GENTLE ACOUSTIC CROSSTALK CANCELATION USING THE SPECTRAL DIVISION METHOD AND AMBIOPHONICS

Jens Ahrens, Mark R. P. Thomas, Ivan Tashev

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
{*jeahrens,markth,ivantash*}*@microsoft.com*

ABSTRACT

We propose the concept of gentle acoustic crosstalk cancelation, which aims at reducing the crosstalk between a loudspeaker and the listener's contralateral ear instead of eliminating it completely as aggressive methods intend to do. The expected benefit is higher robustness and a tendency to collapse less unpleasantly. The proposed method employs a linear loudspeaker array and exhibits two stages: 1) Use the Spectral Division Method to illuminate the ipsilateral ear using constructive interference of the loudspeaker signals. This approach provides only little channel separation between the listener's ears at frequencies below approximately 2000 Hz. 2) There we additionally use destructive interference by Recursive Ambiophonics Crosstalk Elimination (RACE). RACE was chosen because of its tendency to collapse gently. In a sample scenario with realistic parameters, the proposed method achieves around 20 dB of channel separation between 700 Hz and 9000 Hz, which appears to be sufficient to achieve full perceived lateralization when only one ear is illuminated.

Index Terms— Acoustic crosstalk cancelation, Spectral Division Method, Ambiophonics, transaural, loudspeaker array

1. INTRODUCTION

Many spatial audio presentation methods rely on the independent control of the sound pressure at the two ear drums of a listener. The most prominent examples are methods that employ head-related transfer functions (HRTFs) [1]. The crosstalk that arises between the ears of a listener in headphone presentation is negligible in most situations, which makes it particularly well suited for HRTF-based approaches. When signals that carry HRTF information are presented via loudspeakers in free space then *crosstalk cancelation* has to be applied, which refers to reducing or eliminating the signal radiated by a loudspeaker in free space that arrives at the listener's *contralateral* ear. The *ipsilateral* ear is defined here as the ear that is primarily illuminated by the loudspeaker under consideration. The literature on crosstalk cancelation is extensive so that the following citations can only be representative but not complete.

Crosstalk cancelation dates back to the 1960s and typically assumes a two-loudspeaker setup [2, 3, 4] or a setup with a slightly higher number of loudspeakers [5, 6, 7]. It is also referred to as *transaural* presentation. All methods require the listener's head to be either in a predefined location or its position to be tracked. The solution is usually obtained by some sort of numeric inversion of the transfer function of the crosstalk path(s). One common problem is the circumstance that this type of crosstalk cancelation is very sensitive towards inaccuracies of the tracking system or inaccuracies of the estimation of the crosstalk path, which can make them collapse in an unpleasant way. Since all these methods aim at completely eliminating the crosstalk we propose to categorize them as *aggressive* crosstalk cancelation.

Some applications do not require the crosstalk to be eliminated completely. Full lateralization of virtual sound sources can be achieved with a channel separation of slightly more than 20 dB without explicit application of HRTFs [1]. We therefore propose in this paper the concept of *gentle* crosstalk cancelation, the core idea of which is reducing the crosstalk just enough so that a given desired result such as full lateralization is achieved. The expected benefit is an increase of the robustness and a gentler collapse. Examples from the literature that are comparable to the presented approach are [8, 9]. Both approaches are two-stage with the first stage being either a circular loudspeaker array above the listener in order to evoke a sound field that maximizes the natural head shadowing between the ears [8] or a beamforming approach [9], respectively. Both approaches employ an aggressive crosstalk canceler in the second stage and over the entire frequency range.

We propose an approach to gentle crosstalk cancelation that employs linear loudspeaker arrays in order provide more freedom in terms of possible listener positions compared to [8]. There are indications that constructive interference of the sound fields emitted by loudspeakers is more robust than destructive interference in terms of transducer mismatch and the like. We therefore rely on analytic sound field synthesis wherever it is possible. It turns out that a careful design of the synthesized sound field together with the shadowing of the listener's head already provide considerable crosstalk cancelation at frequencies above 2000 Hz. Below 2000 Hz, however, constructive interference does not allow for a sufficiently narrow beam towards to ipsilateral ear and also the shadowing of the contralateral ear is reduced because of diffraction.

