



Audio Engineering Society Convention Paper

Presented at the 131st Convention
2011 October 20–23 New York, USA

This Convention paper was selected based on a submitted abstract and 750-word precis that have been peer reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This convention paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Binaural Reproduction of Stereo Signals Using Upmixing and Diffuse Rendering

Christof Faller¹ and Jeroen Breebaart²

¹*ILLUSONIC LLC, St-Sulpice, Switzerland*

²*ToneBoosters, Eindhoven, The Netherlands*

Correspondence should be addressed to Christof Faller (christof.faller@illusonic.com)

ABSTRACT

In this paper, benefits and challenges related to binaural rendering for conventional stereo content are explained in terms of width of the sound stage, timbral changes, perceived distance and naturalness of phantom sources. To resolve some of the identified issues, a two-stage process consisting of a spatial decomposition followed by dedicated post processing methods is proposed. In the first stage, several direct sound source signals and additional ambience components are extracted from the stereo content. These signals are subsequently processed with dedicated algorithms to render virtual sound sources by means of HRTF or BRIR convolution, and to render an additional diffuse sound field with the correct inter-aural coherence properties based on the extracted ambience signals. It is argued that this approach results in a wider sound stage, more natural and accurate spatial imaging of sound sources, and resolves the “here and now” versus the “there and then” duality for room acoustic simulation in binaural rendering methods.

1. INTRODUCTION

Binaural rendering [1, 2] is relatively widely used for applications which present multi-channel surround audio over headphones, e.g. gaming and mobile video. In these applications, consumers can experience multi-channel surround sound without the need of a dedicated loudspeaker setup, and without disturbing other persons present in the same physi-

cal environment.

Another niche market is the use of binaural rendering for mastering and monitoring with headphones. Professional audio engineers are taught and trained to create mixes and perform the mastering process on loudspeakers, and often switch between different loudspeaker sets to verify whether their work

translates well on these. Such switching requires the availability of a multitude of loudspeaker setups and reproduction environments, which can be costly and space consuming. Additionally, amateur and semi-professional audio engineers are increasingly benefiting from virtual audio reproduction techniques that are commercially available. This group of audio enthusiasts often have a small music studio in their homes, and hence the use of virtual loudspeakers over headphones minimizes the disturbance for their family members.

The rendering of binaural audio over headphones comes with a set of relatively well-known challenges (see [3, 4] for overviews). For the sake of completeness, and to motivate the work described in this paper, the various challenges are reviewed:

Anthropometric inter-subject variability:

Without individualized head-related transfer functions (HRTFs) or individualized binaural room impulse responses (BRIRs), sound source localization can be subject to errors due to a mismatch between the localization cues of the subject’s own ears and those present in the employed transfer functions [5, 6].

Conflicting auditory and proprioceptive cues:

The absence of the correction for head rotation on auditory localization cues can result in an increase in localization errors, especially in the front/back direction [7, 8].

Positioning accuracy versus cost trade-off:

A limited number of (virtual) loudspeakers often necessitates the use of amplitude panning techniques, which come with certain limitations related to their maximum allowed spacing. As a rule of thumb, amplitude panning only works for frontal positions and is not applicable to lateral positions, while the maximum allowed angular distance between two loudspeakers is approximately 60 degrees [9–11].

Perceived externalization: To obtain a convincing out-of-head sound source localization and a correct sound source distance, proper simulation of the acoustic environment is required [12–16]. The use of such simulation may however result in undesirable timbral changes due

to comb-filter interactions between direct sound and reflections.

“Here and now” or “there and then” dualism:

Binaural algorithms often involve room acoustic simulation (“here and now”) to result in a sufficient out-of-head percept. However, this virtual reproduction environment can be significantly different from the environment of the actual recording (e.g., “there and then”), resulting in potentially conflicting or ambiguous room acoustic cues.

