Audio Engineering Society

Convention Paper 9865

Presented at the 143rd Convention
2017 October 18–21, New York, NY, USA

Efficient Structures for Virtual Multi-Channel Immersive Audio Rendering

Jean-Marc Jot[1] and Daekyoung Noh[2]

[1]*Magic Leap, Inc. - Sunnyvale, CA, USA*
[2]*Xperi Corp. - Santa Ana, CA, USA*

Correspondence should be addressed to Daekyoung Noh (sid.noh@xperi.com)

ABSTRACT

New consumer audio formats have been developed in recent years for the production and distribution of immersive multichannel audio recordings including surround and height channels. HRTF-based binaural synthesis and cross-talk cancellation techniques can simulate virtual loudspeakers, localized in the horizontal plane or at elevated apparent positions, for audio reproduction over headphones or conventional loudspeaker playback systems. In this paper, we review and discuss some practical design and implementation challenges of immersive audio virtualization methods, and describe computationally efficient processing approaches and topologies enabling more robust and consistent reproduction of directional audio cues in consumer applications.

1 Introduction

This paper is concerned with the virtual 3D audio reproduction of discrete multi-channel material, where each channel represents a loudspeaker-feed audio signal, as in the legacy 5.1 and 7.1 surround sound formats. Efficient binaural and transaural 3D audio virtualization and cross-talk cancellation algorithms have been proposed during recent decades, including sum and difference "shuffler"-based topologies taking advantage of properties such as the left-right symmetry of channel layouts, minimum-phase models of the head-related transfer functions (HRTFs) and spectral equalization methods, as well as digital IIR filter approximations [1] [2] [3] [4]. As illustration, Figure 1 depicts a simple 4-channel virtual 3D audio reproduction system.

It is straightforward to extend these techniques for application to recently developed immersive audio formats including additional height channels (e.g. [5] [6] [7] [8]), so that virtual elevation effects may be readily reproduced jointly with horizontal-plane lateralization and surround localization effects, at the cost of a linear complexity increase with the number of additional channels. This approach offers the attractive perspective of height-channel reproduction without requiring physical height, ceiling or upward-firing loudspeakers, achieving compatibility with legacy stereo or surround-sound playback systems.

Persistent practical challenges in the development of 3D audio virtualization technologies include: delivering robust subjective localization performance over

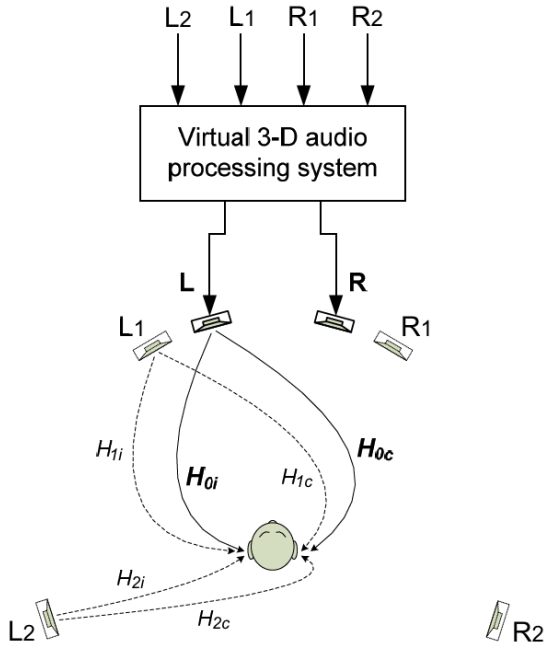

Fig. 1: 4-channel virtualization processing system.

headphones despite individual morphology differences [9] or while allowing flexible listener position in loudspeaker playback [4], and convincingly rendering the due interpolated localization of "phantom sources" panned across two or more input channels [10].

In order to alleviate horizontal localization and timbre reproduction artifacts for listener positions away from the central location (or "sweet spot") in a loudspeaker-based virtual 3-D audio system, some proposed solutions involve neutralizing the virtualization processing above the medium frequency range [4] [11]. For immersive reproduction with height, this approach must be implemented in such a manner as to preserve elevation effects, which are known to depend significantly on high-frequency features of the head-related transfer functions [12].

