Audio Engineering Society

Convention Express Paper 218

Presented at the AES 156th Convention
2024 June 15–17, Madrid, Spain

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 The complete manuscript was not peer reviewed.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. This paper is available in the AES E-Library (http://www.aes.org/e-lib), 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.

Immersive Audio Reproduction and Adaptability for Irregular Loudspeaker Layouts Using Modified EBU ADM Renderer

Toni Hirvonen[1], Carlos Tejeda-Ocampo[2], Ema Souza-Blanes[1], and Sunil Bharitkar[1]

[1]*Samsung Research America*
[2]*Samsung Research Tijuana*

Correspondence should be addressed to Toni Hirvonen (t.hivornen@samsung.com)

ABSTRACT

Immersive and multichannel audio commonly utilize a limited set of defacto output loudspeaker layouts. However, modern applications may result in using arbitrary numbers of speakers in varying directions. Aiming to realize broadcast-quality reproduction in such situations, we modify the EBU ADM Renderer to handle user-defined output layouts beyond the standard 10 available. Listening tests with object-based material are run to assess the performance of the proposed renderer, and to illustrate typical differences that can occur with different example cases. While the results show the utility of the proposed modification, we also recognize the lack of ideal layout-adaptability in typical gain-based immersive reproduction.

1 Introduction

The goal of Next-Generation Audio (NGA) is to provide consumers with listening experiences that are personalized, accessible and immersive. The NGA specification described in [1] lists recommended loudspeaker listening layouts in terms of speaker azimuth and elevation. To utilize these in practice, Audio Definition Model (ADM) is used as a common metadata standard for Advanced Sound Systems and NGA. It is defined in [2] as an open metadata format for production, exchange and archiving of immersive audio, including channel, scene and object-based content.

Under the NGA paradigm, content producers only have to create one singular audio mix with the corresponding ADM metadata. This paper investigates some of the practical limitations of this goal. Theoretically, the mix will be automatically adapted to the customer playback system and loudspeaker layout by a specialized renderer such as the EBU ADM Renderer (EAR) [3]. EAR provides a complete interpretation of ADM metadata and is capable of rendering audio signals to the 10 loudspeaker configurations specified in [1], making it a viable broadcast-quality open software solution. Consequently, EAR is the recommended loudspeaker reproduction method e.g. for the emerging AOMedia Immersive Audio Model and Formats (IAMF) - container [4].

While the canonical 10 loudspeaker layouts supported by EAR in principle cover the most common commercial systems, real-world customer set-ups are expected to deviate from this standard. New consumer layouts

outside those defined in [1] can be developed and gain popularity before being officially adopted to standards. Applications such as wireless speaker systems and automotive audio [5] may result in layouts with arbitrary numbers of speakers placed freely in the listening space. Even though the speakers can be somewhat compensated to a sweet spot regarding e.g. delay, it is not well defined how aspects like layout asymmetry are handled. Despite suggested margins of tolerance regarding loudspeaker positions in the layouts of [1], customer set-ups can exist outside these margins.

To provide a better listening experience to users unable to comply with [1], as well as forward compatibility with future reproduction systems, a flexible open renderer is needed. Furthermore, the renderer performance should be subjectively verified. While listening tests have been conducted in [6] to assess EAR's performance for a subset of the layouts defined in [1], there has been no assessment of the performance of EAR panning algorithms for irregular layouts. Also, as opposed to [6], the use of hidden references and anchors would make some listening tasks potentially clearer.

As mentioned, NGA content can be channel-, scene- or object-based. In case of object-based input, the ADM format allows for defining the spatial properties in either Cartesian-, or polar format. Although EAR constrains the target output layouts, some of it's core rendering algorithms, particularly those for egocentric panning using polar coordinates, are in principle usable for flexible rendering to irregular speaker layouts. The main challenge to utilizing EAR in this way is how to deal with the panning algorithms used for Cartesian input.