Without individualized head related transfer functions (HRTFs) or binaural room impulse responses (BRIRs), and without head tracking, externalization is often limited and front/back confusions can occur. Despite of this, for multi-channel surround audio, the potential benefit of using binaural rendering versus a downmix, e.g. an ITU downmix [17], is obvious: the rear channels are rendered spatially distinctly from the front channels, enhancing the extent of the auditory spatial image compared to stereo presentation.

On the other hand, the enhancement of two-channel stereo headphone playback by means of binaural rendering is not used so widely. This could be attributable to a potentially negative balance between benefits and drawbacks of the existing approaches as listed above. In this paper, we therefore discuss solutions with the goal of targeting the last three challenges.

More specifically, we propose the use of upmixing techniques (spatial decomposition of stereo signals) [18–20] to improve the accuracy of localization cues from phantom images that are encapsulated in the stereo content, and to resolve issues related to room acoustic simulation. The idea of upmixing algorithms as pre-processing stage for HRTF convolution per se is not new (see [21, 22] for some practical examples). Our novelty relies in the fact that we use signals extracted from the stereo content itself for room acoustic simulation as well. For this purpose, we propose that stereo signals are first decomposed into direct and ambient signals, which are further processed and rendered to enhance binaural stereo rendering and enable new trade-offs:

- Direct sound is converted to left, right, and center signals. For these signals, corresponding left, right, and center HRTFs or BRIRs are used, enabling the use of larger opening angles than 60 degrees without degrading the (phantom) center.
- Ambient sound is rendered with left and right HRTFs or BRIRs, possibly with a larger opening angle than direct sound, further widening the resulting auditory spatial image. Alternatively, ambient sound is rendered as diffuse sound field, similarly as proposed in [23].
- Surround ambience channels are generated using upmix techniques and binaurally rendered to further provide auditory spatial image enhancement. Alternatively, surround ambience is rendered as diffuse sound field.

The paper is organized as follows. The upmixing technique we are using to decompose a stereo signal into direct, ambient, and surround signals is summarized in Section 2. Section 3 discusses conventional binaural rendering and its application to upmixed signals, including matrix surround rendering. Section 4 describes the use of different rendering techniques for direct and ambient sound, i.e. binaural and diffuse rendering. Discussion and reporting of informal listening impressions are in Section 5. The conclusions are provided in Section 6.

2. UPMIXING

We are using an upmix, decomposing a stereo signal into three dry direct sound signals: left $d_L(n)$, center $d_C(n)$, and right $d_R(n)$, where n is the discrete time index of the sampled signals. Further, left and right ambience signals, $a_L(n)$ and $a_R(n)$, and left and right surround signals, $s_L(n)$ and $s_R(n)$, are computed.

The signals $d_L(n)$, $d_C(n)$, $d_R(n)$, $a_L(n)$, and $a_R(n)$ are generated as described in [19, 20] for a 3-channel stereo upmix. The surround signals are obtained by

$$\begin{aligned} s_L(n) &= h(n) \star a_L(n - \tau) \\ s_R(n) &= h(n) \star a_R(n - \tau), \end{aligned} \quad (1)$$

where \star denotes linear convolution and $h(n)$ is the impulse response of a first or second order low-pass

filter with a cutoff frequency in a range between 2 and 10 kHz, to mimic the high frequency attenuation property of reverberation resulting from frequency-dependent wall absorptivity. The delay τ ensures that the surround signals are not localizable due to the precedence effect [24]. We found a delay corresponding to 40 ms suitable. Alternatively, or additionally, the filters $h(n)$ may comprise multiple echos or late reverberation, and may be different for the left and right sides, respectively, to result in lower correlation or coherence between the signals $s_L(n)$ and $s_R(n)$.

