Audio Engineering Society

Conference Paper 18

Presented at the AES 5th International Conference on Audio for Virtual and Augmented Reality
2024 August 19–21, Redmond, WA, USA

This paper was peer-reviewed as a complete manuscript for presentation at this conference. 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.

Object Coding Masking Model Evaluation with Opus Codec for Next Generation Audio Applications

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

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

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

ABSTRACT

The performance of the prevailing audio coding methods is not fully established in some object-based audio situations such as six degrees of freedom (6DOF) listening. We implement a modified Opus codec capable of processing object input and analyzing the mutual masking between objects without information about object metadata or listener position. MUSHRA results indicate that the proposed codec is suitable of coding object content at realistic transmission rates while still retaining excellent quality. The model accounting for mutual masking seems relatively suitable for next generation applications involving listener movement and 6DOF situations. We also make available our Opus implementation[1] as an open-source object coding solution and alternative for the current parametric codecs.

1 Introduction

Although the principles of traditional audio coding are established since the work of Brandenburg et al [1], Next Generation Audio (NGA) applications may warrant re-validating these methods. For example, recent pushes in consumer immersive audio and Virtual Reality (VR) [2] have revitalized interest in delivering improved spatial audio.

A characteristic of NGA is the increased amount of data and the number of audio tracks. This is obvious for multichannel formats (5.1, 7.1.4, 22.2), and for scene-based capture utilizing Higher Order Ambisonics (HOA) [3]. Alongside these, object-based audio is one of the new reproduction paradigms [4]. Due to its role as an immersive content creation tool [5], the demands on object audio transmission and storage are also considerable. Thus, efficient audio coding is essential for making NGA feasible. For multichannel-, and HOA audio, such methods have been proposed and standardized [6, 7]. Although the same systems include methods for object audio coding [8, 9], we argue these are not as established in all cases.

The principal idea of lossy audio coding is to exploit hearing system inaccuracies. One such aspect is to analyze how well energetic masking hides certain parts, and assign coding distortion to those time-frequency tiles by using less bits for coding [10]. There is however a question how well the typical masking models applied in channel-based coding are suited for object

[1] https://github.com/thirv/opus-obj (with dependencies to other modified Xiph projects at https://github.com/thirv)

coding. The proper application of masking analysis in object audio depends on the listening situation: In one extreme, the objects can be fully interactive, and in the other, objects are used as a content creation tool and reproduced purely as spatial elements without interaction. In the former case, object playback gains can be modified, and each object can even be listened solo without other interference. Here, joint coding may not benefit much compared to discrete coding instances. In the latter scenario, it is however possible to exploit mutual masking of objects [11].

A question arises what to do in situations that lie between these two extremes. Specifically in VR Six degrees of freedom (6DOF) applications, the listener can move near to one sound source, while other sources are still active. Since the listening situation is not known at the encoding stage, but mutual masking can still occur, it is not clear what is the best coding strategy. Although efficient proprietary methods for audio object coding have been introduced [8, 9], there has not been much consideration on these questions beyond relying on traditional multichannel psychoacoustics.

Our motivation for this paper is twofold: 1) the lack of availability of an open-source joint audio object codec, and 2) the existing closed methods have not been tested for applications like 6DOF. We propose an extension to the popular Opus audio codec [12]. The base Opus codebase is modified to accept simultaneous input of multiple objects, and to assign bitrate between them based on joint time-frequency masking analysis. In addition to traditional quality evaluation, our listening tests include a simulated 6DOF situation: The effect of listener positional movement was examined with two specific listening locations, and by allowing subjects to move freely. Sec. 2. gives more details on the methods, and Secs. 3 and 4 present results and discussion.

