Acoustical Engineering 2015

On the Management of Direct and Reflected Sounds in the 5.0 Surround Sound Reproduction System

P. Kleczkowski*, A. Król and P. Małecki
AGH University of Science and Technology, Department of Mechanics and Vibroacoustics,
Al. Mickiewicza 30, 30-050, Krakow, Poland

Direct and ambient sounds are usually reproduced through the same loudspeakers. In multichannel systems, with the use of anechoic recordings and auralization techniques it is possible to separate these sound components so that they are reproduced via different loudspeakers. Four basic options of management of direct and reflected sounds were investigated in a perceptual experiment. One sound source and the standard 5.0 surround sound system was used. It was found that listeners consistently preferred sound reproduction with direct sound radiated from only one loudspeaker.

DOI: 10.12693/APhysPolA.128.A-11
PACS: 43.38.Md, 43.60–c, 43.55.Lb, 43.66.Lj

1. Introduction

In contemporary audio production, original room acoustics are seldom captured during recording, hence indirect means of rendering spaciousness are the key issue. All methods are based on one common operation: the convolution of a dry sound with a room impulse response (IR), where the latter is obtained from a measurement or a calculation based on a physical model. In audio engineering practice, the aim is often the perception-based spaciousness. Two-channel IRs, i.e. measured with two locations of the sound source, compatible with the stereophonic sound reproduction system are often adequate. In the research, the purpose is replication of acoustic spaces, usually referred to as auralization [1, 2] or virtual acoustics [3, 4]. Those applications require multichannel reproduction systems and multichannel IRs.

Original sets of multichannel IRs of existing spaces are obtained from measurements with a variety of microphone systems, either spaced or coincident. In the first case, the procedure may provide plausible reproduction of acoustic ambience but is rather perception-based, i.e. can be seen as artificial reverberation. With coincident microphone setups, like the Ambisonic microphone [5], the effects are closer to accurate reproduction of a particular acoustic sound field. The sets of IRs obtained with coincident microphone setups are often referred to as the Spatial Impulse Responses (SIRs). Various methods of encoding a measured SIR in B-format (i.e. W, X, Y, and Z components) [5] for reproduction through systems with different numbers of loudspeakers have been developed, from simple [6] to advanced [7–11].

When the acoustics of a room are analyzed, its IRs are divided into three consecutive parts: the direct sound, early reflections, and late reflections (the reverberation tail). The parts have different physical and perceptual properties. In this work, perceptual properties of direct sound (DS) as opposed to all reflected sounds (RSs) are of interest, where the latter comprises both early and late reflections.

A straightforward approach to implement convolution in the rendering of acoustic spaces: $y(t) = s(t) * h(t)$, where $s(t)$ is the anechoic signal and $h(t)$ is the room IR, is to use the complete IR, i.e.:

$$h(t) = h_{\mathrm{dir}}(t) + h_{\mathrm{refl}}(t) = \mathrm{DS} + \mathrm{RSs}. \tag{1}$$

This is a natural consequence of measuring spatial IRs, where the DS is captured by all capsules or microphones, although with different amplitudes. A recent example of this approach in auralization can be found in [12].

Arguments can be put forward that using only RSs in convolution ($h(t) = h_{\mathrm{refl}}(t) = \mathrm{RSs}$) is a better choice for all channels delivering ambience. Indeed, there is only one direction from which the DS reaches the listener. If other directions are represented in a multichannel system, then only reflected sounds should arrive from them.

Another option can be considered: using only the DS part in convolution ($h(t) = h_{\mathrm{dir}}(t) = \mathrm{DS}$) in front channels of a multichannel system, i.e. removing RSs part from them. This option can be justified by the fact that reflections from the sides (especially from directions around $\pm 60\,^{\circ}\mathrm{C}$) are preferred over reflections from the front, as they produce lower interaural cross-correlation and hence increase the apparent source width and spatial impression [13–15]. Leaving out some channels just for the reproduction of DS should also reduce masking in those channels [16, 17] and intermodulation distortion resulting from combining DS and RSs in appropriate channels.

The combination of both above options leads to complete separation of DS from RSs in the reproduction system, i.e. some loudspeakers are dedicated only to reproduction of DS while the others reproduce only RSs.