Cartesian representation has recently increased in popularity with Dolby Atmos [7, 8, 9], and employs the associated allocentric panning algorithms [10]. However, algorithms such as using a separate "panpot" for each axes, have certain symmetry requirements that make the standard allocentric panning unsuitable for irregular layouts [3]. The upside of employing such layout limitations is retaining an aspect of localization: Even though the allocentric paradigm is not in principle limited to a specific listening position, directional perception is predictable for the typical listening area.

While there are some proprietary, and some open solutions applicable for allocentric rendering with arbitrarily positioned loudspeakers, these do not account for the subjective perceived localization. Rather, they aim to allocate loudspeaker gains based on some other weighting function of source positions and speaker locations. For example, two similar distance-based methods DBAP [11] and VDP [12] were introduced as panning procedures for arbitrary layouts. As the listener position is not accounted for, spatial reproduction can be severely hindered by e.g. the precedence effect [13, 14].

As opposed to distance-based methods, our assumption for NGA reproduction is that the speaker layout is aligned to a preferred listening position, at least in terms of delay. Such compensation is in use in many modern reproduction systems, and can be realized independent of the renderer. One can then assume that reproduction based on anechoic amplitude panning gives a directional impression according to the original intention [14]. This is a typical assumption of loudspeaker gain - based immersive reproduction that does not involve more complex methods, like direct-ambient decomposition [15], or wave field synthesis [16]. However, we will also discuss how e.g. room effects in practice can affect gain-based rendering.

At the same time, academic methods, such as the amplitude panning algorithms of [17, 18], are not directly used here in favor of an engineering approach of modifying EAR and retaining as much of the functionality of [3] as possible; even though EAR employs VBAP-like amplitude panning in some cases, it also includes many refinements making it more suitable for real-life NGA applications.

Considering the previous points, and aiming to realize broadcast-quality reproduction for irregular layouts where loudspeakers can be placed in arbitrary directions, we have modified EAR toward this goal. Two listening tests with object-based material are run to assess the performance of the proposed renderer, and to illustrate typical differences that can occur in the spatial impression with different realistic layout example cases. These methods are detailed in Sec. 2, and Secs. 3 and 4 present the results and the related discussion.

2 Methods

2.1 Modified EAR

The EAR reference implementation of [3] in Python is available at `https://github.com/ebu/ebu_adm_renderer`. We modified its user interface and processing to permit flexible rendering for custom layouts. The modifications can be summarized as:

- The command-line interface was modified to accept a custom layout specified in YAML-format with the EAR polar speaker notation [3].
- Force the EAR built-in conversion [3] of the input ADM metadata to polar coordinates in custom layout rendering. The usage of EAR allocentric panning is thus omitted for custom layouts, as these algorithms have constraints that are not applicable. The modified EAR processes Cartesian input to custom layouts with egocentric panning, while being transparent to polar-format input.
- The built-in layout requirement that a speaker is present in every azimuth quarter was circumvented by placing a virtual speaker in any quarter where there are no real speakers present, and downmixing that virtual speaker to the nearest real speaker.

The above enable the modified renderer to operate with minimal changes to the EAR core functionality and panning algorithms. In case the custom layout is not used, or the custom layout is identical to one of the 10 layouts in the standard EAR [3, 1], the modified renderer functionality falls back to that of the standard EAR. The custom layouts are also not limited in that they can consist of any number of speakers (> 0) located in arbitrary directions.

2.2 Informal Comparison with Distance-Based Panning

To compare against an alternative capable of rendering Cartesian-format input to irregular layouts, we conducted informal listening with distance-based panning and EAR. Single static sound sources were rendered to $[-45, -15, 20]^\circ$ azimuth in a standard 5.1 layout using both the modified EAR and VDP [12]. These virtual sources were compared against the reference of a real loudspeaker in the target direction. In the central listening position, both methods were quite similar (despite VDP not accounting for the listener position).