2 Methods

2.1 Joint Masking Analysis for Object Audio

Compared to mono audio, there are additional, and possibly related considerations in multichannel audio compression: 1) Accounting for signal correlations and the perceptual covariance structure between the content tracks. 2) Determining how much other tracks contribute to perceptual masking of the signal fine structure of each track. For 1), known methods range from de facto heuristics like mid-side stereo coding to parametric multichannel coding [6, 7]. While relevant for channel-based audio and HOA, 1) is traditionally thought as not so interesting for object audio since objects are assumed uncorrelated. Therefore, this paper will not address 1). In contrast, our main contribution is to analyze 2) for object audio.

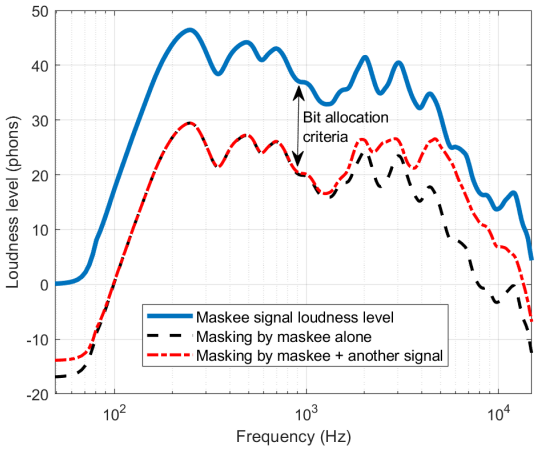


Fig. 1: An example speech signal segment, and the associated masking curves in two different cases: Signal is presented in isolation or the interference from other signals is considered insignificant ("maskee alone"), and in the presence of another high-frequency masker signal ("maskee + another signal"). Ideally, coding distortion is allocated differently in these two cases. For the illustration, we utilized loudness-based modeling [13, 14], and show the results in loudness level - domain along with the maskee signal itself.

Fig. 1 illustrates an idealized masking analysis for audio coding. The main idea is to calculate a masking curve, and use the distance of the target signal level from the curve as a relative measure. Usually the masking curve is not considered an absolute threshold of inaudibility. Rather, coding distortion is allocated by assigning available bits to represent perceptually important time-frequency tiles more accurately. Quantization can be performed in some latent or learned domain, such as in energy-normalized frequency bands [12]. In the example of Fig. 1, it can further be seen that the high frequencies of the maskee signal would get different amounts of coding distortion depending on

how much an another signal is deemed to contribute to the masking curve. It should be noted that theoretical auditory masking analysis alone is not the only factor for bit allocation in audio coding. However, the overall strategy generally holds.

As mentioned in Sec. 1, accounting for the spatial nature of objects is not trivial. The question is how much do differing spatial properties contribute to release from masking between objects? The underlying psychoacoustic model of this work assumes that spatial release from masking is not very notable in a normal listening room [15] or in a simulated aural space. This is different from an ideal anechoic space without reflections, where spatial masking release can be significant (up to 15dB) [16]. Furthermore, the coding artifacts are mostly hidden by energetic, rather than informational masking [17].

This assumption negates the need for pre-rendering, or complicated spatial analysis based on e.g. object positional metadata. In other words, the goal of the proposed method is to remain agnostic to the listening setup, room acoustic properties, etc. Without notable masking release, each object can be assumed to be masked by the sum signal of all audio, and the masking curve is the sum of all objects' masking energies. This is also more efficient than naive discrete coding of assigning all objects equal bitrates and separate mono codec instances.

In [11], the authors introduced a pre-processing method for assigning importance for the elements of object audio content, and using the result to assign bitrates with legacy audio codec instances. The method successfully utilized the previous paradigm of minimal masking release. However, integrating same principles within the frame-based processing of the codec itself allows for a more fine-grained and efficient allocation. The next Section details our proposal to do that.

2.2 Modified Opus Codec for Object Audio

As mentioned, our implementation is a fork from the Opus source code [18]. We copy much from the existing Opus *multistream surround encoder* intended for the most popular multichannel formats (e.g. 5.1). It analyzes mutual masking so that static channels are estimated to contribute to each other's masking by fixed but varying amounts; e.g. frontal channels are emphasized more in bit allocation.

