris [2], consists in complementing the spatial image with a third layer, that uses binaural and transaural processing to position the source. Binaural/transaural processing is indeed particularly efficient at rendering precise lateral sources. In our trial tests we found transaural rendering to be generally more stable with respect to front-back movements than constant-power panning. It also allows some interesting features, such as simulating a source below or above the horizontal plane, or simulating proximity (i.e. a source positioned inside the loudspeaker circle). However, binaural/transaural processing also suffers from two main drawbacks:

- The quality of the rendering is quite sensitive to the geometry and dimensions of the listener's head and torso, as they should ideally perfectly match the morphology of

[1] We consider here that the rear loudspeakers are positioned at +/- 110°

the dummy or human head on which the binaural and transaural filters were measured. In case of a mismatch, both timbral and spatial features will suffer alterations. In this case, the spatial rendering is less precise, and the tone color may be severely damaged. This is particularly true for transaural rendering: as it implies filter inversion, it can introduce unnatural resonances in the mid- and high-frequency range. However, as regards transaural processing, regularization methods during the calculation of the crosstalk filters can reduce these artifacts. Indeed, as described in a recent study by Clément Cornuau [3] at IRCAM, it was shown that applying a compression factor to the magnitude spectrum of the HRTF involved in the inversion process brings a considerable improvement. This so-called "spectral compression" of transaural filters is implemented in Transpan.

- Transaural rendering is quite unstable with respect to left-right movements and head rotations. Indeed, as noticed by Gardner [4], sound sources localization was found to be more stable with respect to front-back movements than to left-right movements and head rotations (with and without head-tracking), which is coherent with our own observations.

The simultaneous use of constant-power panning and binaural/transaural processing consists of compensating the drawbacks of one technique by the advantages of the other and vice-versa. The use of the new binaural/transaural layer with the two traditional layers as described above (or with the constant-power panning layer only if no multichannel microphone array is involved in the recording), leads to similar considerations and constraints as when superimposing a spot-microphone panned with constant-power to a multichannel array: first, the spatial position should be coherent on all involved layers; second, special attention must be given to equalization of each layer and its respective time-alignment.

3. Architecture of the Panner

Figure 2 shows the architecture of the Transpan panner. The diagram shows the processing of a mono spot-microphone, considering that the result is meant to be mixed with the main multichannel array, if available. This architecture combines traditional elements such as a 5.0 constant-power panner (found in all DAWs and mixing boards) and a LFE / low-pass filtered send channel, with the new binaural, transaural, time-alignment and equalization processing elements.

3.1. Binaural Processor

The binaural processor models the acoustic path from a mono sound source to both ears using the so-called *HRTF* (Head-Related Transfer Functions). The two resulting signals, called here Lbin and Rbin, are meant for headphone listening. The binaural processor used here is a part of the Spat~ library [5][6], using HRTF measured at IRCAM on a large number of human subjects.

3.2. Transaural Processor

This processing stage aims at adapting binaural signals for a reproduction on loudspeakers, by compensating the crosstalk from each loudspeaker to the contralateral ear (this is why