However, at 45° with a large gap between the virtual source and loudspeakers' direction, and with off-sweetspot listening, it was found that the inherent property of distance-based methods of assigning some gain to every loudspeaker can result in unwanted effects in terms of spatial accuracy. As mentioned in Sec. 1, this is probably because of the precedence effect dominating localization in non-optimal conditions for amplitude panning.

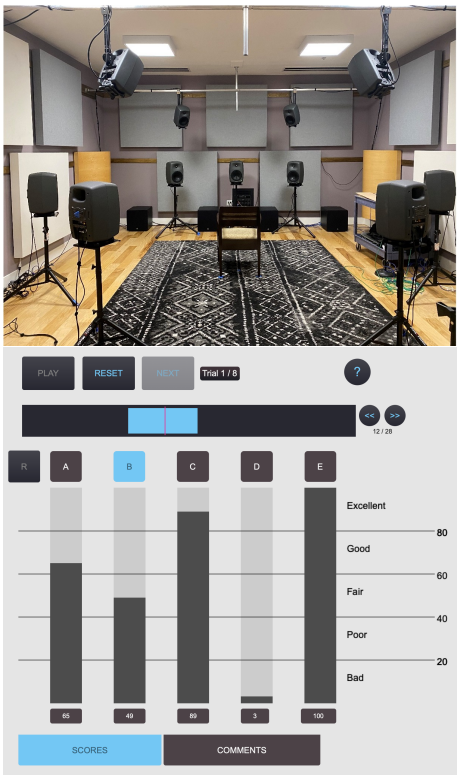


Fig. 1: Listening test room and user interface.

However, since distance-based methods are primarily intended for reproduction in non-distance compensated irregular layouts with no guarantee of directional accuracy, we decided against manufacturing a formal listening test to point out the truism that EAR-style pairwise panning in the horizontal plane can be more stable than the distance-based methods. Regardless, it is mentioned as motivation to utilize the built-in panning algorithms of EAR for more flexible layouts, instead of trying to add new algorithms like DBAP or VDP for the sake of handling Cartesian-format input.

2.3 Listening Tests

2.3.1 Setup

Formal listening tests for the modified EAR took place in a 7 m (L) x 5.33 m (W) x 3.05 m (H) listening room (Fig. 1) equipped with 16 loudspeakers supporting stereo, ITU 5.1 and 7.1.4, and non-ITU 5.1 and 4.1

Layout	Loudspeakers (azimuth)	Rendering
5.1 (href)	as in BS.2051 [1]	EAR
irregular 5.1	[45°, −20°, 15°, 100°, −150°]	EAR
irregular 5.1	[45°, −20°, 15°, 100°, −150°]	modified EAR
5.1 (anchor)	as in BS.2051 [1]	inverted EAR

Table 1: The conditions for Test 1.

regular and irregular layouts for channel-based playback. Loudspeakers were level-matched at the listeners's position and met the BS.1116 [19] specification in terms of room response curve within the 50 Hz -16 kHz frequency range.

A custom Max/MSP double-blind testing program and interface were used to run the tests (see Fig. 1). The interface, running on a tablet, allowed assessors to control the playback of the audio -including a segment loop function-, give ratings for each condition, and write comments.

2.3.2 Test 1

The first listening test employed a MUSHRA-style [20] methodology to assess the resulting localization accuracy of using the modified EAR compared to the reference EAR. As real-world constraints may force consumers to place their speakers in irregular approximations of the ITU layouts [1], we compared two 5-channel layout options: standard 5 non-LFE speakers of ITU 5.1 [1], and the other 5 at irregular positions. The layouts and test conditions can be seen in Table 1.

The reference condition was rendered using EAR and played through the ITU 5.1 layout. The two test conditions were the same audio rendered with the standard and modified EAR through an irregular 5.1 layout. Theoretically, the use of the standard EAR 5.1 rendering for irregular 5.1 should be detrimental, since it wrongly assumes that the loudspeaker locations are at BS.2051 locations [1]. In this test, we wanted to test this hypothesis in a realistic example layout case.