3. BINAURAL RENDERING

3.1. HRTFs, BRIRs, and BRIR models

A pair of head related transfer functions (HRTFs) models the transfer of sound from a source at a specific position to left and right ear entrances of a listener. The first part of binaural room impulse responses (BRIRs), i.e. the transfer functions of direct sound between a source and ear entrances, corresponds to HRTFs. Additionally, BRIRs capture early and late reflections reaching a listener's ears.

Using FIR filters to model BRIRs is computationally expensive. On the other hand, the early reflections and reverberation part of the BRIRs are crucial for externalization. For achieving lower computational complexity, BRIRs are often modeled with HRTFs and a reverberator, see e.g. [12, 16, 25–27]. In the following, unless otherwise noted, when we use the term HRTF we always mean HRTF, BRIR, or BRIR model.

3.2. Conventional binaural stereo rendering

The standard way of rendering stereo signals binaurally is to apply pairs of HRTFs to the left and right stereo signal channels. This corresponds to rendering the left and right channels as left and right sources, illustrated in Figure 1. The angle α is usually chosen to be 60 degrees in accordance with a standard stereo loudspeaker setup. If externalization is limited, as is often the case, such a binaural signal may evoke a more narrow stereo image than the original stereo signal, due to the choice of α being only 60 degrees. As mentioned before, larger angles are problematic with amplitude panned sources in the stereo signal.

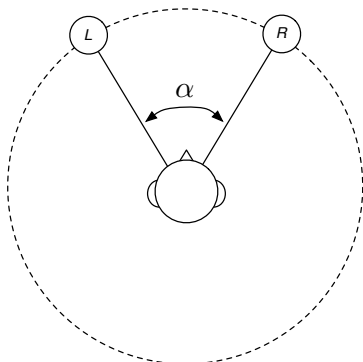


Fig. 1: The left and right stereo channels are rendered as left and right binaural sources.

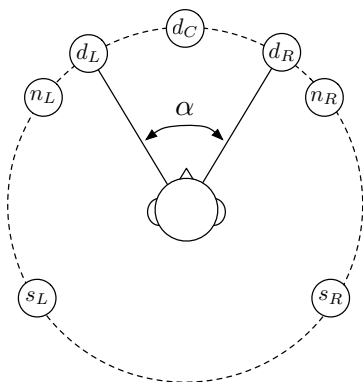


Fig. 2: The upmix channels are rendered as binaural sources at different directions.

3.3. Binaural rendering of the upmix signals

The most straightforward option to render upmix signals is to use a dedicated pair of HRTFs for each signal. This approach is illustrated in Figure 2. The angle α can now be chosen larger because the center direct signal is now reproduced with its own dedicated HRTFs instead of being a phantom source. The left and right ambience signals, a_L and a_R , may be rendered with the same HRTFs as the left and right direct signals. Alternatively, HRTFs more to the side may be used for rendering a_L and a_R , further widening the auditory spatial image. The surround ambience channels, s_L and s_R , are rendered with side/rear HRTFs.

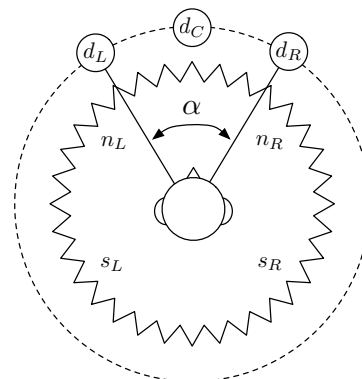


Fig. 3: The direct sound channels are rendered as binaural sources and the ambient and surround channels as diffuse sound field.

3.4. Binaural rendering of matrix surround

Matrix surround [28] downmixes (“Lt, Rt”) can be rendered like multi-channel surround audio signals, by using an enhanced stereo decomposition [29].

In addition to front direct sound channels (d_L , d_R , d_C), also rear direct sound channels (d_{Ls} , d_{Rs}) are computed by considering matrix surround cues (positive and negative amplitude ratios [29]). The d_{Ls} and d_{Rs} signals are rendered using distinct HRTFs.

