Compression of Higher Order Ambisonics with Multichannel RVQGAN

Toni Hirvonen and Mahmoud Namazi

Samsung Research America
[t.hirvonen,nima.namazi]@samsung.com

Abstract

A multichannel extension to the RVQGAN neural coding method is proposed, and realized for data-driven compression of third-order Ambisonics audio. The input- and output layers of the generator and discriminator models are modified to accept multiple (16) channels without increasing the model bitrate. We also propose a loss function for accounting for spatial perception in immersive reproduction, and transfer learning from single-channel models. Listening test results with 7.1.4 immersive playback show that the proposed extension is suitable for coding scene-based, 16-channel Ambisonics content with good quality at 16 kbit/s.

Introduction

Audio compression is an important aspect of storage and transport, in particular with respect to new immersive formats. The size and dimensionality of this type of content motivates the discovery of new, efficient coding methods. This is especially true in the case of Higher Order Ambisonics, a prominent surround sound methodology [1] which has in recent times been a popular solution especially for spatial sound capture. At the same time, the rise of data-driven processing methods and increase in available computational power make it feasible to explore methods beyond conventional audio coding.

This paper presents a simple, yet general and efficient, multichannel extension to the popular RVQGAN audio coding methods suitable the compression of Ambisonics audio scenes. Despite the present paper being anchored in audio coding, i.e. faithfully reproducing an original reference, we also draw inspiration from music generation. In addition to the modified model architecture, we also propose customized training methods. As a novel testing aspect for neural codecs, multichannel loudspeaker listening is used to evaluate if the method is suitable for multichannel neural compression in terms of spatial quality and compression efficiency.

Background

Neural Audio Coding and Music Generation

In this section, we draw commonalities between neural coding and music generation, and survey how both fields relate to the present study.

RVQGAN is a popular model for end-to-end neural audio coding [2, 3]. Methods besides RVQ have been proposed, namely diffusion models [4, 5] as well as transformer-based alternatives [6]. Neural codecs, despite having very good quality at particular low-bitrate range, have still suffer from many issues which impede real-life use. Complexity, especially on the decoder side, is an issue for real-life deployment on mobile devices. End-to-end neural methods can have problems scaling towards complete transparency, hybrid systems address this by using generative neural techniques to enhance the quality of a traditional codec [7].

The majority of neural audio coding research has been focused on single-channel aka mono audio compression. However, most modern audio content has more than one channel, making relevant the spatial and immersive aspects of audio. In recent decades, formats and reproduction systems utilizing more than two channels have emerged, starting from 5.1 surround and culminating with the immersive audio formats of recent years [8, 9, 10, 11, 12]. To our knowledge, the few attempts to apply neural coding methods to multichannel audio coding amount to processing a stereo input as dual-mono, thereby doubling the bitrate [13]. These works do not take advantage of the interchannel correlations present in the data.

For music generation, proposals have been oriented more toward stereo audio. [14, 15, 16] The reason is possibly that compression and faithful reproduction of the reference signal is viewed as most suitable for content with more learnable structure than in unconstrained music, such as speech (which is typically mono). However, music generation uses looser measures of "goodness" that are not applicable to the coding paradigm. Also, as the generators typically have no audio input, not many insights can be obtained for multichannel encoding.

While acknowledging the transparency scaling issues and taking inspiration from the music generation studies, our goal is to remain within the end-to-end audio coding paradigm. We, however, limit this study to ambient background material that is typically more constrained in structure, and is a less critical part of the audio content in mixed presentations of multiple audio elements [11]. Testing other material and model complexity reduction is left as future work.

Ambisonics

Ambisonics is a prominent audio representation widely utilized for spatial capture [1]. This section briefly summarizes the technical aspects of the format. Higher Order Ambisonics (HOA) leverages spherical harmonics coefficients for sound field encoding. These coefficients, in conjunction with spherical harmonics as basis functions, are summed to approximate the original recorded sound field. Mathematically, for a given radius r, angle θ, and wave number k, the pressure in a plane is represented by Equation 1, where $B^{\pm 1}_{mm}$ denote the HOA coefficients.

$$p(r,\theta) = B^{+1}_{00} J_0(kr) + \sum_{m=1}^{\infty} J_m(kr) B^{+1}_{mm} \sqrt{2} cos(m\theta) + \sum_{m=1}^{\infty} J_m(kr) B^{-1}_{mm} \sqrt{2} sin(m\theta) \tag{1}$$

To achieve faithful reproduction of the original sound field, theoretically, an infinite number of spherical harmonics would be needed. However, practical implementations necessitate truncation to a finite order M, where increased order yields enhanced fidelity. The resultant truncated multichannel HOA signal, known as B-format, demands $(M+1)^2$ Ambisonics channels to represent the sound field accurately. For instance, first-order Ambisonics contains 4 channels, second-order Ambisonics contains 9 channels, and third-order Ambisonics contains 16 channels. Decoding these channels to speaker signals involves applying a decoding matrix contingent upon the HOA B- format order and speaker positioning.