In the next section of this paper, we review general filter design principles for 3-D audio reproduction of left-right symmetric virtual loudspeaker pairs. We then extend these methods to develop efficient and practical designs for virtual elevation effects in multi-channel audio systems. Lastly, we include an enhancement employing signal decorrelation processing to improve the reproduction of phantom sources.

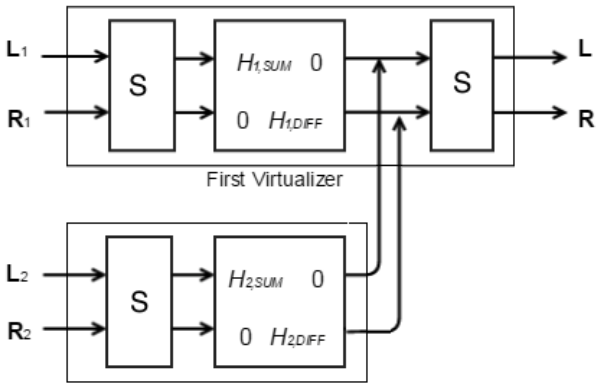

Fig. 2: 4-channel pairwise virtualizer.

2 Pairwise multi-channel 3D audio virtualization

Figure 2 illustrates a realization of the virtualization processing system of Figure 1, reproducing two virtual loudspeaker pairs (L_1, R_1) and (L_2, R_2) for playback over headphones or a physical loudspeaker pair (L, R). Each input pair is processed by an elementary transaural virtualizer assuming that the virtual and physical loudspeaker pairs are symmetrically arranged relative to the median plane. With this assumption, the virtualized output signal for the pair (L_1, R_1) is given by [1] [2] [11]:

$$\begin{bmatrix} L \\ R \end{bmatrix} = \mathbf{S} \cdot \begin{bmatrix} H_{1,SUM} & 0 \\ 0 & H_{1,DIFF} \end{bmatrix} \cdot \mathbf{S} \cdot \begin{bmatrix} L_1 \\ R_1 \end{bmatrix}$$

where the sum/difference "shuffler" matrix $\mathbf{S}$, the sum filter, and the difference filter are defined by:

$$\mathbf{S} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

$$H_{1,SUM} = \{H_{1i} + H_{1c}\} \cdot \{H_{0i} + H_{0c}\}^{-1}$$

$$H_{1,DIFF} = \{H_{1i} - H_{1c}\} \cdot \{H_{0i} - H_{0c}\}^{-1}$$

with the ipsilateral or contralateral HRTF notations of Figure 1. In this paper, $\{H(z)\}$ denotes the minimum-phase transfer function having the same magnitude frequency response as $H(z)$. The dependence on the complex frequency variable z is omitted for simplification. The virtualizer for the input pair (L_2, R_2) is realized similarly, and the approach may be extended to additional input pairs in a multi-channel audio virtualization system.

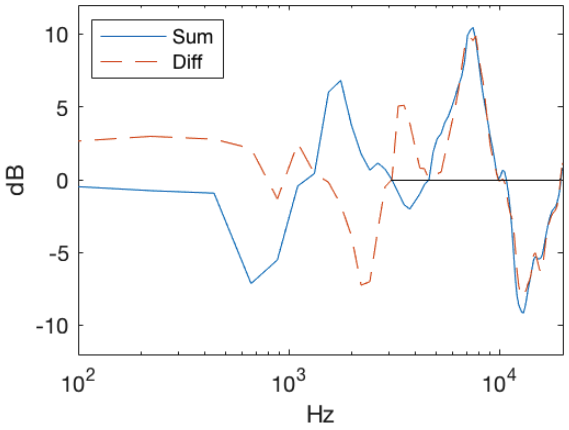

Fig. 3: Sum and difference filter responses for elevated virtual loudspeaker pair at azimuth ±30 degrees and elevation +45 degrees, with physical loudspeaker pair at azimuth ±15 degrees in the horizontal plane.

