19th INTERNATIONAL CONGRESS ON ACOUSTICS
MADRID, 2-7 SEPTEMBER 2007

PERCEPTUALLY-BASED AURALIZATION

PACS: 43.60.Dh

Tsingos, Nicolas
REVES-INRIA, Sophia Antipolis, France; nicolas.tsingos@sophia.inria.fr

ABSTRACT

With increasingly complex environments, the cost of auralization can quickly become a significant bottleneck for interactive applications, such as video games or simulators.
While limitations of the human auditory perception have been successfully leveraged for lossy audio compression, real-time auralization pipelines still implement brute-force processing, independent of the content to process and perceptive capabilites of a human listener.
Several recent studies have shown that listeners are unlikely to perceive a complex auditory environment in its entirety. In this paper, we review a number of recent approaches which can be used to organize and simplify virtual auditory scenes, leveraging both the nature of the contents and limitations of human perception. Such approaches are independent of the spatial reproduction techniques and lead to scalable processing performance with minimal degradation in the perceived spatial and tonal quality.
They can be applied to a variety of applications and allow for rendering highly complex environments at interactive speed.

INTRODUCTION

Handling 3D audio simulation is a key factor for creating convincing interactive virtual environments. The introduction of auditory cues associated to the different components of a virtual scene together with auditory feedback associated to the user interaction enhances the sense of immersion and presence [9,17]. Our spatial auditory perception will be solicited for localizing objects in direction and distance, discriminating between concurrent audio signals and analyzing spatial characteristics of the environment (indoor vs. outdoor contexts, size and materials of the room,…). Typical situations encountered in interactive applications such as video games and simulators require processing of hundreds or thousands of sources, which is several times over the capabilities of common audio dedicated hardware. The main computational bottlenecks are a per sound source cost, which relates to the different effects desired (various filtering processes, Doppler and source directivity simulation, etc.), and the cost of spatialization, which is related to the audio restitution format used (directional filtering, final mix of the different sources, reverberation, etc.). Although a scientifically authentic [19,26] result can be achieved through physical modeling of these steps, the processing of complex sound scenes, composed of numerous direct or indirect (reflections) sound sources, can take advantage of perceptually based optimizations in order to reduce both the necessary computer resources and the amount of audio data to be stored and processed. Several auditory perceptual properties may be exploited in order to simplify the rendering pipeline with limited impact on the overall perceived audio quality. The general approach is to structure the sound scene by (1) sorting the relative importance of its components, (2) distributing properly the computer resources on the different signal processing operations and (3) handling the spatial complexity of the scene (Figure 1). These techniques, derived from psycho-acoustics, perceptual audio-coding and auditory scene analysis introduce several concepts similar to those found in computer graphics: selective, progressive and scalable rendering (e.g., visibility/view-frustum culling and geometrical/shading level-of-detail).

PRINCIPLES OF PERCEPTUALLY-BASED AURALIZATION

Masking and illusory continuity

Selective audio processing approaches build upon prior work from the field of perceptual audio coding that exploits auditory masking. When a large number of sources are present in the environment, it is very unlikely that all will be audible due to masking occurring in the human

auditory system [22]. This masking mechanism has been successfully exploited in perceptual audio coding (PAC), such as the well known MPEG I Layer 3 (mp3) standard [25] and several efficient computational models have been developed in this field. In the context of interactive applications, this approach is thus also linked to the illusion of continuity phenomena [13], although current works do not generally include explicit models for this effect. This phenomenon is implicitly used together with masking to discard entire frames of original audio content without perceived artefacts or "holes" in the resulting mixtures.

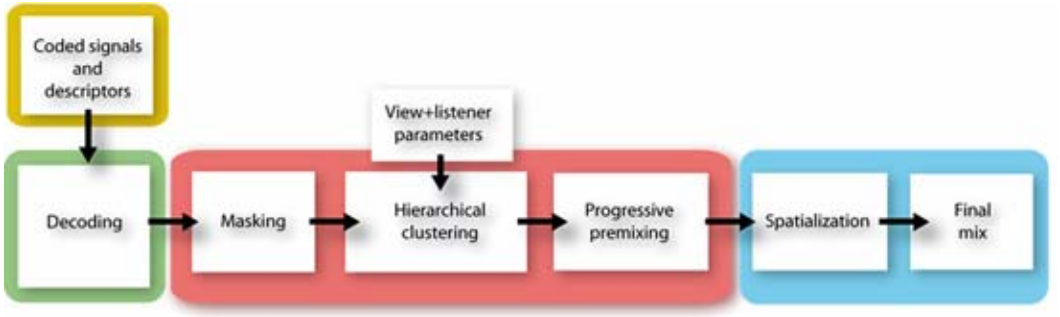

Figure 1.-Overview of a perceptually-based auralization pipeline for VR applications.