*corresponding author; e-mail: kleczkow@agh.edu.pl

Elements of the above options have been applied both in experimental systems [6–8, 18, 19] and in commercially available convolution reverberation plug-ins, but no investigation concerning their perceptual effects has been made. The aim of the work reported here was to investigate perceptual effects of each of the above options, by comparison with the reference option consisting in performing full convolution according to (1) in all channels.

Such an investigation can be performed with many possible sets of independent variables, like the SIRs of different rooms, auditioning in different acoustic environments, using different multichannel systems, different numbers of sound sources, and different methods of evaluation. In this paper, an introductory experiment is presented, with one SIR, listening in an anechoic chamber, one sound source reproduced with the standard 5.0 surround reproduction system [20], and the pairwise comparison method of perceptual evaluation.

2. The method

2.1. Management of signals

The reference option, further referred to as option A, is presented in Fig. 1. The input signal to each channel is the result of the convolution $s(n) * h_X(n)$, where $s(n)$ is an anechoic recording of a sound source and $h_X(n)$ denotes complete IR according to (1), and X denotes a particular channel of the 5.0 system (designated by standard abbreviations L, C, R, LS and RS). For the sake of the clarity of drawings, in Figs. 1 through 4 the general symbols of sound components were used: d for DS and r for RSs. In fact, both d ($h_{\mathrm{dir}}(n)$) and r ($h_{\mathrm{refl}}(n)$) parts were different in each of the channels as obtained from the encoding of a SIR respectively into its $h_{\mathrm{L}}(n)$, $h_{\mathrm{C}}(n)$, $h_{\mathrm{R}}(n)$, $h_{\mathrm{LS}}(n)$, and $h_{\mathrm{RS}}(n)$ components.

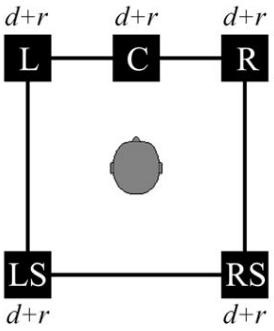

Fig. 1. The layout of the reference option A. Each of the loudspeakers is fed by both direct and reflected components of sound. The simplified symbols d (direct) and r (reflected) are used in Figs. 1–4. The indexes d_{L}, $d_{\mathrm{C}}, d_{\mathrm{R}}$, etc. were omitted to avoid cluttering of the diagrams.

In Fig. 2, option B is shown. It differs from option A in that the component r is removed from the center channel C and added in equal amounts to both front side channels L and R.

In order to stay consistent with option A, channel C in option B should be fed by signal d of the form

$$d(n) = s(n) * h_{\mathrm{Cdir}}(n), \tag{2}$$

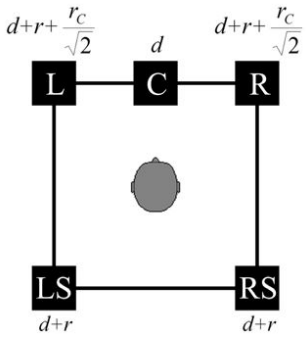

Fig. 2. The layout of option B.

but it was decided to take most of the possible advantage of dedicating a channel to the reproduction of DS by replacing d according to (2) with just the anechoic sound. This corresponds to replacing the $h_{\mathrm{Cdir}}(n)$ by a scaled Kronecker delta and thus eliminating considerable magnitude and phase distortion introduced by the measured IR.

The scaling factor for the Kronecker delta, i.e. the value of the amplitude of the anechoic sound $s(n)$, was found from

$$k_a = \frac{(s(n) * h_{\mathrm{Cdir}}(n))_{\mathrm{RMS}}}{s(n)_{\mathrm{RMS}}} \tag{3}$$

and thus channel C in option B was fed by the signal

$$d(n) = k_a s(n). \tag{4}$$

The signals fed to channels L and R (Fig. 2) were of the form

$$\left(d+r+\frac{r_{\mathrm{C}}}{\sqrt{2}}\right)_{\mathrm{L}} = s(n) * h_{\mathrm{L}}(n) + \frac{1}{\sqrt{2}}\left(s(n) * h_{\mathrm{Crefl}}(n)\right), \tag{5}$$

$$\left(d+r+\frac{r_{\mathrm{C}}}{\sqrt{2}}\right)_{\mathrm{R}} = s(n) * h_{\mathrm{R}}(n) + \frac{1}{\sqrt{2}}\left(s(n) * h_{\mathrm{Crefl}}(n)\right), \tag{6}$$

where $s(n)$ is the anechoic sound and the $\sqrt{2}$ term in the denominator is used to maintain the same total energy of RSs added to L and R channels as was removed from channel C. This is necessary to avoid a bias in perceptual experiment, as otherwise individual listeners might choose a particular option on the basis of its ratio of DS to RSs energy.