For a virtual loudspeaker situated in the median plane, the ipsilateral and contralateral HRTFs are assumed identical and the difference filter vanishes. In the case of headphone reproduction, H_{0c} is substantially zero and H_{0i} denotes the headphone-to-ear transfer function.

Figure 3 displays the frequency responses of the sum and difference filters $H_{2,SUM}$ and $H_{2,DIFF}$ for the virtualization of a height channel pair (L_2, R_2). In this example, HRTF measurements from the CIPIC database [13] were used, and an average over 45 subjects was calculated for each of the sum and difference magnitude spectra $|H_{2i}+H_{2c}|$, $|H_{2i}-H_{2c}|$, $|H_{0i}+H_{0c}|$ and $|H_{0i}-H_{0c}|$, prior to spectral division.

In order to reduce subjective localization artifacts when the listener's head shifts or rotates away from the reference listening position and look direction, band-limiting may be realized by setting the sum and difference filter magnitude responses to unity above mid frequencies (as suggested in [11] for horizontal virtualizer designs). However, as indicated in figure 3, that modification would result in the elimination of salient elevation cues, including a boost in the 8-kHz region. In the following, alternative methods are discussed.

3 Virtual Elevation Filter

In several previous studies, a virtual elevation filter is derived from a HRTF magnitude or power ratio evaluated for a sound source located in the median plane,

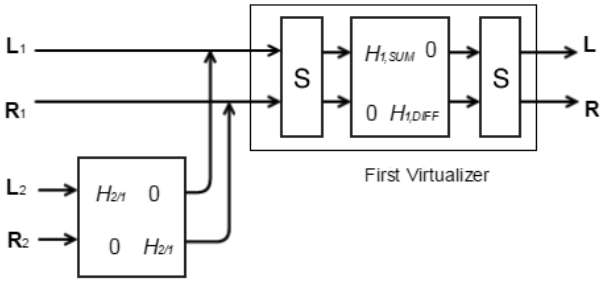

Fig. 4: 4-channel pairwise virtualizer using virtual elevation pre-filtering of the height input pair.

in front of the listener [14] [15] [16] [17]. This approach corresponds to the simplified virtualizer topology shown in Figure 4, where the height channel pair (L_2, R_2) is pre-filtered with a virtual elevation filter $H_{2/1}$ before processing by the horizontal virtualizer shared with input channel pair (L_1, R_1). In this case, the frequency response of the virtual height filter is given by the HRTF magnitude ratio:

$$|H_{2/1}| = |H_{2i}|/|H_{1i}| = |H_{2c}|/|H_{1c}|$$

Figure 5 illustrates the variation of the ipsilateral virtual elevation filter across subjects in the CIPIC database, for a fixed target elevation, at azimuth 30 degrees. Also displayed is the frequency response derived by averaging $|H_{2i}|/|H_{1i}|$ across subjects in the CIPIC database. As reported in previous studies, the high-frequency features, which are affected by pinna shape, vary significantly across subjects [18]. Additionally, it is visible that a rear virtual elevation filter, derived similarly for azimuth 150 degrees, has a different response, notably lacking a high-frequency boost in the 8 kHz region.

Figure 6 illustrates the variation of the averaged elevation filter magnitude frequency response with increasing target elevation angle, at azimuth 30 degrees, for the ipsilateral or contralateral ear. It is observed that elevation cues generally increase monotonically with elevation angle.

4 Virtualizer Design Improvements

The virtualizer topology of figure 4 is applicable if the ipsilateral and contralateral virtual elevation filters can be assumed to be identical. This approximation is justified for median-plane locations (azimuth 0) and, as seen in figure 6, acceptable for azimuth 30 degrees.