Our notable simplification is to skip any such spatial unmasking adjustments. In addition to the psychoacoustic arguments discussed in Sec. 2.1, the dynamic nature of objects makes them more prone to mask each other consistently. Furthermore, the frontal channels will probably carry important audio more consistently than any given object. We therefore hypothesize that the simple masking analysis shown in Algorithm 1 is able account for object content idiosyncrasies.

Algorithm 1 Joint masking analysis for N audio objects in the Opus codec

Require: $n \in [1,N], b, \mathbf{x}$
$e_{n,b} \leftarrow ||\mathbf{x}_{n,b}||_2$
$e_{n,b} \leftarrow weightspread(e_{n,b}, b)$
$m_{n,b} \leftarrow e_{n,b} / \sum_n e_{n,b}$
$(m_{n,b} \leftarrow w_n m_{n,b})$
$r_{n,b} \leftarrow allocation(m_{n,b}, ...)$

where $\mathbf{x}$ indicates the signal frame transform-domain vector, n and b the object-, and frequency band indices, respectively. Frequency-band weighting, across-band spreading, as well as the final bitrate allocation r are determined via unmodified Opus algorithms. Object-specific weighting w_n can be optionally used to adjust the perceptual importance of that object. For this work, we use $w = 1$.

The method works as follows: As in Fig. 1, the ratio of each objects' frequency band energy to the masker signal represents the perceptual importance. The masker signal computation assuming minimal spatial masking release is simply the sum of all the object perceptual energies, and independent of any spatial properties.

Although not examined in this paper, a simple method to introduce perceptual masking release effects would be weighting a specific object via some factor w_n. This weight can be depend on e.g. spatial properties, the signal content, informational masking, playback setup, content creator preferences etc. However, using other values besides $w = 1$ is left for future studies.

The rest of the bit allocation engine is codec-specific, and typically heuristic in nature; in Opus it utilizes e.g. frequency weighting according to hearing sensitivity and other factors, mask frequency spreading, the available bit reservoir etc. for the final decision [12]. We skip any application of the "channel offset" heuristic of the original Opus multichannel masking analysis (i.e. altering the masking of some channels according to

the total number of channels). We also do not couple channels (i.e. group pairs together for efficient mid-side coding etc), since we assume the object waveform signals are uncorrelated.

Other practical aspects considered in the implementation were: 1) the number of audio tracks (objects) can be typically be larger than channels, and 2) objects can be sparse in terms of energy. For 1), the proposed method can in principle process an unlimited number of input tracks, We limit the implementation to 32 at this time and use 21 in our tests. Related to 2), we rely on the Opus frame processing and bit reservoir, but noticed that the requested rate of Opus often deviated from the true rate, moreso than with channel coding. The desired true rate was obtained via iterating the request.

2.3 Comparison to Existing Methods

The pioneering MPEG Spatial Audio Object Coding (SAOC) [8] utilizes a separate pseudo-Quadrature Mirror Filterbank (QMF) for parametric coding. Parameters are sent for each perceptually spaced frequency band, and applied to another signal (i.e. downmix). SAOC aims to reconstruct band object energies, and possibly other parameters from mono or stereo downmixes. SAOC further re-instates the correlations between the rendered output channels, necessitating the use of signal decorrelators. SAOC achieves additional sideinfo bitrates as low as 3 kbit/s per object, but requires joint operation between coding and rendering.

Motivated by Dolby Atmos content [5, 19], AC-4 Joint Object Coding (JOC) [9] proposed further parametric methods for higher-quality coding of individual objects: The QMF frequency band parameters are full least-squares solutions for arbitrary downmixes, and docorrelation is applied directly in the object-domain. JOC also utilizes dynamic spatial clustering techniques both for dealing with large number of objects, and for constructing adaptive downmixes. To our knowledge, JOC nor SAOC have been tested in applications such as 6DOF listening.