Test stimuli consisted of pink noise bursts and speech samples rendered to a selection of azimuth angles distributed in the horizontal plane around the listener position every 30 degrees: [-30, 30, -60, 60, -90, 90, -120, 120, 150, -150, 0, 180]° (see Fig. 2). The elevation in this test was 0°. Each trial corresponded to a single static source at one angle. Assessors were instructed to compare each condition to the reference in each trial, and to rate on a 0-100 MUSHRA scale how close localization-wise they were to the reference, and to try to disregard timbre effects. A low anchor condition consisted of a modified version of the reference signal, where left and right channels, as well as the front and rear channels, were inverted.

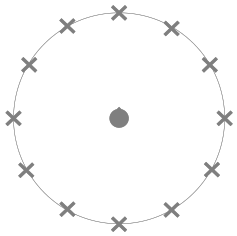


Fig. 2: The set of stimuli azimuth directions in Test 1

2.3.3 Test 2

A second test was conducted to assess the performance of the modified EAR with complex audio scenes and various layouts. These included ITU 7.1.4 and 5.1, and also non-ITU 4.1 and 4.1.2. The latter layouts are assumed to be realistic examples of a consumer wireless loudspeaker system etc. As the standard EAR does not have the option to utilize any 4-, or 6-channel output layouts, these require rendering by the modified EAR. Furthermore, an irregular version of 4.1 was included to compare how well layout asymmetry is handled by the modified EAR.

For this test, we utilized audio samples from the S3A object-based audio drama dataset [21]. This dataset contains three audio scenes in ADM-format (Family, Forest and Protest). For each scene, two short segments

Layout	Loudspeakers (azimuth; elevation)	Rendering
7.1.4 (href)	as in BS.2051 [1]	EAR
5.1	as in BS.2051 [1]	EAR
4.1.2	[(45°; 0°), (−45°; 0°), (110°; 0°), (−110°; 0°), (30°; 30°), (−30°; 30°)]	modified EAR
4.1	[(45°; 0°), (−45°; 0°), (110°; 0°), (−110°; 0°)]	modified EAR
irregular 4.1	[(45°; 0°), (−20°; 0°), (110°; 0°), (−150°; 0°)]	modified EAR
stereo (anchor)	as in BS.2051 [1]	EAR

Table 2: The conditions for Test 2.

were extracted, resulting in six test samples, as detailed in Table 3.

This test employed the MUSHRA methodology. In each trial, listeners would compare the different renderings of a sound scene. The reference stimulus was rendered using EAR and played through the ITU 7.1.4 layout. The test conditions were the same audio rendered with EAR for the case of ITU 5.1 and stereo, and with modified EAR for the two 4.1 layouts and 4.1.2. A standard stereo rendering was employed as the low anchor. The test conditions are summarized in Table 2.

Scene	Content
Family1	Family members arguing across rooms in a house
Family2	Family members arguing across rooms in a house
Forest1	Children playing around in a forest, elevated narration voice, non-diegetic music
Forest2	Child travelling through a river, elevated narration voice, non-diegetic music
Protest1	People chanting in a protest, helicopter flying above head
Protest2	People yelling at each other in the middle of a crowd protesting

Table 3: Descriptions of the six samples for Test 2. Samples were 9-14 seconds in duration.

The assessors were instructed to rate how close each condition appeared to the reference stimulus in terms of overall difference. Since the test material included complex scenes, it was noted that timbre (room modes) in addition to spatial aspects would potentially affect the grades. As the object-based content was dynamic, the final grade represented the average quality over the different segments of each sample.