Channels LS and RS in option B receive identical signals as in option A.

Option C is presented in Fig. 3. In this option, the DS was removed from all but the center channel.

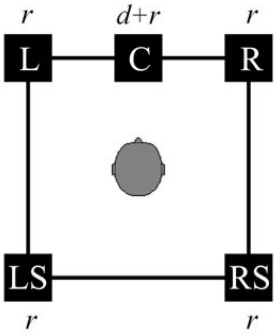

Fig. 3. The layout of option C.

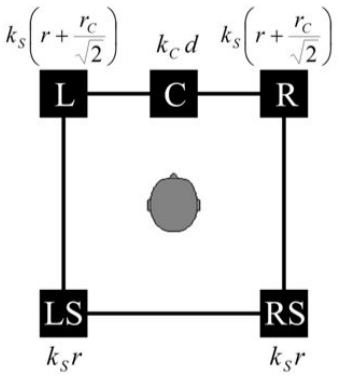

Fig. 4. The layout of option D.

Figure 4 presents option D. In this option complete separation was performed: channel C was fed by the DS while channels L, R, LS and RS were fed by only the RSs. The signal in channel C was

$$d(n) = k_c k_a s(n), \tag{7}$$

where

$$k_c = \frac{\sqrt{(s * h_{\mathrm{Cdir}})^2_{\mathrm{RMS}} + (s * h_{\mathrm{Ldir}})^2_{\mathrm{RMS}} + (s * h_{\mathrm{Rdir}})^2_{\mathrm{RMS}}}}{k_a \cdot s_{\mathrm{RMS}}} \tag{8}$$

(the notation of discrete time signals $x(n)$ was omitted to streamline the formula), and signals in channels L and R were of the form

$$\left(r + \frac{r_{\mathrm{C}}}{\sqrt{2}}\right)_X = \left(s(n) * h_{X\mathrm{refl}}(n) + \frac{1}{\sqrt{2}}\left(s(n) * h_{\mathrm{Crefl}}(n)\right)\right), \tag{9}$$

where X denotes L or R respectively, and k_s is

$$k_s = \frac{\sqrt{(s(n) * h_{\mathrm{Crefl}})^2_{\mathrm{RMS}} + (s(n) * h_{\mathrm{Lrefl}})^2_{\mathrm{RMS}} + (s(n) * h_{\mathrm{Rrefl}})^2_{\mathrm{RMS}} + (s(n) * h_{\mathrm{LSrefl}})^2_{\mathrm{RMS}} + (s(n) * h_{\mathrm{RSrefl}})^2_{\mathrm{RMS}}}}{\sqrt{(s(n)_{\mathrm{LD}})^2_{\mathrm{RMS}} + (s(n)_{\mathrm{RD}})^2_{\mathrm{RMS}} + (s(n) * h_{\mathrm{LSrefl}})^2_{\mathrm{RMS}} + (s(n) * h_{\mathrm{RSrefl}})^2_{\mathrm{RMS}}}}, \tag{10}$$

where LD and RD denote signals computed for channels L and R according to (9).

The multiplications by k_c and k_s in (7), (8), and (9) were used to compensate the shifts of DS and RSs energies.

The DS parts ($h_{\mathrm{dir}}(n)$) of measured IRs are not perfect impulses as they are blurred for various reasons. There are no definite rules of their division into DS and RSs parts. Usually, the first couple of milliseconds of the IR is assumed to be its DS part [8, 21–24]. The authors of this study measured the IR of their measurement setup in an anechoic chamber and obtained the value of about 3 ms.

The SIR used in the experiment was taken from the library developed by one of the authors (PM) [25]. It was measured in a small wooden Orthodox church with an RT60 of 1.1 s and an approximate volume of 750 m^3.

2.2. The listening experiment

The experiment was carried out in the anechoic chamber of the AGH University with an internal volume of 465 m^3 [26].