At this lower frequency range we propose to apply Recursive Ambiophonics Crosstalk Elimination (RACE) [10]. Although RACE relies on destructive interference, it tends to collapse very gently in that the perceived spaciousness is reduced but the timbre of the signal is hardly affected [11].

We will describe this two-stage approach in detail and analyze its properties based on numeric simulations of a sample scenario.

2. SPECTRAL DIVISION METHOD

The Spectral Division Method (SDM) is an analytic approach for sound field synthesis and was proposed in [12]. For the case of linear loudspeaker arrays, it allows for prescribing the synthesized sound pressure along a reference line that is parallel to the array and provides a perfect solution for the case of a continuous distribution of secondary sources of infinite extent. A detailed treatment of the theoretic possibilities in this context is beyond the scope of this pa-

Ahrens, J.; Thomas, M.R.P.; Tashev, I., "Gentle acoustic crosstalk cancelation using the spectral division method and Ambiophonics," Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on, 20-23 Oct. 2013, p. 1 - 4, DOI: 10.1109/WASPAA.2013.6701827

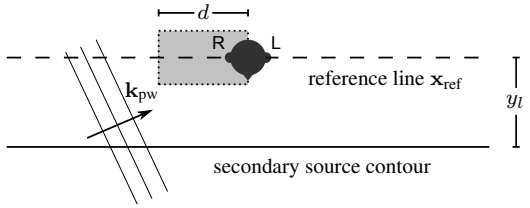


Figure 1: Conceptual illustration of the prescribed sound field: A plane wave with propagation direction $\mathbf{k}_{\text{pw}}$ illuminates only a section of width d of the reference line is created. The dashed line indicates the reference line and the gray area indicates the illuminated part of the reference line. L and R denote the listener's left and right ear respectively.

per. We therefore outline the theory only briefly and then concentrate on a specific sample loudspeaker array. We assume that the secondary source distribution is located along the x-axis, that the listener's ears are located at a known location inside the horizontal plane, and that the listener looks perpendicularly at the secondary source distribution.

The sound pressure $S(\mathbf{x}, \omega)$ evoked by such a continuous secondary source distribution is given by an integration over the driving function $D(\mathbf{x}_0, \omega)$ of the secondary source that is located at $\mathbf{x}_0 = [x_0\, 0\, 0]^T$ and its spatio-temporal transfer function $G(\mathbf{x}, \mathbf{x}_0, \omega)$. The integration is performed along the entire secondary source contour as

$$S(\mathbf{x}, \omega) = \int_{-\infty}^{\infty} D(\mathbf{x}_0, \omega) G(\mathbf{x} - \mathbf{x}_0, \omega)\, \mathrm{d}x_0\ . \tag{1}$$

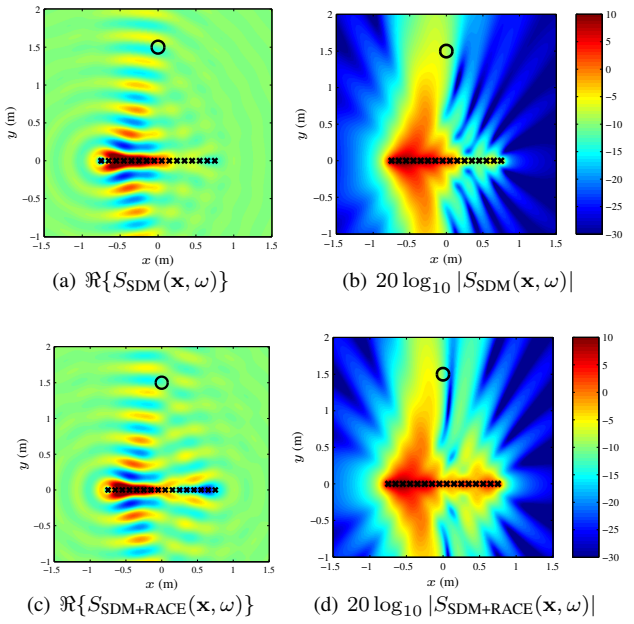
The driving function $\tilde{D}(k_x, \omega)$ for the synthesis of a desired sound field $\tilde{S}(k_x, y = y_l, z = 0, \omega)$ on the reference line $\mathbf{x}_{\text{ref}} = [x\, y_l\, 0]^T$ can then be determined in wavenumber domain as [12]