Our proposed method is inherently not parametric and we do not rely on a downmix. Parametric reconstruction from downmix frequency bands is difficult to scale to transparent quality due to rigid signal models. Also, the optimal downmix arrangement as a function of rate has to be decided somehow. In contrast, we quantize

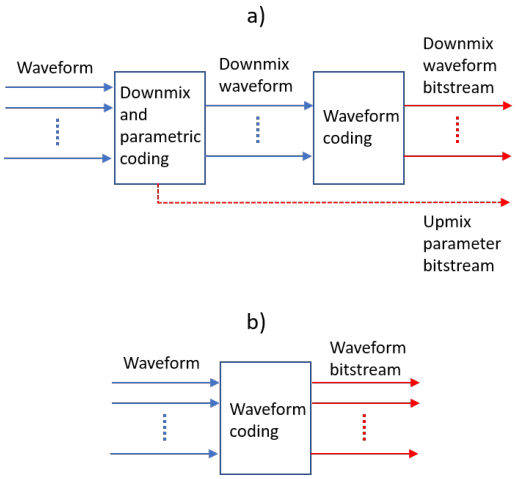


Fig. 2: Conceptual difference between a) parametric coding utilizing a downmix, and b) the present proposal of direct joint waveform coding. Solid arrows represent audio tracks, and dotted arrow auxiliary metadata. Red color indicates compression.

the object signals directly in the waveform coding Modified Discrete Cosine Transform (MDCT) - domain. Fig. 2 illustrates this conceptually. The method offers a simple way to scale toward transparency by increasing the total bitrate, and omits the use of additional filterbanks or decorrelators. While SAOC and JOC implicitly exploit the mutual masking between object signals under similar mutual masking assumptions, our method is arguably more explicit on how precisely the bits are allocated per object and why. The parametric methods also do not offer an easy way to emphasize individual objects in case of partial unmasking.

Additionally, neither JOC nor the present method are dependent on the rendering similarly as SAOC. Lastly, some NGA object coding proposals have utilized the naive bit allocation of assigning constant equal waveform coding bitrate to all objects. We do not consider such methods here since the present study relates to perceptual audio coding.

2.4 Subjective Tests

The implemented object codec was evaluated in a series of listening tests. As material, we utilized audio from the S3A object-based audio drama dataset [20]

and Netflix open content [21]. For each listening test sample, a 10 - 15 second segment of 20 critical objects was selected from these databases. The utilized audio scenes are described in Table 1. A total of 8 test samples were used. For each 20-object sample, a special static speech object was added to -90° azimuth to provide a critical target for the 6DOF-aspect of our test.

Scene	Content
"Family"	Family members arguing across rooms in a house.
"Forest"	Children in a forest, elevated narration voice, non-diegetic music.
"Protest"	People chanting and yelling in a protest, helicopter flying.
"Sol Levante"	Orchestral cinematic music, dramatic and bombastic.

Table 1: Descriptions of object audio scenes used in the listening tests. Eight 10 - 15 second clips of 20 critical objects were selected from these scenes.

The purpose of adding a static object to -90° relates to our goal of testing the 6DOF-aspect of NGA reproduction. Although we are not utilizing VR per se, the potential extent of spatial masking release and the validity of the proposed object masking model can be examined in a real listening room by allowing different listening locations. We adopt two static listening positions as seen in Fig. 3: One central and one closer to the special static object. Additionally, an additional test allowed free listener movement inside the loudspeaker layout. As result, the subjects repeated the same test three times with randomized session orders.

The central listening position can be viewed as the ideal situation for the simple masking model not accounting for any spatial masking release. The second listening position near the -90° source represents a realistic case where the listener moves closer to a prominent object. The loudness of the static -90° object was on average slightly lower compared to the rest of the scene. Thus, it was audible and subject to attention at all listening positions. Listener moving closer would allow coding distortion to be unmasked more in case the hypothesis of low masking release is not valid. This would then be observable as a bigger difference in the quality vs bitrate in this listening position compared to central.