The standard 5.0 system setup [20] was used at $\pm 120^\circ$ angles for surround speakers. The radius of the system was 2.5 m. The listener's head was precisely positioned in the middle of this setup. Genelec 6010A two-way self-powered loudspeakers were used, with their tweeters positioned 1.2 m above the floor. The sensitivities of all monitors were calibrated with the use of pink noise at 80 dB (A) and the Svantek SVAN 959 sound pressure level meter. The audio interface was the RME Fireface 800. The experiments were run with a custom script written in MATLAB, providing a screen-and-mouse user interface. The audio excerpts were pre-computed and replayed from the computer's hard disk at the level of 80 dB (A). Each listener was given written instructions before the experiment.

Sixteen listeners participated, most of them aged 20 to 25. They had various levels of experience. None reported any hearing problems. Their audiometric thresholds were not measured, as according to [27] there is no effect of the listener's audiometric threshold on his or her performance in listening tasks with test material well above the threshold.

The experiment was run as a series of two-interval forced choice comparisons (2IFC), where each pair consisted of one audio excerpt reproduced according to two different options. There were six possible pairs of options: A vs. B, A vs. C, A vs. D, B vs. C, B vs. D, and C vs. D. Each trial was repeated ten times. The sequence of 60 trials (ten repetitions of six pairs) was randomized, as was the sequence within each interval.

The listener activated a pair of sounds by a software button and could listen to it only once. The break between the intervals in a pair was 100 ms. Next, he or she was asked the question: "Which version sounded better?". Their choice was saved by the click of one of two appropriately marked software buttons. This procedure was repeated for three different audio excerpts. The first was a phrase from J.S. Bach's *Bourrée* played on the guitar and lasting 9 s, the second was a phrase from H. Purcell's *Trumpet Voluntary* lasting 5 s, both taken from the Bang & Olufsen's "Music for Archimedes" anechoic recording CD that comes with the CATT Acoustics software [28]. The last was a phrase from W.A. Mozart's aria of Donna Elvira from the opera *Don Giovanni*, played on the cello and lasting 6 s, from anechoic recordings of

symphonic music [29]. Each listener evaluated 6 pairs $\times 10$ repetitions $\times$ 3 sources = 180 intervals, which took nearly an hour. Listeners were advised to take breaks.

3. Results and discussion

During the listening test, the trials containing a given pair were randomly distributed among the others. On that basis, individual presentations of each pair were assumed independent.

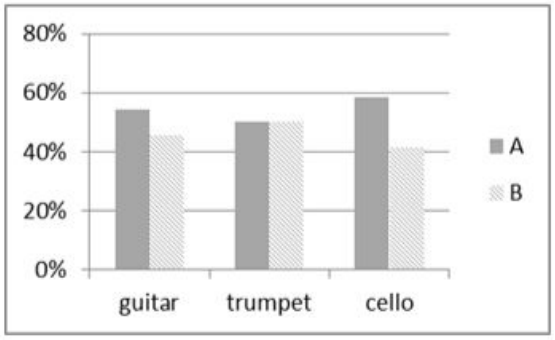

Fig. 5. The ratios of the total number of preferences in the perceptual comparison: option A vs. option B, for each of the sound sources.

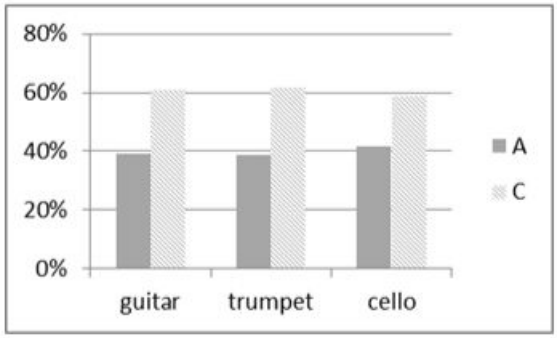

Fig. 6. The ratios of the total number of preferences in the comparison A vs. C.