$$\tilde{D}(k_x, \omega) = \frac{\tilde{S}(k_x, y = y_l, z = 0, \omega)}{\tilde{G}(k_x, y = y_l, z = 0, \mathbf{x}_0 = [0\,0\,0]^T, \omega)}\ , \tag{2}$$

which can then be transferred to time-frequency domain or time domain via numeric Fourier transforms. $\tilde{G}(\cdot)$ in (2) may not exhibit zeros, which is fulfilled for omnidirectional secondary sources. The employment of discrete loudspeakers in practice constitutes a spatial sampling of the continuous secondary source distribution. The consequences of this spatial discretization have been treated extensively in the literature [13]. We discuss the aspects that are relevant in the current context in Sec. 4. Further details on the implementation of SDM are discussed in [14].

We choose to illuminate the listener's ipsilateral ear with a plane wave. Other virtual sound fields may also be useful. In order to exploit natural shadowing due to the listener's head we limit the extent of the illuminated part of the reference as indicated in Fig. 1: We choose the illuminated part to be of width d and center the listener's head at one of its boundaries. All other parts of the reference line are chosen to be quiet. $\tilde{S}(\cdot)$ is (2) can be obtained via a numerical Fourier transform from $S(\cdot)$, which is described analytically as outlined above.

Under ideal assumptions, a sharp transition between the illuminated part and the quiet parts of the reference line can indeed be achieved. However, this requires substantial evanescent sound field components, which have to be avoided in practice as transducer mismatch can render the result unusable. We therefore trigger only the propagating components of the desired sound field by setting $\tilde{D}(k_x, \omega) = 0 \forall\, |k_x| > \left|\frac{\omega}{c}\right|$ in (2) [13]. The absence of evanescent



(a) $\Re\{S_{\text{SDM}}(\mathbf{x}, \omega)\}$ (b) $20 \log_{10} |S_{\text{SDM}}(\mathbf{x}, \omega)|$

(c) $\Re\{S_{\text{SDM+RACE}}(\mathbf{x}, \omega)\}$ (d) $20 \log_{10} |S_{\text{SDM+RACE}}(\mathbf{x}, \omega)|$

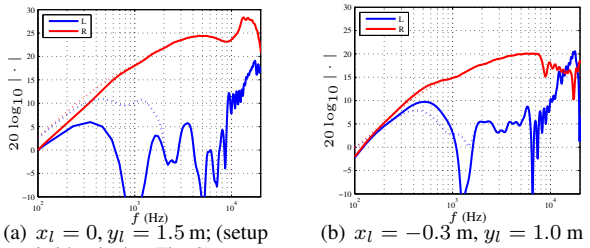
Figure 2: Synthesized monochromatic sound fields for $f = 1000$ Hz and listening position $x_l = 0$, $y_l = 1.5$ m. The marks indicate the positions of the loudspeakers; the circle indicates the assumed position of the listener's head. The scattering of the latter is not considered in the simulation. Fig. 2(a) and (c) show the real part of the synthesized sound field, (b) and (d) show the magnitude on a logarithmic scale.

sound field components as well as the sampling and truncation of the secondary source distribution will smear the energy and the intended quiet parts of the reference line will not be perfectly quiet. At this stage, it would be reasonable to chose $d = \infty$. However, as will be shown, it is important to chose d to be relatively small. We set $d = 0.5$ m for convenience and assume an array of 16 equally spaced omni-directional loudspeakers that are arranged symmetrically with respect to the y-z-plane.

Fig. 2(a) and (b) show the synthesized sound field for illumination of the right ear for $f = 1000$ Hz at listening position $x_l = 0$, $y_l = 1.5$ m. The azimuth of the plane wave propagation direction is $\theta_{\text{pw}} = 76.0°$. Note that all sound field components that are apparent in Fig. 2 other than the beam that illuminates the ipsilateral ear are artifacts mostly due to spatial truncation of the secondary source distribution.

The resulting channel separation at the 'ears' of a rigid spherical head for the chosen parameters is indicated by the dotted lines in Fig. 3(a) again for illumination of the right ear. The channel separation is in the order of 20 dB between 2000 Hz and 9000 Hz and lower elsewhere.