4. BINAURAL AND DIFFUSE RENDERING

Alternatively to what has been described in Sections 3.3 and 3.4, the stereo and/or surround ambience channels, a_L , a_R , s_L and s_R , may be rendered under the assumption that they represent a diffuse sound field. This is illustrated in Figure 3.

Diffuse sound appears as left and right ear input signals which are nearly fully correlated at low frequencies with decreasing correlation as frequency increases [23, 30]. It has also been shown that the correct correlation is required for a proper distance percept [31]. In the sequel, it is assumed that for left and right ambience and surround signals, $x_L = a_L + s_L$, $x_R = a_R + s_R$, the expected energies are equal and the signals are assumed to be independent:

$$\langle x_L^2 \rangle = \langle x_R^2 \rangle, \quad (2)$$

$$\langle x_L x_R \rangle = 0. \quad (3)$$

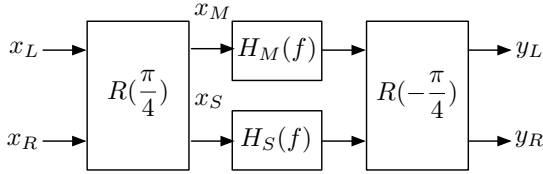


Fig. 4: Diffuse rendering: cascade of rotation, filtering, and inverse rotation to realize a frequency-dependent coherence function.

Under these assumptions, a desired coherence characteristic $\rho(f)$ as a function of frequency f can be realized by the subsequent steps of stereo rotation, filtering, and inverse rotation as visualized in Fig. 4.

The signal pair x_L, x_R is first rotated using a rotation matrix $R(\alpha)$ with an angle $\alpha = \pi/4$ to obtain a mid/side decomposition x_M, x_S :

$$\begin{bmatrix} x_M \\ x_S \end{bmatrix} = \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{bmatrix} \begin{bmatrix} x_L \\ x_R \end{bmatrix}. \quad (4)$$

Subsequently, the signals x_M, x_S are filtered with filters h_M, h_S with corresponding frequency responses $H_M(f)$ and $H_S(f)$, respectively. The filters are designed such that they adhere to an overall energy preservation requirement:

$$H_M(f)H_M^c(f) + H_S(f)H_S^c(f) = 2, \quad (5)$$

where $(\cdot)^c$ is the complex conjugation operator. The second requirement is that the diffuse rendered signals y_L, y_R ,

$$\begin{bmatrix} y_L \\ y_R \end{bmatrix} = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix} \begin{bmatrix} h_M \star x_M \\ h_S \star x_S \end{bmatrix}, \quad (6)$$

adhere to the desired coherence function $\rho(f)$. This can be formulated in terms of filter responses as:

$$\rho(f) = \frac{(H_M(f) + H_S(f))(H_M(f) - H_S(f))^c}{H_M(f)H_M^c(f) + H_S(f)H_S^c(f)}. \quad (7)$$

In the case that the phase responses of $H_M(f)$ and $H_S(f)$ are identical (for example when using linear-phase filters), this simplifies to:

$$|H_M(f)| = \sqrt{1 + \rho(f)}, \quad (8)$$

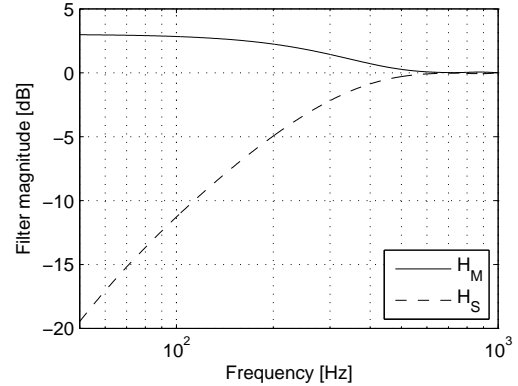


Fig. 5: Magnitude for linear-phase filters $H_M(f)$ and $H_S(f)$ using the coherence model adopted from [32].