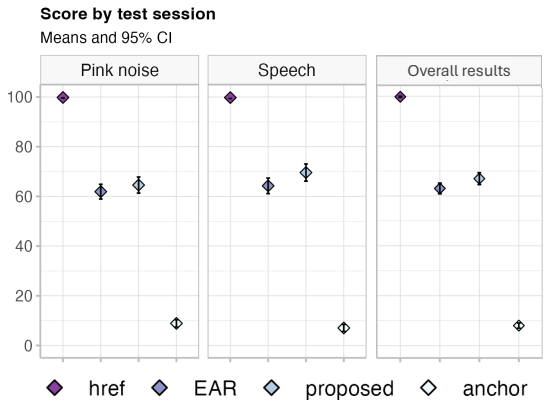


Fig. 3: Overall results for Test 1

3 Results

3.1 Test 1

12 subjects, including 9 experienced listeners, participated in Test 1. Each subject completed the rating task for all twelve angles for the pink noise and speech stimuli in separate sessions. The order of presentation of the test sessions was evenly distributed among the panel of assessors. The combined results are shown in Fig. 3.

The data were inspected for outliers, as well as for the ANOVA assumptions, which were not met ($p < 0.05$

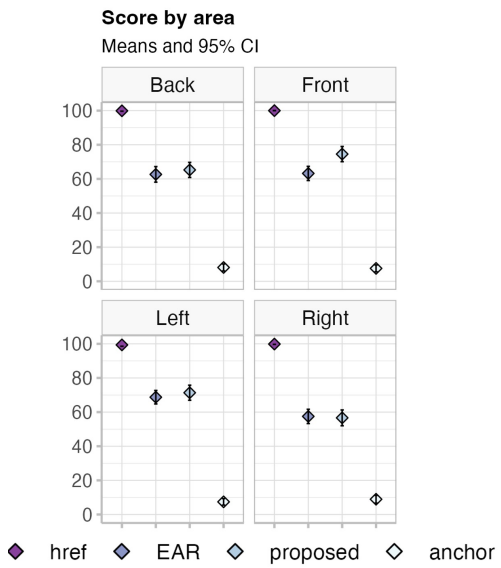


Fig. 4: Detailed results by source azimuth directional area for Test 1.

for Shapiro Wilk's test and Levene's test). Consequently, the non-parametric Friedman's test was employed to assess the effect of the angle and condition factors. Both were found to have a significant effect on the scores, along with their interaction (($p_{condition} = 2.98e^{-8}$, $p_{angle} = 1.72e^{-5}$ and $p_{interaction} = 2.2e^{-16}$).

Overall, the modified EAR scored higher than the regular EAR for the irregular layout. This trend seems to be different depending on the area the angle was located. The set of angles used in the test was grouped into 4 categories: angles superior to 30° and inferior to 120° were labeled as left or right respectively, while the remaining were grouped as front or back respectively. For the sides and back areas, the modified EAR scored either as high as the regular EAR or higher, as shown on Fig. 4. The front area presents an improvement brought by the modified EAR.

Pairwise t-tests were applied to compare the EAR-irregular and modified EAR-irregular conditions, after ensuring the subset of data for these two conditions followed a normal distribution. The difference in ratings between EAR-irregular ($M = 63$, $95\%CI = 2.16$) and flexEAR-irregular ($M = 67$, $95\%CI = 2.39$) was found significant, with $p = 0.000282$. Less discrimination was achieved with pink noise stimuli than with speech stimuli, with a significant difference found

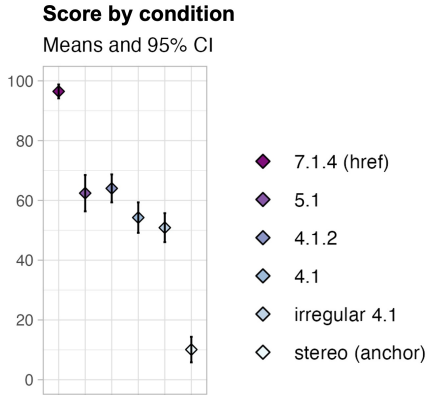


Fig. 5: Overall results for Test 2.

between the same pair of conditions for the speech sample ($p_{speech} = 0.00217$), while this difference was not found significant for the noise signal ($p_{pinknoise} = 0.0476$).