The spatial sampling of the secondary source distribution becomes significant at higher frequencies. Simulations show that propagating spatial aliasing artifacts of significant amplitude arise above $f_{\text{a}} \approx 2200$ Hz. These artifacts are copies of the desired field that propagate in different directions. The higher the considered frequency is the more copies arise and the more these copies propagate 'away' from the array and therefore get closer to the listener. Refer to Fig. 4 for examples. For the considered parameters and loudspeaker spacing, it is above approximately $f_{\text{limit}} = 9000$ Hz



(a) $x_l = 0$, $y_l = 1.5$ m; (setup is identical to Fig. 2) (b) $x_l = -0.3$ m, $y_l = 1.0$ m

Figure 3: Magnitude of the transfer function from the loudspeaker array to the surface of a rigid sphere of similar radius like a human head ($r = 8.5$ cm); R refers to the sound pressure at the location that its equivalent to the location of the right ear; L refers to the pressure at the left ear; dotted lines: SDM only; solid lines: SDM + RACE

that the aliasing artifacts that carry considerable energy illuminate the contralateral ear and therefore reduce the channel separation. This circumstance is also apparent in Fig. 3(a), where the channel separation shows a significant step around this frequency. Choosing a small loudspeaker spacing or smaller width d of the illuminated area will further increase f_{limit}.

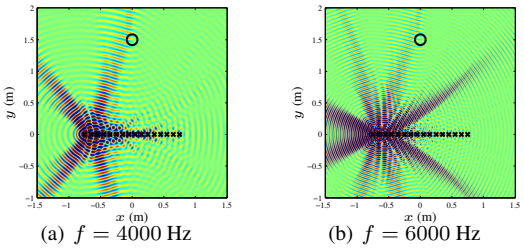
We will not attempt to further increase the channel separation above 9000 Hz – or reduce the crosstalk, respectively – because the considered wave lengths are very short and the system therefore becomes vulnerable to changes in the speed of sound, loudspeaker misplacement, and the like. Additionally, most signals like speech or music exhibit very little energy in this frequency range so that crosstalk may be assumed to be tolerable there.

The crosstalk below 2000 Hz, on the other hand, cannot be tolerated. The circumstance that the crosstalk is higher here is due to the fact that pure constructive interference does not allow for creating a sharp transition between illuminated and quiet areas and additionally, more diffraction around the listener's head occurs. We therefore have to apply destructive interference here as outlined in Sec. 3. We choose to employ Ambiophonics because if its tendency to collapse gently [11].

3. AMBIOPHONICS

The heart of Ambiophonics is RACE [10], which is a heuristic approach to crosstalk cancelation, and is designed for symmetric two-loudspeaker setups. It consists essentially in canceling the crosstalk at a given contralateral ear by a delayed and attenuated copy of the signal that caused the crosstalk. The cancelation signal has opposite sign and is emitted by the considered ear's ipsilateral loudspeaker. The delay Δt with respect to the original signal accounts for the longer path to the ear under consideration and the attenuation Δa accounts for head shadowing. RACE is typically applied in the frequency range between 250 Hz and 5000 Hz. Remarkably, frequency independent delay and attenuation seem to be sufficiently accurate, which makes the implementation of RACE straightforward.

There occurs of course also crosstalk with respect to the cancelation signal so that an according cancelation of the cancelation signal has be performed and so forth. Each recursion is attenuated by a few dB so that the cancelation signals become inaudible after a handful of recursions. RACE can indeed achieve impressive results [11] but it requires a carefully tuned setup of carefully chosen



(a) $f = 4000$ Hz (b) $f = 6000$ Hz

Figure 4: Synthesized sound field with aliasing artifacts apparent; the color scale was set different from Fig. 2(a) and (c) for convenience;

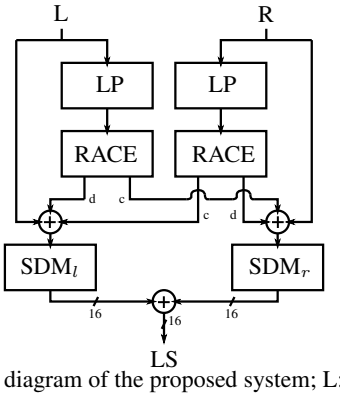


Figure 5: Block diagram of the proposed system; L: left channel of input signal; R: right channel of input signal; LP: low-pass filter; d: RACE direct path; c; RACE crosstalk path; SDM$_i$: SDM that creates the sound field that illuminates the indexed ear; LS: loudspeaker array

hardware and it requires the listener to be located on the symmetry plane between the two loudspeakers. The proposed employment of a loudspeaker array partly remedies these challenges.