Importance and saliency of sound sources

Evaluate all possible solutions to the optimization problem required for optimal rendering of a sound scene would be computationally untractable. An alternative is to use greedy approaches which first require estimating the relative importance of each sources in order to get a good starting point. A key aspect is also to be able to dynamically adapt to the content. Several metrics can be used for this purpose such as energy, loudness or the recently introduced saliency. Recent studies have compared some of these metrics showing that they might achieve different results depending on the nature of the signal (speech, music, ambient sound "textures"). Loudness has been found to be generally leading to better results while energy is a good compromise between complexity and quality.

Limitations of spatial hearing in complex soundscapes

Human spatial hearing limitations, as measured through perceivable distance and angular thresholds [1] can be exploited for faster rendering independently of the subsequent signal processing operations. This is useful for applications where the reproduction format is not set in advance, Recent studies have also shown that our auditory localization is strongly affected in multi-source environments. Localisation performances decrease with increasing number of competing sources [3] showing various effects such as pushing effect (the source localization is repelled from the masker) or pulling effects (the source localization is attracted by the masker) which depend on the time and frequency overlapping between the concurrent sources [2]. As a result, spatial simplification can probably be performed even more aggressively as the complexity of the scene, in particular the number of sound sources, grows.

TECHNIQUES FOR DYNAMIC MASKING OF SOUND SOURCES

Interactive applications bring several additional constraints since the masking does not concern a premixed audio stream but has to be evaluated between several concurrent sound signals in course of their processing. Since scenes are generally highly dynamic, masking thresholds have to be continuously predicted and updated according to the instantaneous characteristics of the source signals and their position in space relative to the listener.

Dynamic masking of sound sources was introduced by Tsingos et al. [30] in order to limit the spatial rendering only to audible sources. The method takes advantage of pre-computed signal characterisation (power spectrum distribution and tonality index [25]) associated with each individual audio sample throughout its duration. At runtime, this information is accessed dynamically in order to predict the source's instantaneous loudness [23] anticipating for subsequent frequency-dependent attenuation linked to the source directivity, source-listener distance and possible occlusion or scattering effects. At each frame, sources can thus be sorted according to their loudness contribution at the listener's ears and summed up until they mask the remaining sources which can then be simply discarded from the rendering pipeline.

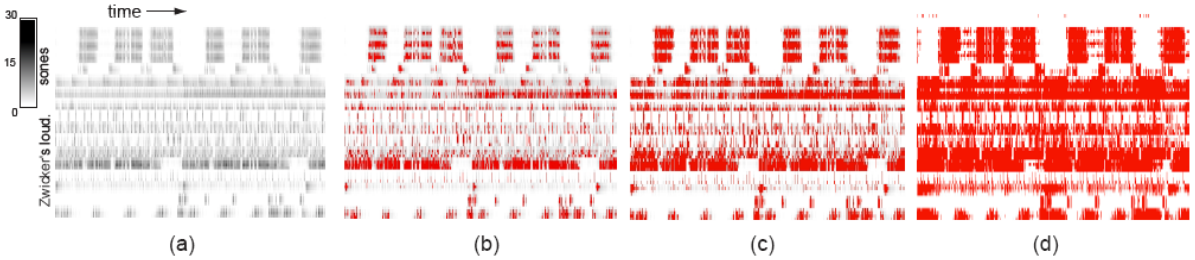

Figure 2.- (a) Loudness values (using Zwicker's loudness model) through time for the 17 tracks of a musical mix. Each track was selectively filtered and processed into 4 frequency sub-bands resulting in 68 signals to prioritize. (b) Priority map showing the first 12.5% most important frames highlighted in red. (c) First 25% most important frames and (d) first 50% most important frames.

SELECTIVE AND PROGRESSIVE SIGNAL PROCESSING

When large numbers of sound sources are still present in the scene after the auditory culling (masking) or for systems with limited processing power, per-source processing (e.g., studio-like effects [31]; Doppler effect; distance, occlusion, reverberation filtering; sub-mixing, etc.) can still represent a strong bottleneck of the audio rendering pipeline.