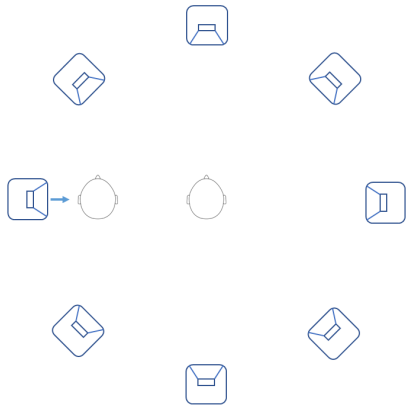


Fig. 3: A schematic illustration of the listening test setup with a symmetric 8-speaker layout. Two static listening positions were used: A traditional central position, and an off-center position 1 meter from the loudspeaker at -90° azimuth. Additionally, the subjects were able to move freely within the loudspeaker setup in a third test. The material was realistic dynamic object audio, with a special static speech object added to -90° (blue arrow).

Additionally, a third test variation allowed the listener to move freely within the area of the loudspeaker layout, while giving their grades on a wireless tablet interface. As they could get close to each physical speaker, this is almost equal to evaluating the channels of the rendered signal in isolation. The goal was to provide a worst-case low limit for the quality. However, it is arguably not a typical realistic use case; sometimes the subjects could effectively isolate very quiet parts of the sound scene, instead of moving toward any actual discrete objects. Nevertheless, this situation is included to fully push the masking model.

Otherwise, the tests employed the normal MUSHRA [22] methodology with a hidden reference and a low anchor. In each trial, listeners would compare the different renderings of a sound scene at three different total VBR bitrates: 600, 450 and 300 kbit/s (29, 21, and 14 kbit/s per object on average). Comparison with naive discrete object coding is not included, as the typical bitrates for high quality would exceed 1 mbit/s (assuming equal 48-64 kbit/s per object).

Fig. 3 encapsulates the listening test setup by illustrating the loudspeaker layout and the two static listening

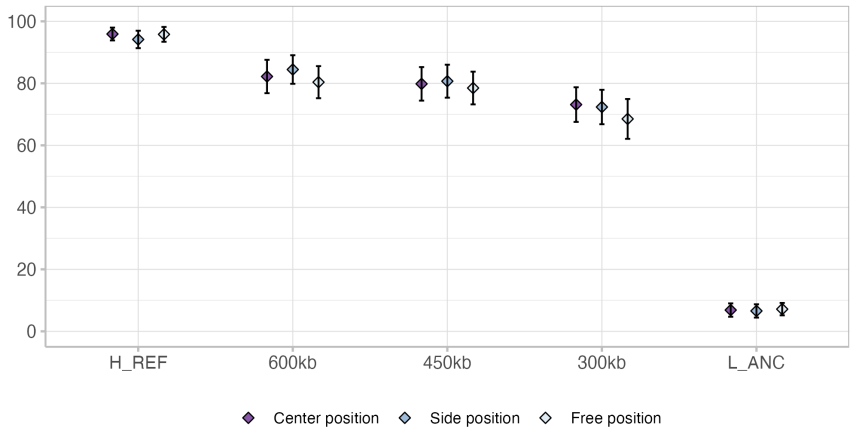


Fig. 4: Proposed codec MUSHRA results for 8 subjects repeated in three different listening positions: "Center" and "Side" indicate the central and the offset listening position in Fig. 3, respectively. "Free" indicates that subjects could move freely within the loudspeaker setup. The tested conditions included three total bitrates (600, 450 and 300 kbit/s) for the 21-object content, a hidden reference (H_REF) and a low-pass anchor (L_ANC).

positions used. The speaker layout was symmetrical, rather than a typical one employed in NGA multichannel formats (5.1., 7.1 etc) in order to better simulate a VR-like situation and equalize all possible listening locations. Height was omitted since that could not be varied by moving the listening position.