$$|H_S(f)| = \sqrt{1 - \rho(f)}. \quad (9)$$

The resulting magnitude spectra of $H_M(f)$ and $H_S(f)$ are shown in Fig. 5. The coherence function $\rho(f)$ used to generate these filter responses was adopted from [32]. The filter $H_M(f)$ (solid line) which is applied to the mid signal x_M has a gain of approximately +3 dB at low frequencies, and gradually decreases toward 0 dB when frequency increases. The filter $H_S(f)$ (dashed line) which is applied to the side signal x_S , has a high-pass character with a -3 dB cut-off frequency of approximately 260 Hz.

Another option is to only render the surround signals s_L and s_R as diffuse sound and render the stereo ambience signals a_L and a_R using HRTFs, illustrated in Figure 6. This can be motivated as follows: the surround signals are generated with the goal to only represent room signals. This can not be generally said about the ambience signals. The ambience signals, as obtained by the stereo decomposition, are often full-band and contain also un-correlated direct sound.

The scheme shown in Figure 7 illustrates the various rendering options that have been described so far. The stereo signal is decomposed into direct sound, ambient, and surround signals. Direct sound signals are rendered with binaural rendering (rendering with HRTFs, BRIRs, or BRIR models). Ambient and

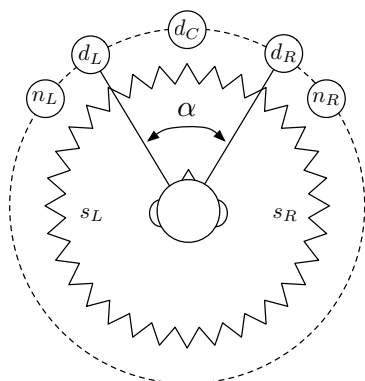


Fig. 6: The direct and ambience signals are rendered as binaural sources and the surround signals as diffuse sound field.

surround signals are either rendered with binaural or diffuse rendering.

5. DISCUSSION

The approach of using upmixing techniques with binaural and diffuse rendering has several potential advantages over the use of binaural stereo rendering alone or the use of additional means of acoustic room simulation to achieve a sufficient level of externalization:

- The filter transfer functions $H_m(f)$ and $H_s(f)$ can be realized with relatively simple, low-order filters. Consequently, the computational complexity of this approach is often significantly lower compared to HRTF or BRIR convolution, or room acoustic simulation as additional module.
- By using the ambience components from the original recording instead of a re-creation of a virtual playback environment, the “here and now” versus “there and then” dualism is solved.
- Diffuse rendering resolves the issue of having to select azimuth and elevation angles for these components if they would have to be reproduced by acoustic sources with a specific direction or position.

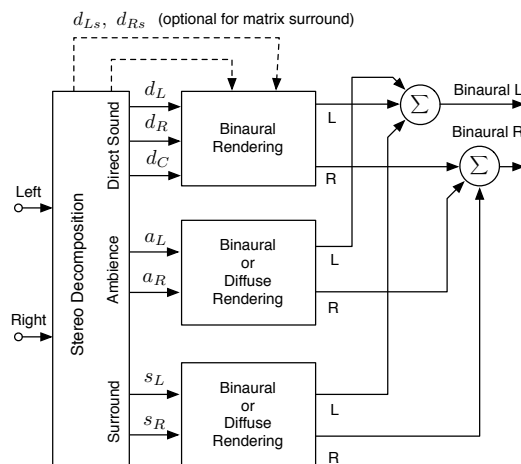


Fig. 7: Scheme for binaural rendering based on stereo decomposition.