We are using RACE only below 2000 Hz where we want to increase the channel separation whereby the cancelation signal is obtained from the input via an equiripple linear-phase low-pass filter. The cancelation signal is then transmitted by a plane wave the illumination area as well as the propagation direction of which are mirrored with respect to the initial plane wave. A block diagram of the resulting system is depicted in Fig. 5.

4. RESULTS

As can be seen from Fig. 2(c) and (d), applying RACE in the present context creates a corridor of very low amplitude. Appropriate choice of the delay and attenuation will make the location of this corridor include contain the location of the contralateral ear so that cancelation/reduction is achieved. We set the heuristic values of $\Delta t = 104\,\mu$s and $\Delta a = -8$ dB in this example. The low-amplitude corridor expands approximately perpendicularly to the loudspeaker array. This suggests that a certain amount inaccuracy in the estimation of the listener's distance by the tracking system is tolerable. This finding applies to other approaches as well [3, 4, 6, 7].

The resulting channel separation can be deduced from the solid lines in Fig. 3(a). The chosen parameters cause a slight reduction of the amplitude at the ipsilateral ear but also increase the channel separation by more than 15 dB at certain frequencies. The transfer function above 2000 Hz is unaffected. The channel separation is comparable to what is achieved in [8, 9] but lower than for ag-

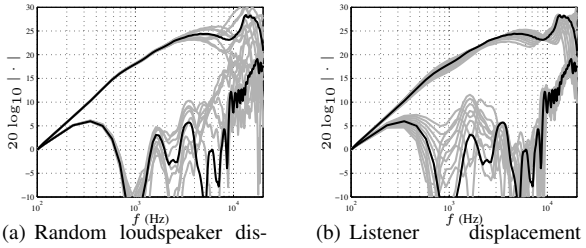
gressive methods, e.g. [7]. The perceptual localization experiments presented in [3, 6] show that this amount of channel separation is not sufficient in order to achieve comparable performance like headphone presentation. However, informal listening suggests that the achieved crosstalk reduction is indeed sufficient to achieve full lateralization.

Using symmetric sound fields for the recursive cancelation seems to be preferable from a perceptual point of view although this does not exploit the maximum possible natural head shadowing for off-center listening positions. One sample result for an off-center listening position is depicted in Fig. 3(b). The listening position $x_l = -0.3$ m, $y_l = 1.0$ m is shown and the RACE parameters were set heuristically to $\Delta t = 521$ μs and $\Delta a = -14$ dB. The azimuth of the propagation direction the illuminating the plane wave is $\theta_{\text{pw}} = 77.3^\circ$. Illuminated areas were shifted in direction of the ipsilateral ear by 0.05 m compared to the illustration in Fig. 1.

The result from Fig. 3(b) suggests that the channel separation drops significantly compared to the central listening position when the listener is located close to one of the boundaries of the loudspeaker array. The reason for this is a significant reduction in the natural head shadowing that can be evoked.

Fig. 6 illustrates the robustness of the presented approach with respect to random loudspeaker displacement (Fig. 6(a)) as well as listener displacement (Fig. 6(b)) from the locations assumed by the driving signals. The loudspeakers in Fig. 6(a) are displaced randomly with respect to both the x and y axes by values obtained from a normal distribution with mean 0 and standard deviation $\sigma = 1$ cm. These parameters are rather generous as loudspeaker arrays can be manufactured with millimeter precision. Fig. 6(a) suggests that according loudspeaker displacements do not reduce the channel separation by a substantial amount in the range below 4000 Hz.

Fig. 6(b) depicts the ear signals for displacement of the listener along the x-axis in the interval $[-5 \text{ cm}, 5 \text{ cm}]$. Such displacements occur due to inaccuracies or latency of the employed listener tracking system. As expected, the channel separation at low frequencies is rather vulnerable to listener displacement and the channel separation can drop below 10 dB. This suggests that a high-accuracy and low-latency listener-tracking system should be employed. This seems to be an inevitable requirement due to the relatively short distance between the ears of around 15 cm. Remarkably, it is primarily the contralateral side below 3000 Hz that is affected by this type of displacement, which suggests that only a moderate perceptual impairment occurs.