In recent years, several contributions were introduced that aim to bridge the gap between perceptual audio coding and audio processing in order to make audio signal processing pipelines more efficient. Fouad et al. [6] propose a level-of-detail progressive audio rendering approach in the time-domain; by processing every n^{th} sample, artefacts are introduced at low budget levels. Wand and Straßer [31] introduce an importance sampling strategy using random selection, but ignore the signal properties, thus potentially limiting the applicability of this method. A family of approaches proposed to directly process perceptually coded audio signals [15,5,28] yields faster implementations than a full decode-process recode cycle. In the context of long FIR filtering for reverberation processing, the recent work by Lee et al. [18] shows that significant improvement can be obtained by estimating whether the result of the convolution is below hearing threshold, hence reducing the processing cost.

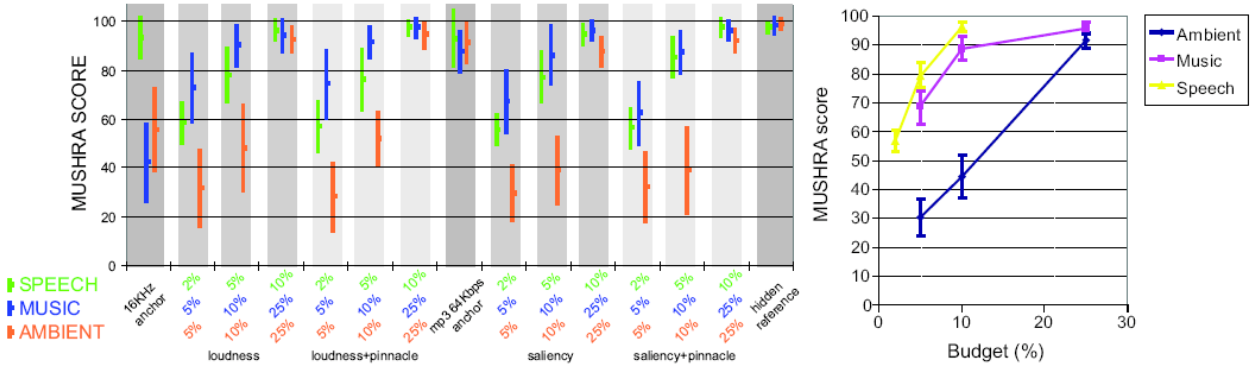

Figure 3.-Left: Average MUSHRA scores and 95% confidence intervals for our progressive processing tests. Right: Average MUSHRA scores and 95% confidence intervals as a function of budget. Note how perceived quality does not vary linearly with the processing budget and also varies depending on the type (i.e., sparseness) of the sounds.

Recent studies have proposed extensions of these approaches by concurrently prioritizing subparts of the original signals to process to guarantee a minimal degradation in the final result [13,7,21,29] (see Figure 2).

Key to such approaches is the choice of a signal representation that allows its progressive encoding and reconstruction. Tsingos et al. [29] introduce a progressive signal processing technique where the coefficients of the short-time Fourier transform (STFT) of each signal are pre-computed and stored in decreasing modulus. In real-time during the simulation, the algorithm prioritizes the signals and allocates a number of coefficients to process for each source so that a predefined budget of operations is respected. Different perceptual metrics can be used to determine the cut-off point in the list of STFT coefficients, leading to different trade off between computer efficiency and perceptual quality (see Figure 3). Moeck et al. [21], complements a loudness-based metrics with the tonality index [25]. Loud tonal or speech signals, for which the STFT representation is sparser, will require fewer coefficients than a

weaker noisier signal for transparent reconstruction. They compare the merits of this metrics with other models such as the audio saliency map proposed by Kayser et al [12]. The two metrics lead to very similar results. However, for intermediate budgets and for cases where both tonal and noisier signals are used, loudness-based prioritization improves perceived quality relative to the saliency-based alternative (Figure 3).