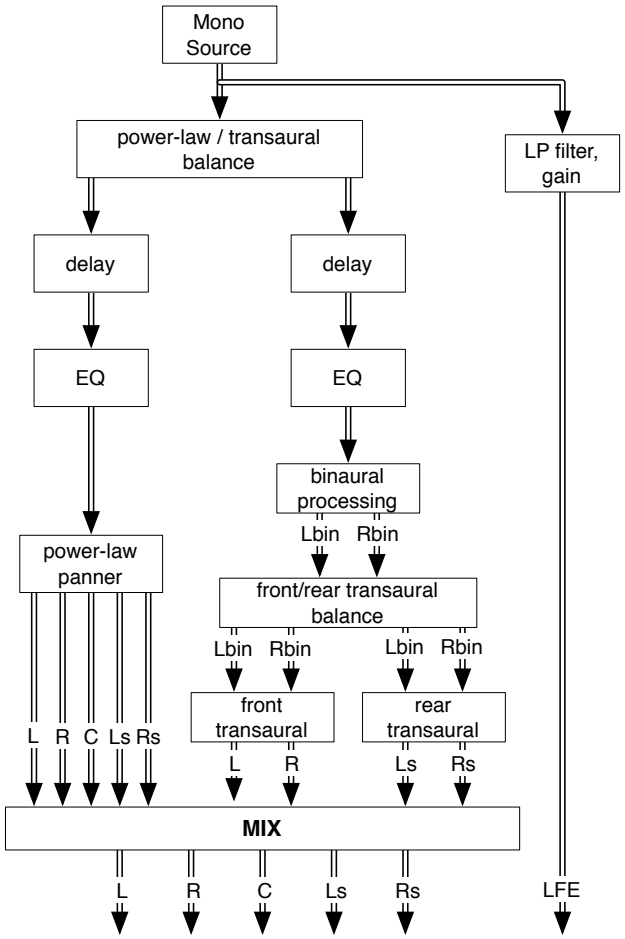

Figure 2: Architecture of the Transpan panner

transaural is also known as *crosstalk cancellation*). Initially proposed by Schroeder and Atal [7], this can be implemented using several possible architectures [4][8]. Following investigations made by Fabio Kaiser [9] at IRCAM on various crosstalk cancellation methods, the so-called "feed-forward asymmetric" technique, as proposed by Gardner [4] was chosen for Transpan, as it provides the best results regarding spatial precision and coloration.

Transaural processing was initially developed for 2-channel stereophony. However, it can be applied to any set of speakers [10], considering that it works best with symmetric loudspeaker setups with respect to the front-back axis. Therefore, two transaural processors are used here: one using the L-R pair, one using the Ls-Rs pair. The main advantage is to overcome a typical problem of the classical (frontal) transaural reproduction, i.e. the instability of non-frontal sources with left-right head movements and head rotations: for significant movements, a rear virtual source is not anymore perceived at its desired position, but will flip into the frontal hemifield. By introducing the rear pair, this can be at least partially compensated, so that the frontal and rear sources remain close to their respective hemifield when the listener moves, even if they lose spatial precision. Moreover, the same principle can also be extended to lateral sources: during our tests, we found that a proper balancing between front and rear transaural information allows sources on the sides to be more stable with respect to movements than a simple transaural pair. The mixing engineer

can thus balance between the front and the rear pair in order to fine-tune the position of the final source.

This balance between front and rear transaural, as well as of the balance between constant-power panning and binaural/transaural panning with respect to source position, can be automatically adjusted to facilitate the workflow. This was recently investigated and implemented by Julien Carton [11].

3.3. Time-alignment

Two delays, one before the constant-power panner, the other before binaural processing, allow to time adjust both layers with respect to each other. As it will be explained below, this allows for the fine-tuning of the spectral and spatial result by correcting phase relationships and by taking advantage of the precedence effect.

3.4. Equalization

Equalizing each layer is used to:

- adjust the final tone color of the source;
- fine-tune the spatial image, for example by boosting one layer in the mid-range (most important for the spatial image in the horizontal plane), in order to make its spatial features more prominent;
- reduce undesired colorations due to binaural/transaural processing;
- minimize comb-filtering or echo detection if the layers are not synchronized.

4. Configuration examples

For the time-alignment between constant-power and binaural/transaural panning, three typical configurations may be envisioned:

1. **Tight synchronization**

   Perfect synchronization is not possible, as binaural processing implies delays between left and right channels (which add to the natural delays between the loudspeakers and both ears in the case of transaural processing). However, for lateral sources, the ipsilateral channel (the channel on the side of the source) is significantly louder than the contralateral channel (the channel on the side opposed to the source). Therefore, as long as the constant-power-panned layer is synchronized with the ipsilateral signal, comb-filtering between the constant-power-panned layer and the contralateral signal is in practice not a problem.