To render the test samples, we utilized the Audio Definition Model (ADM), which is the common open standard for metadata for NGA [19]. All the object-based material and the added static object represented their spatial properties with ADM. To produce the final 8-channel outputs, the ADM format was processed by the EBU ADM Renderer (EAR) [23]. We utilized a customized version of EAR [24], as the standard EAR is limited in output layout selection.

3 Results

Fig. 4 presents the mean MUSHRA scores and confidence intervals for 8 experienced listeners. As evidenced from the low scores for hidden reference and anchor, the listeners were relatively critical. Generally, the quality is at the MUSHRA "excellent" range for both 600 and 450 kbit/s (29 and 21 kbit/s per object), and "good" at 300 kbit/s (14 kbit/s per object).

As illustrated in Fig. 5, the quality varied somewhat depending on the material. It can be seen that e.g. a speech-heavy scene ("Family") was less critcal than a scene including various ambient and dynamic natural sounds, such as rain ("Forest"). For some critical samples, 300 kbit/s did introduce quality issues. However, many cases were also close to transparent at higher rates, as they were mistaken for the reference.

To assess the formal significance of different test aspects, a robust ANOVA [25] was performed using 3 factors: *Condition*, *Sample* and *Position*. The listening position did not significantly affect the scores (*Position* $p = 0.08$), nor did its interactions. (*Condition*Position* $p = 0.55$, *Sample*Position* $p = 0.58$). Unsurprisingly, there are highly significant ($p < 10^{-5}$) effects from *Condition*, *Sample* and their interaction.

4 Discussion

The listening test results and analysis indicate that the simple mutual masking hypothesis works overall similarly in different listening situations. The rate/quality trade-off was not statistically affected by the listening position. This implies that the spatial unmasking effects in a typical room / VR space with reflections are not a primary concern in audio coding. While the results may not fully generalize to anechoic spaces, the ITU room is relatively dry and critical.

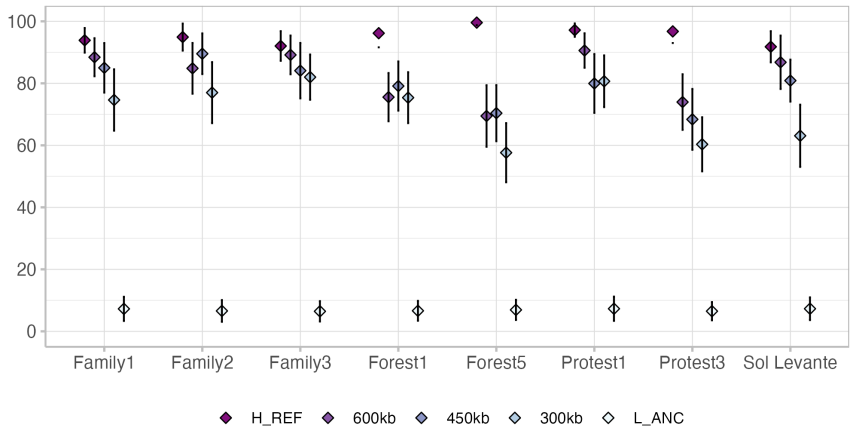


Fig. 5: Proposed codec MUSHRA results per test sample. Results are averaged over all listening positions.

The off-center listening position represented a realistic situation of focusing on, and moving closer to an important discrete object in 6DOF listening. While free movement within the layout is an interesting case, may not be fully corresponding to VR reprodution: If two real loudspeakers pan a source, or contribute to a background scene, that object or scene would in VR probably be reproduced with binaural rendering without the ability to isolate virtual source components such as loudspeakers. We hypothesize that the present free listening represents the worst-case low quality limit, with VR listening typically being less critical.