HIERARCHICAL CODING OF SPATIAL CUES IN SCENE-SPACE

One of the primary goals of spatial audio rendering is to reconstruct to the ears of the listeners the perceptual cues which are responsible for localizing sound in direction. The approaches described above can be used to simplify the signal processing operations required by specific 3D audio rendering algorithms. However, it is sometimes desirable to compress the spatial information of the scene in a way independent of the chosen reproduction technique to enable flexible reproduction. In this case, the coding/simplification of the spatial information must be performed in scene-space. This solution which is referred to as *scalable spatial audio rendering* can be divided into three categories:

- **Fixed basis functions/clustering**: the first set of techniques encodes the spatial cues using a number of fixed basis functions or clusters. For instance, Ambisonics [20,4] uses a spherical harmonics decomposition of the incoming sound pressure at the listening point. For binaural listening, there are approaches that decompose the HRTF onto a basis of eigen-filters corresponding to principal directions [16,11], and approaches that operate in object space explicitly grouping neighboring sound sources belonging to the same cone of directions [10], or using a hierarchical structure [31].
- **Dynamic per-object clustering**: The clustering proposed by Sibbald is an object-based method [27]. Sound sources related to an object or an area are grouped according to their distance to the listener. In the near field, secondary sound sources are created and dynamically uncorrelated in order to improve the spatial sensation. In the far field, sources are clustered together, accelerating the spatial rendering. The drawback of the method is that the clustering is evaluated on a per-object basis and does not consider all the elements of the scene.
- **Dynamic global clustering**: Tsingos et al. [30] introduced a dynamic source clustering method based on both the geometry of the scene and the signals emitted by each source (Figure 4). This is especially useful for scenes where sounds are frequently changing in time varying their shape, energy as well as in location. The algorithm flexibly allocates the required number of clusters; thus clusters are not wasted where they are not needed. The dynamic clustering is derived from the Hochbaum-Shmoys heuristic. The cost-function used for clustering combines instantaneous loudness, distance and angle. An equivalent signal for the cluster is then computed as a mixture of the signals of the clustered sound sources. A representative loudness-weighted centroid is used to spatialize the cluster according to the reproduction setup. This technique has been shown to lead to efficient rendering while maintaining very good rendering quality and minimal impact on localization-task performance, even with a small number of clusters.

APPLICATIONS

Auralization for interactive virtual environments

Figure 1 illustrates the combined use of all previous techniques in the context of a 3D audio rendering engine for complex virtual environments, such as the ones found in video games or simulators. The set of signals for all sound sources are partially decoded so as to retrieve the descriptors and are first tested for masking. Audible sources are clustered and all per-source operations (or premixing) are performed using a progressive approach. Full or scalable decoding of the source signal can be delayed up to this stage, avoiding the cost of streaming/decompressing inaudible signals from the storage media. Finally the obtained signal from each cluster is spatialized using the location of its representative and mixed into the sound output. Note that the framework accommodates both primary sources and secondary sources arising from sound scattering off surfaces.

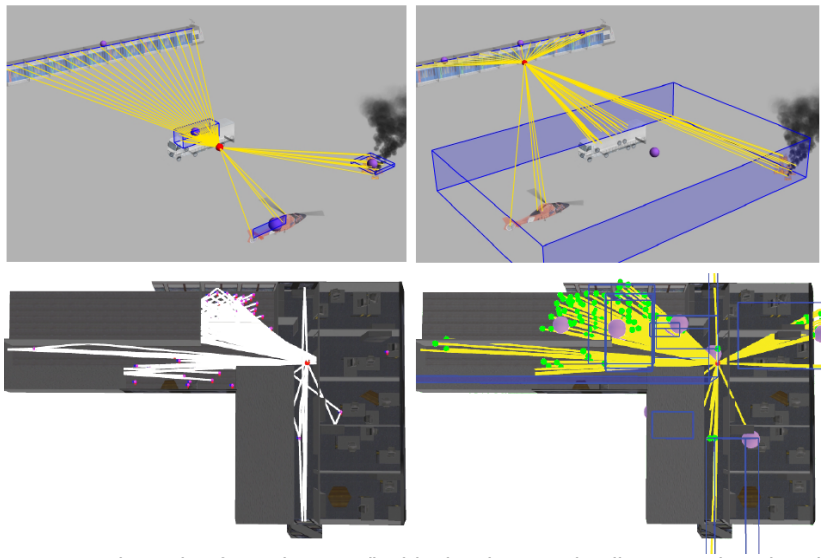

Figure 4.-Top row: note how the four clusters (in blue) adapt to the listener's location (shown in red). Bottom row: a clustering example with image-sources in a simple building environment (seen in top view). The audible image-sources, shown in green in the right-hand side image, correspond to the set of reflection paths (shown in white) in the left-hand side image