3.2 Test 2

Twelve listeners participated to the experiment. Results are displayed in Fig. 5. Data were checked for outliers, and one assessors was excluded from the analysis. As the first round of tests, the data did not meet the ANOVA requirements for normal distribution of residuals and homoscedasticity ($p < 0.05$ for Shapiro Wilk's test and Levene's test), and a non-parametric Friedman's test was used to estimate the effect of the condition (layout) and sample factors, with their interaction. Both the condition factor and the interaction between condition and sample had a significant effect on the ratings ($\chi^2_{condition}(5, N = 48) = 33.7$, $p_{condition} = 2.68e^{-6}$; $\chi^2_{interaction}(35, N = 288) = 223.6$, $p_{interaction} = 2.43e^{-29}$), while the sample factor was not found significant ($\chi^2_{sample}(5, N = 48) = 8.23$, $p_{sample} = 0.144$).

Wilcoxon paired comparisons were performed to assess the difference between mean scores of each pair of conditions. All conditions were rated significantly lower than the hidden reference, and higher than the low anchor. The 4.1 irregular condition was overall rated lower than the other layouts. The pairwise Wilcoxon comparisons revealed that although this condition had been rated significantly lower than the 4.1 and 4.1.2 conditions, the difference between the 4.1 regular and irregular was not significant ($p = 1$).

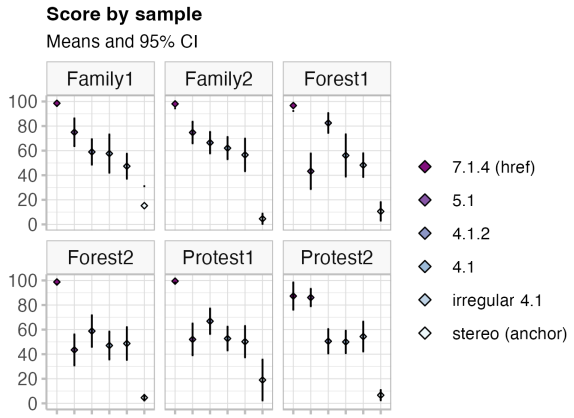


Fig. 6: Results per sample for Test 2.

Although the layout scores are similar, the overall mean score trend seems to slightly correlate with the respective loudspeaker counts (4, 5 and 6), as would be intuitive. Some variation in trends can be observed for the different pieces of content (see Fig. 6). The 4.1.2 layout scored higher for the tracks containing elevated content (both Forest and Protest1 scenes). The irregular 4.1 layout was rated slightly lower than the regular 4.1 condition in the Family tracks that featured more audio objects in the back area, and closer in scenes where the action was located more towards the front area.

4 Discussion

Test 1 showed some of the intuitive benefit in using the modified EAR that can take into account the actual layout loudspeaker directions. However, using a standard ITU 5.1. rendering for a layout deviating from that assumption resulted in a surprisingly small impairment. The reason is speculated to be that room modes and timbre effects interfere with spatial perception and virtual source directional quality, making localization more difficult than in anechoic conditions [14]. This can also explain why broadband pink noise is more spatially ambiguous than speech here; the noise virtual sources were less coherent likely due to the increased room effects on the contributing loudspeakers.

The effect of the listening space also played a role in Test 2. The promise of NGA and gain-based loudspeaker rendering being layout-agnostic does not fully hold, as all tested layout renderings were perceived significantly different from the 7.1.4 reference using the same object content. On the other hand, the results do not indicate that the 4-, and 6-channel layouts were alarmingly worse than the standard 5.1 rendering.

It may seem obvious that layouts like 5.1 are perceived different from 7.1.4 because of the missing height speakers. However, Test 2 did include a 4.1.2 layout with frontal height speakers that overlapped with the 7.1.4 layout. Furthermore, based on verbal listening impressions, the differences compared to the 7.1.4 were difficult to quantify merely as a sum of specific directional mismatches, but rather the overall timbre of the complex scene also differed. Hence, we hypothesize that achieving full transparency across layouts in a normal room is difficult with gain-based reproduction.