- The direct signals left $d_L(n)$, center $d_C(n)$, and right $d_R(n)$ can be rendered with HRTFs with a larger angle than 60 degrees due to the extraction of the center channel $d_C(n)$, and their positions can be modified independently from the ambience components. This property is of great interest when employing a head tracker to resolve conflicting auditory and proprioceptive cues. In such case, the direct signals $d_L(n)$, $d_C(n)$, $d_R(n)$ can be rendered at a position depending on the position of the head, while the ambience signals can be generated as described above without any dependency on the head orientation.

Several informal listening experiments were conducted with the approach described in Section 4. From listeners’ feedback, we learned that listeners appreciate the wider, and more accurate sound stage and the more natural room simulation as compared to conventional binaural rendering methods employing a reverberation module. This informal result is also in line with more formal tests conducted by [4], that also showed that a sound stage that is wider than 60 degrees is often preferred, provided that the center channel is properly reproduced.

6. CONCLUSIONS

In this paper, we explained the benefits and drawbacks of binaural rendering for stereo audio content. It was explained that there exists a trade-off between sound field width (i.e., the opening angle of the virtual speakers) and the naturalness of the center source. Furthermore, when HRTFs are applied to stereo signals, room acoustic simulation is likely to be required for a proper distance and out-of-head percept. Such room simulation may interfere with the room acoustic properties encapsulated in the recording itself, and may result in undesirable timbre changes due to comb-filter interactions between reflections and the direct sound.

To overcome and resolve these trade-offs, the use of spatial decomposition (or upmixing) methods is proposed, in combination with dedicated post processing algorithms. The spatial decomposition method splits the incoming audio signals into several direct and ambience components to allow independent post processing of these signals. Furthermore, the decomposition method reduces the need for amplitude panning and phantom imaging by extracting more than two direct channels, allowing a wider sound stage when compared to (virtual) stereo reproduction without degradation of the center channel.

The direct signals can be rendered using pairs of HRTFs associated with virtually any sound source position, while the ambience components can be processed to simulate a diffuse sound field as captured by the two ears. This can be achieved by relatively simple operations such as stereo field rotation and filtering.

Informal listening tests confirmed that the proposed method results in an improved attractiveness for the application of binaural rendering algorithms with conventional stereo content. In future research we will focus on further refinement of the proposed approach.

7. REFERENCES

- [1] B. B. Bauer, "Stereophonic earphones and binaural loudspeakers," *J. Audio Eng. Soc.*, vol. 9, pp. 148–151, 1961.
- [2] J. Huopaniemi, *Virtual Acoustics and 3D Sound in Multimedia Signal Processing*, Ph.D. thesis, Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland, 1999, Rep. 53.
- [3] D. R. Belgault, "Challenges to the successful implementation of 3-d sound," *J. Audio Eng. Soc.*, vol. 39, pp. 864–870, 1991.
- [4] J. Breebaart and E. Schuijers, "Phantom materialization: A novel method to enhance stereo audio reproduction on headphones," *IEEE Trans. on audio, speech and language processing*, vol. 16, no. 8, pp. 1503–1511, 2008.
- [5] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, 1996.
- [6] F. L. Wightman and D. J. Kistler, "Individual differences in human sound localization behavior," *J. Acoust. Soc. Am.*, vol. 99, pp. 2470–2500, 1996.
- [7] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Am.*, vol. 105, pp. 2841–2853, 1999.
- [8] U. Horbach, A. Karamustafaoglu, R. Pellegrini, P. Mackensen, and G. Theile, "Design and applications of a data-based auralization system for surround sound," in *Proc. 106th AES convention*, Munich, Germany, 1999.
- [9] G. Theile and G. Plenge, "Localization of lateral phantom sources," *J. Audio Eng. Soc.*, vol. 25, no. 4, pp. 196–200, 1977.
- [10] J. C. Bennett, K. Barker, and F. O. Edeko, "A new approach to the assessment of stereophonic sound system performance," *J. Audio Eng. Soc.*, vol. 33, no. 5, pp. 314–321, May 1985.
- [11] V. Pulkki, "Localization of amplitude-panned sources I: Stereophonic panning," *J. Audio Eng. Soc.*, vol. 49, no. 9, pp. 739–752, 2001.
- [12] D. R. Belgault, "Perceptual effects of synthetic reverberation on 3-D audio systems," in *Proc. 91th AES convention*, New York, USA, 1991.