Bandwidth management

The proposed dynamic masking strategy can also be used for bandwidth management, for instance to optimize voice streaming in voice-over-IP applications. In this case, the masking estimation is integrated into a forwarding bridge. Each participant sends coded frames of audio data together with additional descriptors (energy and tonality in a small set of frequency sub-bands) to the forwarding bridge. A dynamic masking procedure is performed for each participant and only the streams audible to each participant are sent downwards. This allows for optimizing bandwidth while maintaining flexible decoding on the client side. Details can be found in [24] and demos at the following URL: http://www-sop.inria.fr/reves/OPERA/

DISCUSSION AND CONCLUSIONS

Although the perceptually-motivated approaches use simple models rather than full binaural hearing models, they were found to work very well in practice, as supported by a number of perceptual studies in [7,21,30] and are already used in some commercial game applications [21]. Their simplicity is key in providing a good trade-off between the time required for decision-making in order to optimize resources vs. the cost of the actual processing. Even with the increased computational power of recent multi-core processors, we believe scalable approaches are still likely to be required in order to tune computing resources or manage bandwidth. They make authoring easier by allowing sound designers to use as many sources as required by the application without thinking about rendering limitations. This is of particular interest for sources linked to physical simulations, such as contact sounds. A strong benefit of these approaches is also to be content-adaptive. An issue is the extra space required for storing additional sound descriptors which might become an issue on systems with limited memory. However, this information can be made quite compact (about 1 Kbyte per sec. of data) and required storage space should not be a strong limitation for a wide range of platforms.

Several extensions could be made to the approaches. First, most approaches for masking or saliency do not account for the 3D location of the sources or use very crude approximations. Spatial unmasking effects inducing variations of masking thresholds due to the relative location of masker and maskee are likely to play a major role in the context of 3D audio perception. Extending current masking approaches to account for this phenomena would be of major interest. Extending the clustering techniques to account for the effect of reverberation would also be of primary interest. Finally, while the main target applications of 3D audio rendering are simulation and gaming, no spatial audio rendering work to date evaluates the influence of combined visual and audio restitution on the required quality of the simulation. However, a vast amount of literature in neurosciences suggests that cross-modal effects, such as ventriloquism, might significantly affect 3D audio perception [8]. A preliminary study has been conducted in [21] but additional work is required in this domain.

Acknowledgements
The author would like to thank all the people involved at various stages of this research: E. Gallo, G. Lemaitre, T. Moeck, N. Bonneel, F. Firsching, G. Drettakis and I. Viaud-Delmon. Parts of this research were funded by the EU IST projects *CREATE* and *CROSSMOD* and the French RNTL project *OPERA*. See also http://www-sop.inria.fr/reves/audile.