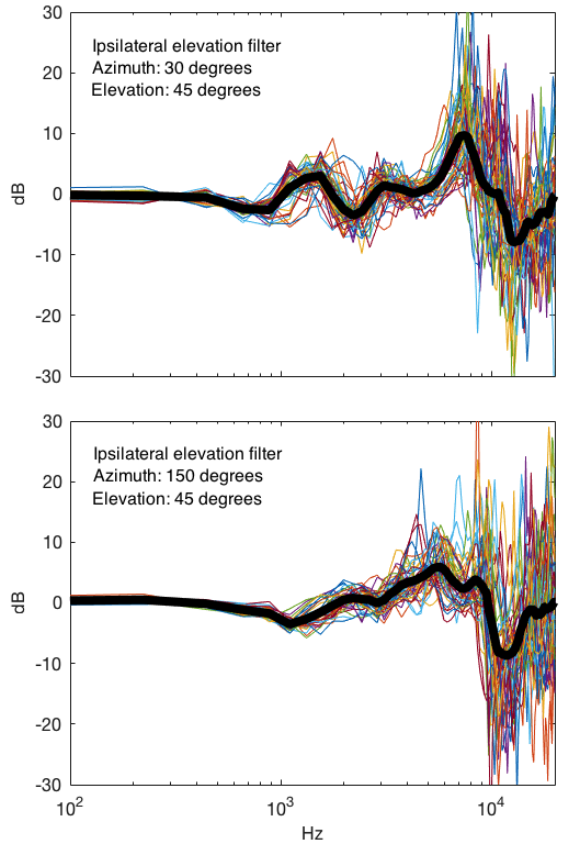

Fig. 5: Variation and average of front or rear virtual elevation filter response across subjects.

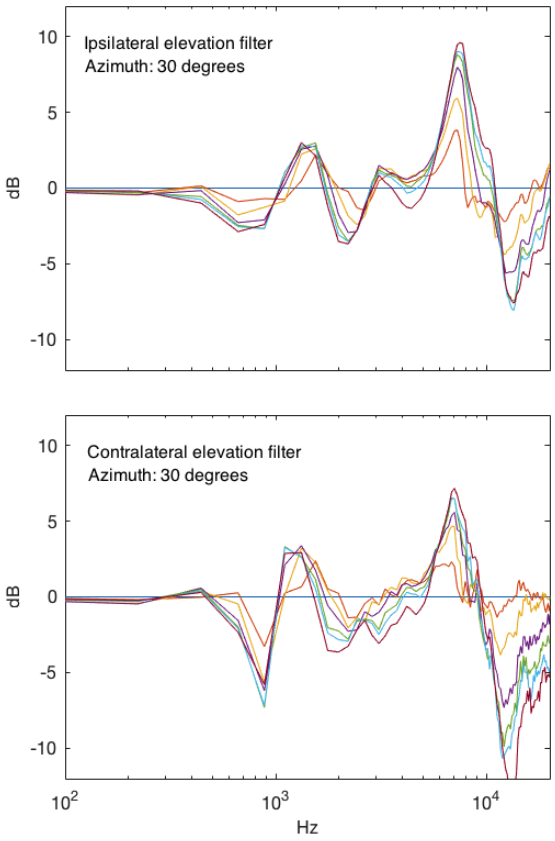

Fig. 6: Variation of virtual elevation filter response with target elevation angle (0.0, 11.2, 22.5, 33.7, 45.0, 50.6, and 61.8 degrees).

However, as shown in figure 7, it becomes less valid for wider azimuth angles. Furthermore, the frequency response of the ipsilateral virtual elevation filter varies significantly with azimuth angle.

Figure 8 depicts a general dual-stage cascaded virtualizer topology, which allows accounting for the difference between ipsilateral and contralateral elevation cues and making the virtual elevation filter dependent on the due azimuth of the virtual sound source relative to the listener's look direction. The second virtualizer, applied to the input pair (L_2, R_2), includes sum and difference pre-filters defined by:

$$H_{2/1,SUM} = \{H_{2i} + H_{2c}\} \cdot \{H_{1i} + H_{1c}\}^{-1}$$

$$H_{2/1,DIFF} = \{H_{2i} - H_{2c}\} \cdot \{H_{1i} - H_{1c}\}^{-1}$$

with notations introduced in section 2 and Figure 1.