Thus, despite operating without the need of object metadata or listener position, the codec seems usable against unmasking effects caused by typical 6DOF listening. Although we did not investigate full interactivity in terms of modifying the playback gains, the results give indication toward reasonable suitability in that application as well. The proposed object-specific importance weighting in Algorithm 1 can further add robustness against unmasking.

The present implementation for audio object coding offers an open-source alternative to the previous parametric methods. "Excellent" quality on the MUSHRA scale are achieved with realistic bitrates. We skip additional filterbanks and decorrelators in favor of mutual waveform coding. The implementation also serves as a baseline for further evaluations. Although built within Opus, the masking model and the waveform coding principles can be utilized with other core coding engines. Despite not investigated, the method is not limited to dynamic object content; since no positional metadata is required for the analysis, the input can also static audio with unusual/irregular channel layouts etc.

References

[1] Brandenburg, K., "MP3 and AAC Explained," in *AES 17th International Conference on High Quality Audio Coding*, Audio Engineering Society, 1999.

[2] `https://aomedia.org/iamf/`, 2024.

[3] Daniel, J., Moreau, S., and Nicol, R., "Further Investigations of High-Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging," in *Audio Engineering Society Convention 114*, 2003.

[4] ITU-R BS.2051-3, "Advanced sound system for programme production," Standard, International Telecommunication Union, Geneva, CH, 2022.

[5] 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.

[6] ISO/IEC 23008-3:2022, "Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio," Standard, International Organization for Standardization, Geneva, CH, 2022.

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

[8] Engdegård, J. et al., "MPEG Spatial Audio Object Coding - The ISO/MPEG Standard for Efficient Coding of Interactive Audio Scenes," in *AES 129th Convention*, Audio Engineering Society, 2010.

[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] Painter, T. and Spanias, A., "Perceptual coding of digital audio," *Proceedings of the IEEE*, 88(4), pp. 451–515, 2000.

[11] Hirvonen, T., Tejeda-Ocampo, C., Souza-Blanes, E., and Bharitkar, S., "Perceptually Motivated Bitrate Allocation for Object-Based Audio Using Opus Codec," in *AES 155th Convention*, Audio Engineering Society, 2023.

[12] Valin, J.-M., Maxwell, G., Terriberry, T. B., and Vos, K., "High-quality, low-delay music coding in the Opus codec," *arXiv preprint arXiv:1602.04845*, 2016.

[13] ISO 532-2:2017, "Acoustics - Methods for calculating loudness - Part 2: Moore-Glasberg method," Standard, International Organization for Standardization, Geneva, CH, 2022.

[14] van de Par, S., Kohlrausch, A., Charestan, G., and Heusdens, R., "A new psychoacoustical masking model for audio coding applications," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, 2002.

[15] Koenig, H., Allen, J. B., Berkley, D. A., and Curtis, T. H., "Determination of masking-level differences in a reverberant environment," *J. Acoust. Soc. Am.*, 61, p. 1374–1376, 1970.

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

[17] Kidd, G., Mason, C., Brughera, A., and Hartmann, W., "The Role of Reverberation in Release from Masking Due to Spatial Separation of Sources for Speech Identification," *Acta Acustica united with Acustica*, 91, pp. 526–536, 2005.

[18] `https://github.com/xiph/opus/`, 2024.

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

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

[21] `https://opencontent.netflix.com/`, 2024.

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

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

[24] Hirvonen, T., Tejeda-Ocampo, C., Souza-Blanes, E., and Bharitkar, S., "Immersive Audio Reproduction for Irregular Loudspeaker Layouts Using Modified EBU ADM Renderer," in *AES 156th Convention*, Audio Engineering Society, 2024.

[25] Feys, J., "Nonparametric Tests for the Interaction in Two-way Factorial Designs Using R," *R J.*, 8, p. 367, 2016.