As conclusion, we postulate that the test results show the utility of the proposed modification in flexible rendering beyond the canonical 10 ITU layouts of the standard EAR [3]. If limited to gain-based reproduction, the modified EAR is usable, and the resulting spatial quality is within the expected limits with custom layouts of various speaker counts and locations.

While NGA covers channel-, scene- and object-based audio paradigms, this paper focused only on the object-based case. In EAR, the Point Source Panner used for object rendering is also employed for channel and scene-based rendering [3]. Thus, we considered object-based rendering to be the most critical case, at least in the context of EAR, and targeted our research accordingly. Examining the two other cases is left for future work.

5 Acknowledgements

Samsung Electronics and Samsung Research America (SRA) supported this work. The authors would like to thank the SRA Audio Lab staff for listening test participation and insights related to this research.

References

[1] ITU-R, "Advanced Sound System for Programme Production," Recommendation BS.2051-3, International Telecommunication Union, Geneva, 2022.

[2] ITU-R, "Audio Definition Model," Recommendation BS.2076-2, International Telecommunication Union, Geneva, 2019.

[3] ITU-R, "Audio Definition Model Renderer for Advanced Sound Systems," Recommendation BS.2127-1, International Telecommunication Union, Geneva, 2023.

[4] "What is IAMF?" `https://aomedia.org/iamf/`, 2024.

[5] Dupont, S., Vindrola, L., Melon, M., Chamard, J.-C., Roussarie, V., and Gazengel, B., "Wideband sound zone system for automotive applications," in *AES 153rd Convention*, Audio Engineering Society, 2022.

[6] "EAR Listening Test Report- Using the MS-IPM Method," Technical Report 43, European Broadcasting Union, Geneva, 2018.

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

[8] ETSI TS 103 190 parts 1 2, "Digital Audio Compression (AC-4) Standard," Standard, European Telecommunications Standards Institute, Geneva, 2018.

[9] Purnhagen, H., Hirvonen, T., Villemoes, L., Samuelsson, J., and Klejsa, J., "Immersive Audio Delivery Using Joint Object Coding," in *AES 140th Convention*, Audio Engineering Society, 2016.

[10] Riedmiller, J. C. and Tsingos, N., "Recent Advancements in Audio – How a Paradigm Shift in Audio Spatial Representation Delivery Will Change the Future of Consumer Audio Experiences," in *Spring Proceedings*, The Technical Forum, 2015.

[11] Lossius, T., Baltazar, P., , and de la Hogue, T., "DBAP–distance-based amplitude panning," in *International Computer Music Conference (ICMC)*, International Computer Music Association, 2009.

[12] Kostadinov, D. and Reiss, J. D., "Spatial Audio Matlab Toolbox," 2009.

[13] Wallach, H., Newman, E. B., and Rosenzweig, M. R., "The precedence effect in sound localization," *The American Journal of Psychology*, 62, p. 315–336, 1949.

[14] Blauert, J., "Spatial Hearing: The Psychophysics of Human Sound Localization (Revised Edition)," 1997.

[15] Pulkki, V., "Spatial Sound Reproduction with Directional Audio Coding," *J. Audio Eng. Soc*, 55(6), 2007.

[16] Berkhout, A. J., "A Holographic Approach to Acoustic Control," *J. Audio Eng. Soc*, 36(12), 1988.

[17] Clark, H., Dutton, G., and Vanderlyn, P., "The "Stereosonic" recording and reproducing system," *IRE Transactions on Audio*, AU-5(4), pp. 96–111, 1957.

[18] Pulkki, V., "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *Journal of Audio Engineering Society*, 45(6), pp. 456–466, 1997.

[19] ITU-R, "Methods for the subjective assessment of small impairments in audio systems," Recommendation BS.1116-3, International Telecommunication Union, Geneva, 2015.

[20] ITU-R BS.1534-2, Standard, International Telecommunication Union, Geneva, CH, 2014.

[21] Woodcock, J., Pike, C., Coleman, P., Franck, A., and Hilton, A., "S3A object-based audio drama dataset," 2020.