Figure 9 displays the sum and difference filter responses in this 2-stage virtualizer, for rendering the height channel pair (L_2, R_2) with the same virtual and physical loudspeaker positions as in the example of figure 3. In this case, the first virtualizer, which is also applied to the horizontal channel pair (L_1, R_1), is band-limited at 3 kHz. The height channels are pre-filtered by the second virtualizer, resulting in combined sum and difference filter frequency responses that compare to those of a single-stage virtualizer as displayed in Figure 3.

The staged virtualizer topology of Figure 8 is computationally advantageous compared to the parallel topology of Figure 2 because it allows sharing horizontal-plane virtualization processing between the two input channel pairs, (L_1, R_1) and (L_2, R_2). Furthermore, the separation of virtual elevation processing from

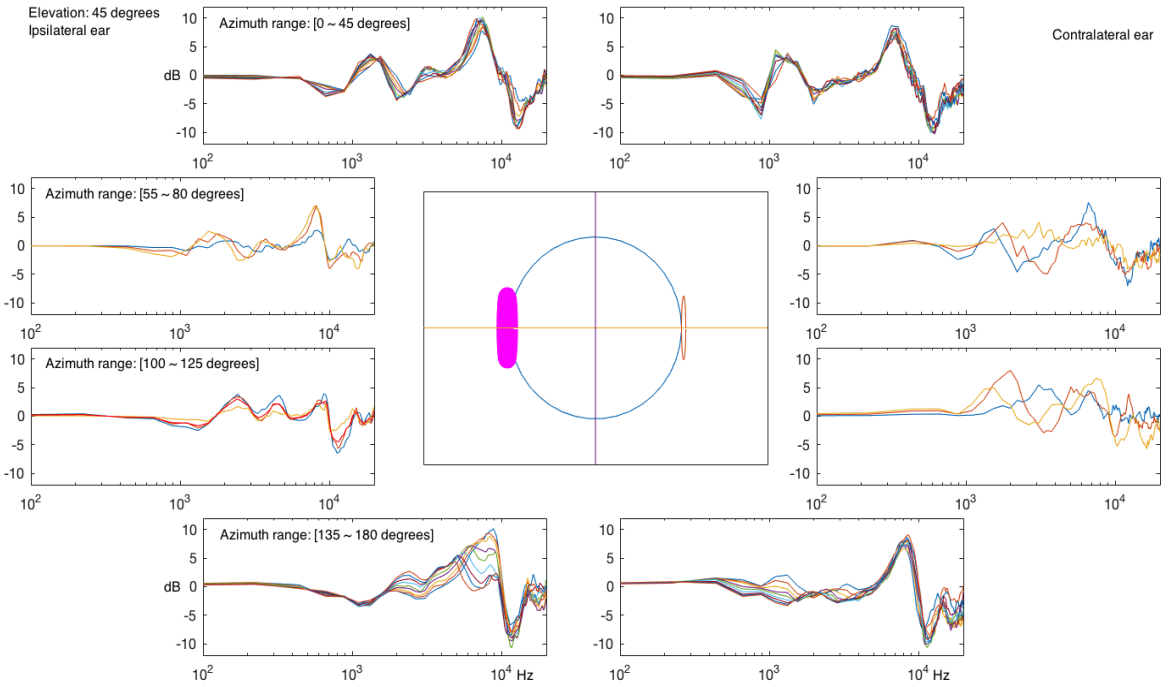

Fig. 7: Variation of the elevation filter across azimuth angles, after averaging across 45 subjects form the CIPIC HRTF database. The range of azimuth angles spanning 0 to 180 degrees is split into four angular sectors.

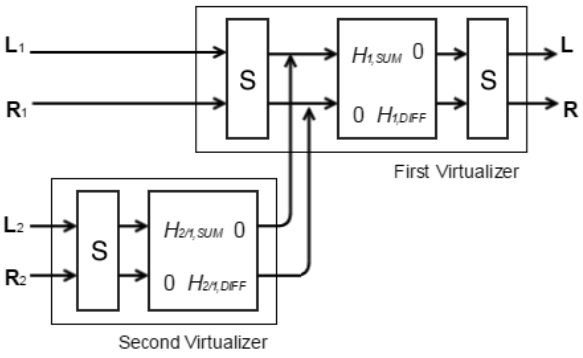

Fig. 8: 4-channel pairwise virtualizer using shared filtering stage, equivalent to the virtualizer of figure 2.

horizontal-plane virtualization processing allows optimizing the reproduction of these two localization dimensions separately: tuning the virtual elevation effect and preserving it even at listening positions away from the “sweet spot”, irrespective of horizontal localization effect design trade-offs.