- [13] Sren H. Nielsen, "Auditory distance perception in different rooms," *J. Audio Eng. Soc.*, vol. 41, no. 10, pp. 755–770, 1993.
- [14] A. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," *Nature*, vol. 397, pp. 517–520, Feb. 1999.
- [15] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *The Journal of the Acoustical Society of America*, vol. 111, pp. 1832, 2002.
- [16] B. G. Shinn-Cunningham, "The perceptual consequences of creating a realistic, reverberant 3-D audio display," in *Proc. of the international congress on acoustics*, Kyoto, Japan, April 2004.
- [17] Rec. ITU-R BS.775, *Multi-Channel Stereophonic Sound System with or without Accompanying Picture*, ITU, 1993, <http://www.itu.org>.
- [18] C. Avendano and J.-M. Jot, "Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix," in *Proc. ICASSP, Orlando, Florida*, May 2002, vol. 2, pp. 1957–1960.
- [19] C. Faller, "Multi-loudspeaker playback of stereo signals," *J. of the Aud. Eng. Soc.*, vol. 54, no. 11, pp. 1051–1064, Nov. 2006.
- [20] J. Breebaart and C. Faller, *Spatial Audio Processing: MPEG Surround and Other Applications*, Wiley, Jan. 2008.
- [21] M.R. Bai and Geng-Yu Shih, "Upmixing and downmixing two-channel stereo audio for consumer electronics," *Consumer Electronics, IEEE Transactions on*, vol. 53, no. 3, pp. 1011–1019, Aug. 2007.
- [22] Michael Goodwin and Jean-Marc Jot, "Spatial audio scene coding," in *Proc. 125th Audio Engineering Society Convention*, Oct. 2008.
- [23] F. Menzer and C. Faller, "Stereo-to-binaural conversion using interaural coherence matching," in *Preprint 128th Conv. Aud. Eng. Soc.*, May 2010.
- [24] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654, Oct. 1999.
- [25] William Gardner, "Reverberation algorithms," in *Applications of Digital Signal Processing to Audio and Acoustics*, Mark Kahrs and Karlheinz Brandenburg, Eds., vol. 437 of *The Kluwer International Series in Engineering and Computer Science*, pp. 85–131. Springer US, 2002.
- [26] F. Menzer and C. Faller, "Binaural reverberation using a modified jot reverberator with frequency-dependent interaural coherence matching," in *Preprint 126th Conv. Aud. Eng. Soc.*, May 2009.
- [27] F. Menzer, *Binaural Audio Signal Processing Using Interaural Coherence Matching*, Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, Apr. 2010, Thesis No. 4643, <http://library.epfl.ch/theses/?nr=4643>.
- [28] J. M. Eargle, "Multichannel stereo matrix systems: An overview," *IEEE Trans. on Speech and Audio Proc.*, vol. 19, no. 7, pp. 552–559, July 1971.
- [29] C. Faller, "Matrix surround revisited," in *Proc. 30th Int. Conv. Aud. Eng. Soc.*, March 2007.
- [30] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Digital Signal Processing, 2009 16th International Conference on*. IEEE, July 2009, pp. 1–4.
- [31] A. W. Bronkhorst, "Effect of stimulus properties on auditory distance perception in rooms," in *Physiological and psychophysical bases of auditory function*, D. J. Breebaart, A. J. M. Houtsma, A. Kohlrausch, V. Prijs, and R. Schoonhoven, Eds., pp. 184–191. Shaker, Maastricht, The Netherlands, 2001.
- [32] M. Jeub, M. Schäfer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1732–1745, 2010.