It would not be correct to combine responses of an individual subject to all three audio excerpts as they were dependent, hence the results were analyzed for each excerpt. The null hypothesis was that listeners did not prefer either of the options (H_0: $p_1 = p_2$). The aim of this research was to find out whether any of the options was preferred to others, therefore two-tailed hypothesis testing was not appropriate and two independent tests were conducted, for $p_1 < p_2$ and $p_1 > p_2$, where p_1 and p_2 denote the probability of choosing the first or second element in the pair, respectively, determined from the results. The test of equal proportions based on normal statistics was performed for each sample consisting of the results of one audio excerpt, i.e. 16 (listeners) $\times$ 10 (repetitions) = 160 results. The required number of subjects was $n \geq 20$. The raw results are presented in Figs. 5 through 10. The histograms

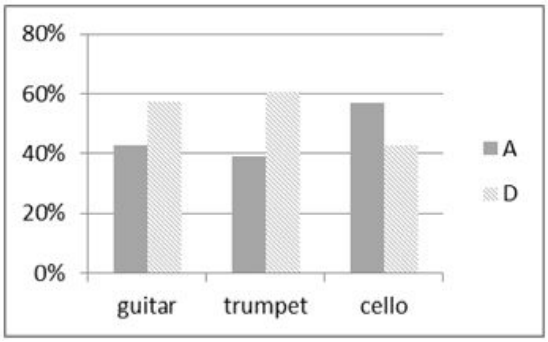

Fig. 7. The ratios of the total number of preferences in the comparison A vs. D.

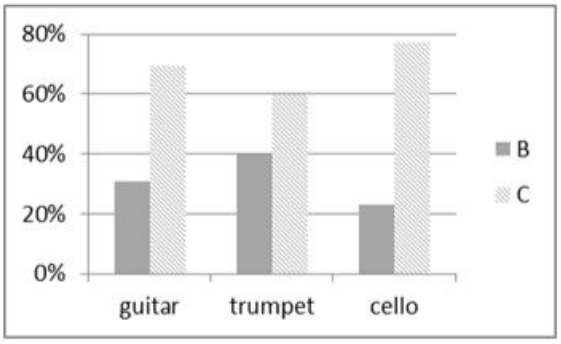

Fig. 8. The ratios of the total number of preferences in the comparison B vs. C.

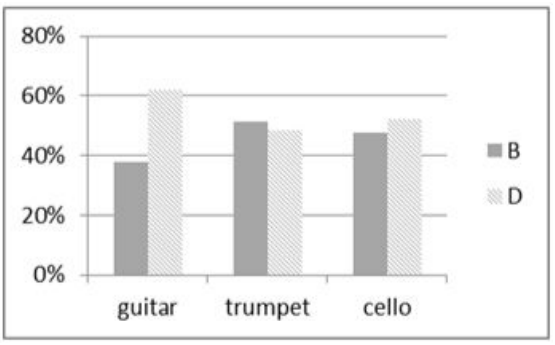

Fig. 9. The ratios of the total number of preferences in the comparison B vs. D.

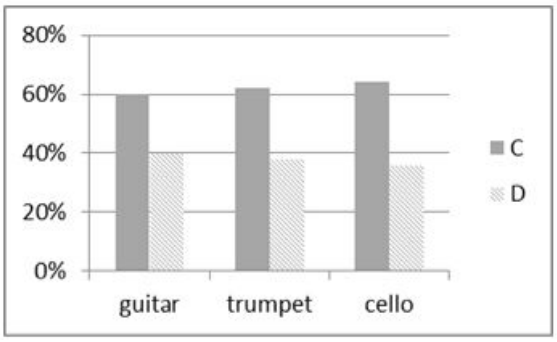

Fig. 10. The ratios of the total number of preferences in the comparison C vs. D.

show ratios of the numbers of preferences in perceptual comparisons, for each investigated pair of options, i.e. A vs. B, A vs. C, A vs. D, B vs. C, B vs. D, and C vs. D.

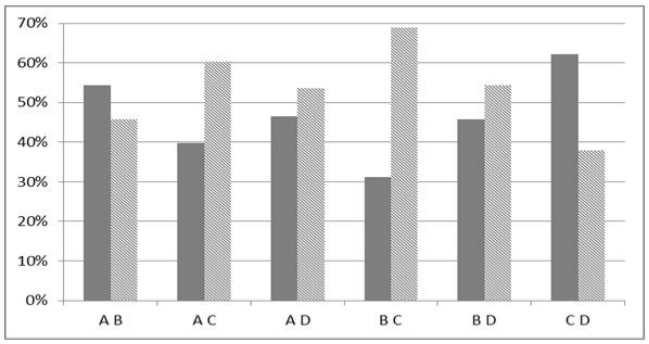

Fig. 11. The ratios of the total number of preferences in the perceptual comparison in all pairs. In each pair, the preferences for all three audio excerpts were summed up.