5 Virtualization of Phantom Sources

Multichannel audio signals commonly include sound components that are "panned" across two or more audio channels, in order to reproduce sound localizations that do not coincide with loudspeaker positions. Such panned sounds are often referred to as "phantom sources". When the input channels are reproduced via virtual loudspeakers, we may refer to the perceived result as a "virtual phantom Source". Even when virtual loudspeaker processing faithfully reproduces the due localization of each input channel signal auditioned individually, it is observed that virtual phantom sources often incur audible defects in loudness or timbre preservation or in localization accuracy (including elevation errors, front-back confusion, or in-head localization).

A particular case of interest is the reproduction of a virtual phantom height center channel by feeding an identical signal to the height input pair (L_2, R_2). While the due sound localization is, in this case, the mid point between the two elevated virtual loudspeaker positions, it is observed that an audio signal reproduced in this manner typically sounds less elevated than the same signal feeding a single virtual elevated loudspeaker.

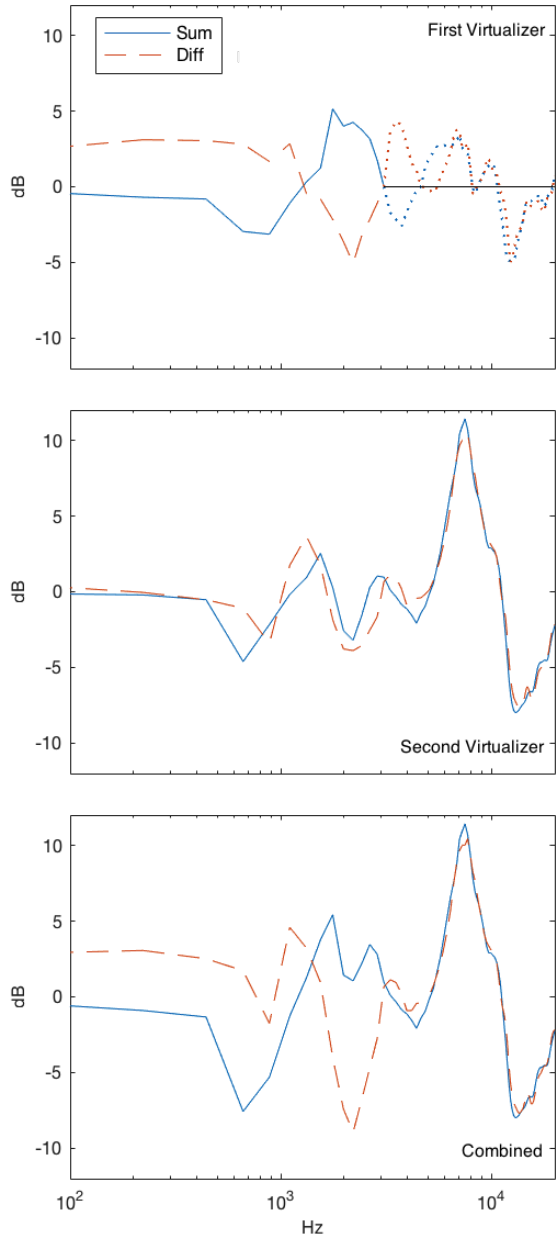

Fig. 9: Sum and difference filter responses for the same virtual and physical loudspeaker positions as in figure 3, using the virtualizer topology of figure 8 with band-limited horizontal virtualizer.