2. **Very small delay**

   By delaying one layer with respect to the other, as compared to case 1 (tight synchronization), by a few samples (either negative or positive delay), significant modifications of the spatial image can be achieved. This has to be compared with the time synchronization between the main microphone array and the constant-power-panned spot microphones, for which advancing or delaying the spot microphones of a few samples with respect to perfect synchronization allows the mixing engineer to fine-tune the spatial and timbral image. Of course, desynchronizing both layers implies comb-filtering that modifies the tone color and may damage it. To minimize this, the binaural/transaural layer has to be significantly softer than the constant-power-panned layer. A low-pass or a high-shelf filter with a negative delay on the

binaural/transaural layer also helps reducing destructive interferences. Such filtering makes sense however only if the delay is very small, otherwise comb-filtering occurs on the whole audible range.

3. **Significant delay (a few tens ms)**

In this case, the binaural/transaural layer is the latest, with a delay varying between 5ms and 30ms. It behaves as an early reflection, considering that the direct sound is provided either by the multichannel array, or by the constant-power-panned layer. As the direct sound provides most spatial and timbral cues, colorations due to binaural/transaural processing are reduced thanks to temporal masking (i.e. *precedence effect* [1]). The counterpart is that the spatial effect of the binaural/transaural layer is also less efficient than in cases 1 and 2. This may still be the right choice if the image provided by the multichannel array and/or the constant-power-panned spot microphone is quite stable, and need only for fine adjustment. In case of sound sources with sharp transients, the echo threshold can be raised by reducing the high-frequency content of the binaural/transaural layer (again using a low-pass or a high-shelf filter).

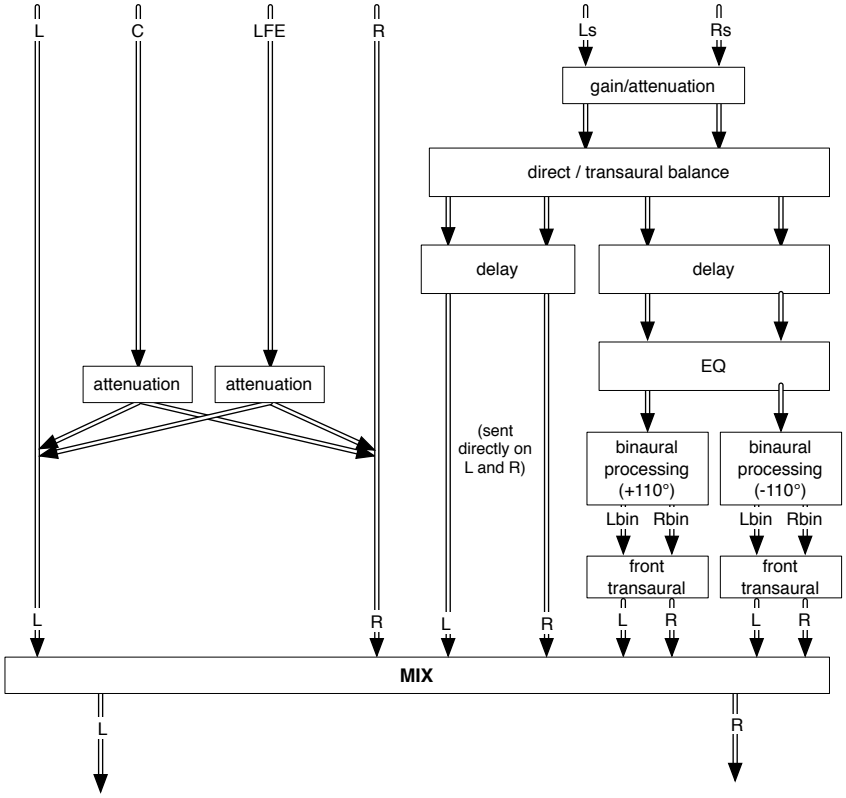

Figure 3: 5.1 to 2.0 downmix using binaural/transaural processing

5. 5.1 to 2.0 downmixing

Another promising application of the binaural and transaural modules used in Transpan is downmixing from multichannel to 2-channel. As shown in figure 3, the downmix approach suggested here differs from the traditional techniques by the introduction of

binaural/transaural processing for the rear channels; instead of being routed respectively to L and R, Ls and Rs channels are spatialized as *virtual loudspeakers*. The main advantage of this approach is that a significant amount of the spatial information of the original 5.1 mix may be preserved: on a traditional 2-channel setup, lateral sources, rear sources, or even sources out of the horizontal plane can be perceived. This works however only on a limited area, which is a little bit smaller than the sweet spot for traditional 2-channel mixes. A snapshot of the downmixer is shown on figure 4.