Figure 11 jointly presents histograms of percentages of preferences in all pairs, where in each pair the preferences for all three audio excerpts were summed up. The results including statistical evaluation are given in Tables I and II.

TABLE I

The results of hypothesis testing with the test of equal proportions based on normal statistics (H_0: $p_1 = p_2$), (H_1: $p_1 < p_2$), where p_1 and p_2 denote the probability of choosing the first or second element in the pair, respectively. $z_{\text{critical}} = -0.674$. Asterisks denote pairs in which H_0 was not rejected. Bold: H_1 accepted for all three audio excerpts.

Option	Instrument	z	p-value	Conclusion
A vs. B	guitar	1.434	0.924	*
	trumpet	0.000	0.500	*
	cello	2.869	0.998	*
A vs. C	guitar	-3.586	1.70×10^{-4}	**A < C**
	trumpet	-3.825	6.55×10^{-5}	**A < C**
	cello	-2.869	0.002	**A < C**
A vs. D	guitar	-2.390	0.008	A < D
	trumpet	-3.586	1.68×10^{-4}	A < D
	cello	2.390	0.992	*
B vs. C	guitar	-6.454	5.44×10^{-11}	**B < C**
	trumpet	-3.347	0.000	**B < C**
	cello	-9.084	5.25×10^{-20}	**B < C**
B vs. D	guitar	-4.064	2.41×10^{-5}	B < D
	trumpet	0.478	0.684	*
	cello	-0.717	0.237	*
C vs. D	guitar	3.347	1.000	*
	trumpet	4.064	1.000	*
	cello	4.781	1.000	*

Significant preference for all three sound excerpts was found for option C over option A, for option C over option B, and for option C over option D. This demonstrates consistent advantage of option C over the others, thus proving considerable listeners' preference for the removal of DS from all but the center channel.

TABLE II

The results of hypothesis testing for the cases where no significance was found in Table I, with the direction of testing inverted (H_1: $p_1 > p_2$).

Scheme	Instrument	z	p-value	Conclusion
A vs. B	guitar	-1.434	0.076	*
	trumpet	0.000	0.500	*
	cello	-2.869	0.002	A > B
A vs. D	cello	-2.390	0.008	A > D
B vs. D	trumpet	-0.478	0.316	*
	cello	0.717	0.763	*
C vs. D	guitar	-3.347	4.1×10^{-4}	**C > D**
	trumpet	-4.064	2.4×10^{-5}	**C > D**
	cello	-4.781	8.7×10^{-7}	**C > D**

This confirms the assumption given in the Introduction, that using only RSs in convolution is a better choice for all channels delivering ambience. Such an implementation of multichannel systems has been used in the Ambiophonics system, in its extension supplementing ambience in reproduction of stereo recordings [6, 18]. It has also been implemented in hardware convolvers for audio. It was also used in the Spatial Impulse Response Rendering (SIRR) multichannel sound reproduction method, where non-diffuse (DS plus early reflections) sounds are radiated from selected directions only, while the diffuse sounds are emitted by all loudspeakers [7, 8]. According to subjective opinions of some audio engineers, the removal of the DS may be advantageous or not, depending on circumstances in a particular recording. Farina et al. [13] noticed that removing DS and some early reflections in the Ambisonics rig resulted in poor localization of the sound arriving from the frontal stage. They mentioned an informal test comparing IRs with and without the DS plus early reflections. It indicated that the difference was not very evident. Apart from this remark, no other experimental evidence in this subject is known to the authors.

The removal of RSs from the center channel with its injection into front left and right channels (option B) was consistently evaluated as inferior to other options, but this was significant only in the B vs. C pair. This could indicate that the operation in option C is advantageous, while the operation in option B is not. However, closer analysis of the signals used revealed that there could be a problem of signal scaling. The formulae (5), (6), (8), (9), and (10) are based on the assumption that signals are uncorrelated. Yet there is some correlation between the IRs of frontal channels, especially in low frequencies, due to the low spatial resolution of the first-order Ambisonics microphone. Therefore, when the RSs of the center channel were added to left and right channels, the amplitude

in these channels might be over-sized, thus violating the assumption of maintaining constant ratio between energies of DS and RSs. Perhaps another approach for scaling signals should be applied.