Improved techniques for the virtualization of phantom sources were proposed in [10] and [19], where frequency-domain spatial analysis/synthesis is used for selectively "re-rendering" them at their due localization. In [20] and [21], additional decorrelation processing

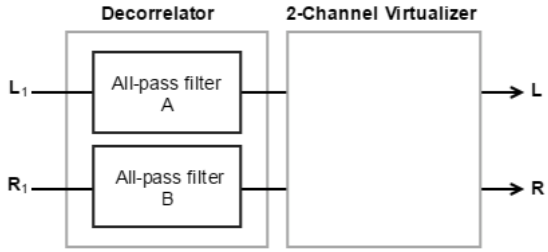

Fig. 10: Decorrelation pre-processing for improving the virtualization of phantom sources.

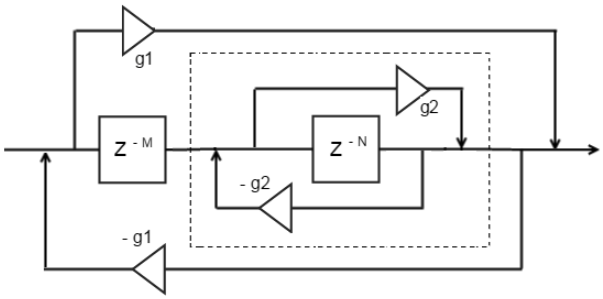

Fig. 11: Nested all-pass filter topology.

is proposed for improving the reproduction of phantom center components. We propose an a solution which, unlike these previous methods, does not involve frequency-domain processing of the input signals.

In figure 10, decorrelation filters are applied to the input signals prior to virtualization processing. The intent of this decorrelation is to ensure that source signals panned between the input channels will be heard by the listener at virtual positions substantially located on the shortest arc segment centered on the listener's position and joining the due positions of the corresponding virtual loudspeakers. As anticipated by previous studies, this effect is accompanied by a broadening (or "smearing") of the perceived sound image [20] [22], which we deem preferable, in the present case, to ineffective virtual elevation processing.

An example realization for the decorrelation filters is the nested all-pass filter topology shown in Figure 11 [23]. A bank of two or more inter-channel decorrelators may be realized by choosing, for the parameters M, N, g_1 and g_2 of each nested all-pass filter, different settings between channels. Variants and improvements addressing timbre and spectral balance preservation with time-domain all-pass filter designs suited for this application are described e.g. in [20].

6 Summary and Discussion

We have described methods for virtual 3-D audio rendering of multi-channel audio material, specifically targeting applications where the listener's look direction is determined (for instance by position sensors or because the listener is assumed to be facing a fixed direction). Such applications include home-theater sound bar loudspeakers or surround-sound systems, and headphone-based immersive virtual reality audio systems that incorporate head-tracking technology.

In these applications, it is not restrictive to assume that virtual loudspeaker locations are fixed in reference to the listener's head, and either in the median plane or arranged in left-right symmetric pairs — which allows the use of efficient parallel pairwise virtualizer topologies and filter designs. Here, we build upon prior studies to propose new topologies employing cascaded virtualization stages, and demonstrate this approach in the case of virtual elevation processing for immersive audio reproduction over conventional stereo loudspeaker playback systems.

Separating virtual elevation from horizontal surround-sound virtualization effects offers computational efficiency and subjective performance advantages, by (1) the possibility of sharing virtualization processing stages between several input pairs, and (2) facilitating subjective performance optimization by allowing independent virtualizer design trade-offs between horizontal and vertical localization attributes (lateralization, elevation and front-back discrimination). Optionally, the proposed solution also includes low-complexity inter-channel decorrelation pre-processing that can improve the subjective localization performance for phantom sources.

7 Acknowledgments

We would like to thank Edward Stein for contributing filter design tools used in this study, Oveal Walker for collaboration in listening evaluation experiments, Ryan Cassidy, Rick Oliver and Themis Katsianos for fruitful discussions and feedback during this research.

This work was carried out while the first author was employed by DTS, Inc. (now a subsidiary of Xperi Corp.).

References

[1] Cooper, D. H. and Bauck, J. L., "Prospects for transaural recording," *Journal of the Audio Engineering Society*, 37(1/2), pp. 3–19, 1989.

[2] Jot, J.-M., Larcher, V., and Warusfel, O., "Digital signal processing issues in the context of binaural and transaural stereophony," in *Proc. 98th Audio Engineering Society Convention*, 1995.