(a) Random loudspeaker displacement

(b) Listener displacement along the x-axis in steps of 1 cm

Figure 6: Equivalent to Fig. 3(a) but for random loudspeaker displacement (Fig. 6(a)) and for listener displacement (Fig. 6(b)). The setup is identical to Fig. 2. The black line indicates the result without displacement, i.e., the solid line from Fig. 3(a)

The overall energy radiated by a system driven with the presented approach is significantly lower than for some of the aggressive cancelers, e.g. [6], which is desirable as it avoids excessive reverberation from the listening room. Comparison of the robustness of the presented approach with the literature is difficult because data on the robustness are either not available or reported in incompatible ways. The available data in [3, 4, 6, 7] suggest that different systems tend to behave differently with respect to listener displacement so that an ultimate judgement requires a perceptual comparison.

5. CONCLUSIONS

We presented the concept of gentle crosstalk cancelation using a linear loudspeaker array. The approach uses purely constructive synthesis above approximately 2000 Hz where the natural head shadowing provides sufficient channel separation. Below 2000 Hz we additionally apply Recursive Ambiophonics Crosstalk Elimination. The achievable channel separation is somewhat lower than for aggressive methods and is dependent on the listener position and is significantly lower for listener positions close to the ends of the loudspeaker array. A reliable comparison with respect to the robustness can only be performed based on a comparative perceptual study.

Future work includes considering the actual loudspeaker directivity in the approach as demonstrated in [13] as well as rotations of the listener's head.

6. REFERENCES

[1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1996.

[2] B. B. Bauer, "Stereophonic earphones and binaural loudspeakers," *JAES*, vol. 9, no. 2, pp. 148–151, Apr. 1961.

[3] W. G. Gardner, "3-D Audio Using Loudspeakers," Ph.D. thesis, Massachusetts Institute of Technology, 1997.

[4] T. Takeuchi, P. A. Nelson, and H. Hamada, "Robustness to head misalignment of virtual imaging systems," *JASA*, vol. 109, no. 3, pp. 958 – 971, Mar. 2001.

[5] Bauck and Cooper, "Generalized transaural stereo and applications," *JAES*, vol. 44, no. 9, pp. 683 – 705, Sep. 1996.

[6] M. R. Bai, C.-W. Tung, and C.-C. Lee, "Optimal design of loudspeaker arrays for robust cross-talk canellation using the Taguchi method and the genetic algorithm," *JASA*, vol. 117, no. 5, pp. 2802 – 2813, May 2005.

[7] Y. Lacouture Parodi and P. Rubak, "Objective evaluation of the sweet spot size in spatial sound reproduction using elevated loudspeakers," *JASA*, vol. 128, no. 3, pp. 1045 – 1055, Sep. 2010.

[8] D. Menzel, H. Wittek, G. Theile, and H. Fast, "The Binaural Sky: A Virtual Headphone for Binaural Room Synthesis," in *Tonmeistersymposium*, Hohenkammer, Germany, 2005.

[9] M. Guldenschuh and A. Sontacchi, "Transaural Stereo in a Beamforming Approach," in *Proc. DAFx-09*, Como, Italy, 2009, pp. 1–6.

[10] R. Glasgal, "360° localization via 4.x RACE processing," in *123rd Convention of the AES*, 2007, p. 7301.

[11] ——, "Ambiophonics," Demonstration given at the Ambiophonics Institute, Rockleigh, NY, USA, 2007.

[12] J. Ahrens and S. Spors, "Sound field reproduction using planar and linear arrays of loudspeakers," *IEEE Trans. on Audio, Sp., and Lang. Proc.*, vol. 18, no. 8, pp. 2038 – 2050, Nov. 2010.

[13] J. Ahrens, *Analytic Methods of Sound Field Synthesis*. Berlin/Heidelberg: Springer, 2012.

[14] J. Ahrens, M. R. P. Thomas, and I. Tashev, "Efficient Implementation of the Spectral Division Method for Arbitrary Virtual Sound Fields," in *IEEE WASPAA*, New Paltz, NY, Oct. 2013.