The results for the pair A and D were significant for all excerpts, but contradicting and thus inconclusive (D preferred to A in two excerpts and A preferred to D in one). Should the supposition of the previous paragraph be true, this would bias the perceptual evaluation of option D as well.

4. Conclusions

The results clearly demonstrate the listeners' preference for option C, i.e. sound reproduction with DS radiated from only one channel. This observation is limited to the case of only one sound source, but still is an important hint for designing multichannel reproduction systems and for practical mixing techniques of audio material for these systems.

The removal of RSs from the center channel (option B) seems to be perceptually inferior, but because of possible problems with signal scaling, options B and D require further analysis and perhaps further perceptual evaluation.

Acknowledgments

This research was partly supported by AGH University grant no. 11.11.130.995. PM's work was supported by the Dean's grant no. 15.11.130.748.

References

[1] M. Kleiner, B.I. Dalenbäck, P. Svensson, *J. Audio Eng. Soc.* **41**, 861 (1993).

[2] M. Vorländer, *Auralization, Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, Springer Verlag, Berlin-Heidelberg, 2008.

[3] W. Woszczyk, T. Beghin, M. de Francisco, D. Ko, *Proc. 127th AES Conv.*, New York 2009, preprint 7856.

[4] W. Woszczyk, B. Leonard, D. Ko, *Proc. AES 40th International Conference*, Tokyo 2010.

[5] M. A. Gerzon, *J. Audio Eng. Soc.* **21**, 2 (1973).

[6] A. Farina, R. Glasgal, E. Armelloni, A. Torger, *Proc. 19th AES Conference on Surround Sound*, Schloss Elmau, Germany 2001.

[7] J. Merimaa, V. Pulkki, *J. Audio Eng. Soc.* **53**, 1115 (2005).

[8] V. Pulkki, Merimaa, *J. Audio Eng. Soc.* **54**, 3 (2006).

[9] F. Zotter, M. Frank, *J. Audio Eng. Soc.* **60**, 807 (2012).

[10] S. Tervo, J. Patynen, A. Kuusinen, T. Lokki, *J. Audio Eng. Soc.* **61**, 17 (2013).

[11] D. Scaini, D. Arteaga, *Proc. 55th AES Conference on Spatial Audio*, Helsinki, Finland 2014.

[12] S. Tervo, P. Laukkanen, J. Patynen, T. Lokki, *J. Audio Eng. Soc.* **62**, 30 (2014).

[13] Y. Ando, *J. Acoust. Soc. Am.* **62**, 1436 (1977).

[14] F. Toole, *Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers and Rooms*, Focal Press, Oxford, 2008.

[15] H. Imamura, A. Marui, T. Kamekawa, M. Nakahara, *Proc. 134th AES Conv.*, Berlin, Germany 2014, e-Brief 134.

[16] C.J. Moore, *Proc. 16th AES International Conference on spatial sound reproduction*, Munich, Germany 1999.

[17] P. Kleczkowski, *Arch. Acust.* **37**, 355 (2012).

[18] R. Glasgal, *Proc. 111th AES Conv.*, New York 2001, preprint 5426.

[19] A. Farina, R. Ayalon, *Proc. 24th International Conference: Multichannel Audio*, Banff, Canada 2003.

[20] ITU-R BS.775-3: Multichannel Stereophonic Sound System with and without Accompanying Picture, 2012.

[21] M. Kuster, *J. Acoust. Soc. Am.* **124**, 982 (2008).

[22] M. Kuster, *J. Audio Eng. Soc.* **57**, 403 (2009).

[23] M. Dunn, D. Protheroe, *Proc. 137th AES Conv.*, Los Angeles, USA 2014, preprint 9157.

[24] F. Menzer, C. Faller, H. Lissek, *Proc. of the IEEE,* **19**, 396 (2011).

[25] P. Malecki, Ph.D. Thesis, AGH University of Science and Technology, Krakow, 2013.

[26] A. Pilch, T. Kamisiński, *Arch. Acoust.* **36**, 955 (2011).

[27] P. Kleczkowski, M. Pluta, *Acta Phys. Pol. A* **121**, A120 (2012).

[28] Music for Archimedes, http://www.ramsete.com/Public/Aurora_CD/Anecoic/Archimedes/CD-cover/Archimedes.htm, accessed 2014 Nov. 10.

[29] J. Pätynen, V. Pulkki, T. Lokki, *Acta Acustica u. Acustica* **94**, 856 (2008).