[3] Huopaniemi, J. and Karjalainen, M., "Review of digital filter design and implementation methods for 3-D sound," in *Proc. 102nd Audio Engineering Society Convention*, 1997.

[4] Gardner, W. G., *3-D audio using loudspeakers*, Springer Science & Business Media, 1998.

[5] Hamasaki, K., Hiyama, K., and Okumura, R., "The 22.2 multichannel sound system and its application," in *Proc. 118th Audio Engineering Society Convention*, 2005.

[6] Jot, J.-M. and Fejzo, Z., "Beyond surround sound—creation, coding and reproduction of 3-D audio soundtracks," in *Proc. 131st Audio Engineering Society Convention*, 2011.

[7] Robinson, C. Q., Mehta, S., and Tsingos, N., "Scalable format and tools to extend the possibilities of cinema audio," *SMPTE Motion Imaging Journal*, 121(8), pp. 63–69, 2012.

[8] Herre, J., Hilpert, J., Kuntz, A., and Plogsties, J., "MPEG-H audio—the new standard for universal spatial/3D audio coding," *Journal of the Audio Engineering Society*, 62(12), pp. 821–830, 2015.

[9] Wightman, F. L. and Kistler, D. J., "Individual differences in human sound localization behavior." *The Journal of the Acoustical Society of America*, 99(4), pp. 2470–2500, 1996.

[10] Breebaart, J. and Schuijers, E., "Phantom materialization: A novel method to enhance stereo audio reproduction on headphones," *IEEE transactions on audio, speech, and language processing*, 16(8), pp. 1503–1511, 2008.

[11] Walsh, M. and Jot, J.-M., "Loudspeaker-Based 3-D Audio System Design Using the MS Shuffler Matrix," in *Proc. 121st Audio Engineering Society Convention*, 2006.

[12] Han, H., "Measuring a dummy head in search of pinna cues," *Journal of the Audio Engineering Society*, 42(1/2), pp. 15–37, 1994.

[13] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C., "The CIPIC HRTF database," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 99–102, 2001.

[14] Kim, S., Ikeda, M., Takahashi, A., Ono, Y., and Martens, W. L., "Virtual ceiling speaker: elevating auditory imagery in a 5-channel reproduction," in *Proc. 127th Audio Engineering Society Convention*, 2009.

[15] López, J. J., Cobos, M., and Pueo, B., "Experiments on the Perception of Elevated Sources in Wave-Field Synthesis Using HRTF Cues," in *Proc. 128th Audio Engineering Society Convention*, 2010.

[16] Lee, K., Son, C., and Kim, D., "Immersive Virtual Sound for Beyond 5.1 Channel Audio," in *Proc. 128th Audio Engineering Society Convention*, 2010.

[17] Lee, Y. W., Kim, S., Jo, H., Park, Y., and Kim, J., "Virtual height speaker rendering for Samsung 10.2-channel vertical surround system," in *Proc 131th Audio Engineering Society Convention*, 2011.

[18] Algazi, V. R., Duda, R. O., and Satarzadeh, P., "Physical and filter pinna models based on anthropometry," in *Proc. 122nd Audio Engineering Society Convention*, 2007.

[19] Goodwin, M. M. and Jot, J.-M., "Binaural 3-D audio rendering based on spatial audio scene coding," in *Proc. 123rd Audio Engineering Society Convention*, 2007.

[20] Vickers, E., "Fixing the phantom center: diffusing acoustical crosstalk," in *Proc. 127th Audio Engineering Society Convention*, 2009.

[21] Jot, J.-M. and Walsh, M., "Center-Channel Processing in Virtual 3-D Audio Reproduction over Headphones or Loudspeakers," in *Proc. 128th Audio Engineering Society Convention*, 2010.

[22] Kendall, G. S., "The decorrelation of audio signals and its impact on spatial imagery," *Computer Music Journal*, 19(4), pp. 71–87, 1995.

[23] Gardner, W. G., "A realtime multichannel room simulator," *J. Acoust. Soc. Am*, 92(4), p. 2395, 1992.