References

[1] BEGAULT D. 1994, 3D Sound for Virtual Reality and Multimedia. Academic Press Professional.
[2] BEST, V., SHAIK, A. V., JIN, C., AND CARLILE, S. 2005. Auditory spatial perception with sources overlapping in frequency and time. Acta Acustica 91, 421–428.
[3] BRUNGART, D. S. AND SIMPSON, B. S. 2005. Localization in the presence of multiple simultaneous sounds. Acta Acustica 91, 471–479.
[4] DANIEL, J. AND MOREAU, S. 2004. Further study of sound field coding with higher order Ambisonics. In Proceedings of the 116th Audio Engineering Society Convention, preprint 6017.
[5] DARLINGTON, D., DAUDET, L., AND SANDLER, M. 2002. Digital audio effects in the wavelet domain. In Proceedings of COST-G6 Conference on Digital Audio Effects, DAFX2002.
[6] FOUAD, H., HAHN, J., AND BALLAS, J. 1997. Perceptually based scheduling algorithms for real-time synthesis of complex sonic environments. In Proc.of the 1997 Intl. Conf. on Auditory Display, Palo Alto, USA.
[7] GALLO, E., LEMAITRE, G., AND TSINGOS, N. 2005. Prioritizing signals for selective realtime audio processing. In Proceedings of International Conference on Auditory Display (ICAD 2005).
[8] HAIRSTON, W. D., WALLACE, M. T., VAUGHAN, J. W., STEIN, B. E., NORRIS, J. L., AND SCHIRILLO, J. A. 2003. Visual localization ability influences cross-modal bias. Journal of cognitive neuroscience 15, 20–29.
[9] HENDRIX, C. AND BARFIELD, W. 1996. The sense of presence within auditory virtual environments. Presence: Teleoperators and Virtual Environments 5, 290–301.
[10] HERDER, J. 1999. Optimization of sound spatialization resource management through clustering. Journal of Three Dimensional Images, 3D-Forum Society 13, 59–65.
[11] JOT, J.-M., WARDLE, S., AND LARCHER, V. 1998. Approaches to binaural synthesis. In Proceedings of the 105th Audio Engineering Society Convention, preprint 4861.
[12] KAYSER, C., PETKOV, C. I., LIPPERT, M., AND LOGOTHETIS, N. K. 2005. Mechanisms for allocating auditory attention: An auditory saliency map. Current Biology 15, 1943–1947.
[13] KELLY, M. AND TEW, A. 2002. The continuity illusion in virtual auditory space. In Proceedings of the 112th Audio Engineering Society Convention, preprint 5548.
[14] KLEINER, M., DALENBACK, B.-I., AND SVENSSON, P. 1993. Auralization - An overview. Journal of Audio Engineering Society, 861–875.
[15] LANCIANI, C. A. AND SCHAFER, R. W. 1997. Psychoacousticallybased processing of mpeg-i layer 1-2 encoded signals. In Proceedings IEEE Signal Processing Society 1997 Workshop on Multimedia Signal Processing. 53–58.
[16] LARCHER, V., JOT, J.-M., GUYARD, G., AND WARUSFEL, O. 2000. Study and comparison of efficient methods for 3D audio spatialization based on linear decomposition of hrtf data. In Proceedings of 108th Audio Engineering Society Convention, preprint 5097.
[17] LARSSON, P., V¨ASTFJ¨ALL, D., AND KLEINER, M. 2002. Better presence and performance in virtual environments by improved binaural sound rendering. Proc. of the 22nd AES Conf. on Virtual, Synthetic and Entertainment Audio, 31–38.
[18] LEE, W.-C., LIU, C.-H. Y. C.-M., AND GUO, J.-I. 2003. Perceptual convolution for reverberation. In Proceedings of 115th AES Convention.
[19] LOKKI, T., HIIPAKKA, J., AND SAVIOJA, L. 2001. A framework for evaluating virtual acoustic environments. In Proceedings of the 110th Audio Engineering Society Convention, preprint 5317.
[20] MALHAM, D. AND MYATT, A. 1995. 3d sound spatialization using Ambisonic techniques. Computer Music Journal 19, 58–70.
[21] MOECK, T., BONNEEL, N., TSINGOS, N., DRETTAKIS, G., VIAUD-DELMON, I., AND ALLOZA, D. 2007. Progressive perceptual audio rendering of complex scenes. Proceedings of ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games, I3D2007.
[22] MOORE, B. C. 1997. *An introduction to the psychology of hearing*. Academic Press, 4th edition.
[23] MOORE, B. C. J., GLASBERG, B., AND BAER, T. 1997. A model for the prediction of thresholds, loudness and partial loudness. Journal of the Audio Engineering Society 45, 224–240.
[24] NAGLE A., TSINGOS N., LEMAITRE G. AND SOLLAUD A., 2007. On-the-fly Auditory Masking for Scalable VOIP Bridges. Proc. of the 30th AES Intl. Conf. on Intelligent Audio Environments, Saariselka, Finland,.
[25] PAINTER, E. M. AND SPANIAS, A. S. 2000. Perceptual coding of digital audio. In Proceedings of the IEEE, 88. 451–515.
[26] PELLEGRINI, R. 2001. Quality assessment of auditory virtual environments. In Proc. of the 2001 Intl. Conf. on Auditory Display.
[27] SIBBALD, A. 2001. MacroFX algorithms. White paper. www.sensaura.co.uk/whitepapers/.
[28] TOUIMI, A. B., EMERIT, M., AND PERNAUX, J.-M. 2004. Efficient method for multiple compressed audio streams spatialization. In Proceeding of ACM 3rd International. Conference On Mobile and Ubiquitous multimedia.
[29] TSINGOS, N. 2005. Scalable perceptual mixing and filtering of audio signals using an augmented spectral representation. In DAFx'05 Conference Proceedings. 277–282.
[30] TSINGOS, N., GALLO, E., AND DRETTAKIS, G. 2004. Perceptual audio rendering of complex virtual environments. In Proc. of SIGGRAPH '04, ACM Press, New York, NY, USA, 249–258.
[31] WAND, M. AND STRASSER, W. 2004. Multi-resolution sound rendering. In Eurographics Symposium on Point-Based Graphics.
[32] ZOLZER, U. 2002. DAFX - Digital audio effects. John Wiley & Sons.