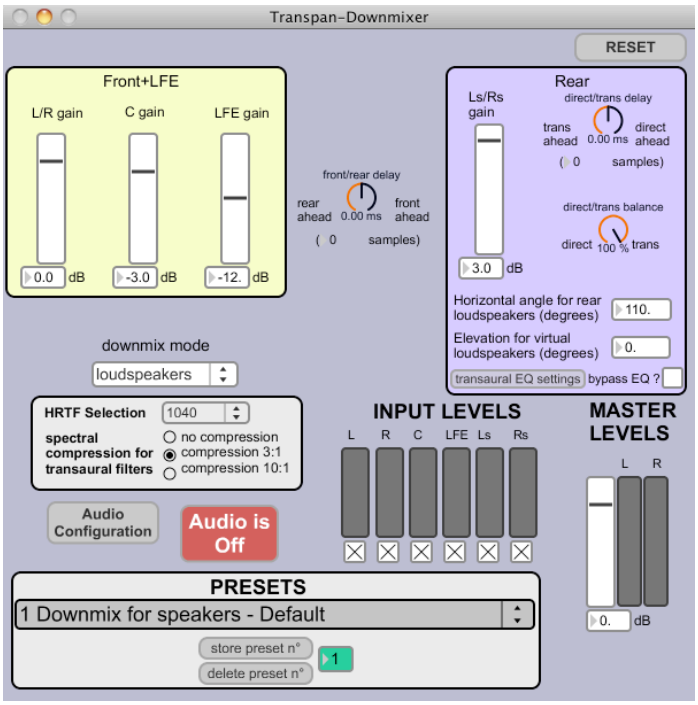

Figure 4: GUI of the Downmixer application (alpha-version)

6. Conclusions

The work presented here aims at enhancing the stability and precision of lateral phantom sources in the 5.1 standard by introducing a new spatialization layer superimposed on the traditional layers (microphone arrays and constant-power-panned spot microphones), utilizing binaural and transaural processing. The proposed technique has been successfully applied in many 5.1 mixes, proving its efficiency. The implemented binaural/transaural engine can be used as well to perform a high-quality 2.0 downmix that preserves as much as possible the spatial content of the original multichannel mix.

The method developed here is not limited to 5.1, and can be applied to any multichannel standard, as many of them present similar weaknesses in regards to stability and the precision of lateral sources.

In its current state of development, this technique is proposed either as a Max/MSP module, or as two standalone applications, which are currently in alpha version. The next step is to provide it as plugins that could be directly integrated in a sequencer for a smoother workflow.

7. References

[1] Blauert, J.: "Spatial Hearing - The Psychophysics of Human Sound Localisation". *The MIT Press, Cambridge, Mass, 1996, ISBN 0-262 02413-6*

[2] Lyzwa, J.-M., Baskind A.: "Utilisation de techniques binaurales et transaurales en prises de son et en post-productions multicanales 5.1". *Report (in French) of the 7th conference of AES Brasil, Sao Paolo 2009. Can be found on* http://www.conservatoiredeparis.fr/la-recherche/ressources-pour-la-recherche/le-service-audiovisuel/

[3] 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*

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

[5] Larcher V., Jot J.-M., Guyard J., Warusfel O.: "Study and comparison of efficient methods for 3D audio spatialization based on linear decomposition of HRTF data". *Proc. AES 108th convention, 2000, preprint n°5097*

[6] Jot J.-M., Larcher V. and Warusfel O.: "Digital signal processing issues in the context of binaural and transaural stereophony". *Proc. Audio Eng. Soc. Conv., 1995*

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

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

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

[10] Bauck J. L. and Cooper D. H.: "Generalized transaural stereo and Applications". *Journal of the Audio Engineering Society, vol. 44, no. 9, 1996*

[11] Carton, J.: "Intégration et exploitation de traitements transauraux pour la production au format multicanal 5.1". *Diploma Thesis (in French), Formation supérieure aux métiers du son, Conservatoire National Supérieur de Musique et de Danse de Paris, March 2012*