Despite its efficacy in capturing immersive audio environments, HOA poses challenges regarding data compression and transmission. The high channel count inherent in HOA entails substantial bitrate requirements for transmission. Thus, efficient compression schemes tailored to HOA data are imperative for practical deployment in various applications. Previous methods have relied on classical signal processing techniques involving linear combinations of lower-order channels or past channels [17] or using techniques such as SVD to reduce the dimensionality of the Ambisonics signal [18, 19, 20, 21, 22, 23, 24, 25, 26, 27]. These compression methods are able to reduce bitrates from tens of megabits/second to the order of hundreds of kilobits/second. However, as proposed in this work, it may be possible to further reduce bitrates using a data-driven neural approach.

One mitigating aspect for real-life application may be that the material typically captured with HOA is more truncated in terms of probability distribution than unconstrained general audio. Modeling a general audio domain such as "all music" is a hard problem, and many neural coding systems opt to test only a truncated distribution like speech in order to get good quality. We propose that the typical use of HOA in transmission systems such as [11] would also often involve a truncated distribution: spatial capture is often applied to background ambience and overall scene of e.g. alive event. Also, in modern transmission and content creation, the background can be complemented with separately coded discrete elements [8, 11]. Therefore, real-life application of the neural methods for scene-based audio can benefit from the low rates offered by data-driven compression, and can be viable even without scaling to high-quality transparency.

Methods

Model Architecture

The basic idea of the present multichannel architecture extension can be summarized as simply increasing the channel count of the first and last layer or RVQGAN convolutional layers to match the number of audio channels. In the case of 3rd-order Ambisonics material, this channel count is 16. With such a change, it is possible to keep the dimensionality, and therefore the compression efficiency of the model bottleneck. Besides straightforward architectural changes, this implies non-trivial changes to the typical loss functions (see next Section).

The standard convolution layer used in DNNs [28] utilizes multiple channels and kernels, which can be also thought of as FIR filters. Time-domain PCM audio signals are typically treated as 1-D time series signals, i.e. the kernels are one-dimensional. In this case, the layer output value with input size (C_{in}, L) and output (C_{out}, L_{out}) is

$$\text{out}(C_{out}) = \text{bias}(C_{out}) + \sum_{k=0}^{C_{in}-1} \text{kernel}(C_{out}, k) * \text{input}(k) \tag{2}$$

where * is the valid cross-correlation operator, C denotes a number of channels and L the length of the signal sequence. By using the standard arithmetic, the operation each channel of the original audio is processed with their own dedicated kernel tailored to the specifics of that signal, and the results are summed to the next layer signal. The kernel filters are optimized to both find the inter-channel structure, and to best represent the linear combination of channels in the output sequences. In previous neural codecs with monaural input, the first layer rather only performs an upmixing operation.

In addition to a brute-force solution of running a monaural model separately in each channel, an alternative method could be to capture the inter-channel structure via two-dimensional kernels (as in 2-D convolutional layers). However, this would increase the computational complexity, and introduce many separate notions of channels in the model, which are arguably more difficult to interpret.

Justifications for the present solution of using convolutional arithmetic to handle inter-channel aspects of the audio can be found in perceptual spatial hearing studies [29]. Much of auditory localization can be explained by interaural time- and level differences, and the overall spatial envolopment is tied to interaural correlation [29]. In non-neural codecs, these aspects can be represented by preserving the inter-channel covariance structure [9, 10]. Also, spatial attributes can be quantized heavily, and are typically assigned less bitrate compared to waveform coding, and can be more generative in nature. We will discuss the signal covariance structure more in the next Section in relation to loss functions.

Loss Functions

The original Descript RVQGAN model was trained with the combination of adversarial-, feature matching-, VQ codebook-, and reconstruction losses [3]. Since our model output is multichannel, we adjust the relevant losses to operate on each channel independently and finally take the expected value over channels as the final loss. For the adversarial discriminator, this requires to change the multi-scale (MSD) and multi-period and waveform discriminators (MPD), as well as the multi-resolution spectrogram discriminators (MRSD) to produce multichannel output.

We found that despite not looking at the channel interactions specifically, the previous paradigm already produces a good spatial impression for the present HOA content. For fine-tuning of spatial quality, we consider a loss for the mismatch of the interchannel covariance structure, which has been found to be strongly related to perceived interaural coherence, and therefore an important descriptor of perceptual spatial impression or the rendered output [29, 30, 31]. Attributes this loss aims to preserve are e.g. envelopment and diffuseness of the material. For content where these are important to emphasize, the covariance loss is a simple way to do that. It is also more general than the specialized loss functions proposed for stereo audio [15, 32].

The covariance loss is calculated as the expected L1 loss between the normalized channel-wise covariance matrices of the original and the model f_θ reconstruction time-domain signals:

$$L_{cov} = \frac{1}{2}\sum_{i=0}^{n}\sum_{j=0}^{n} || \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}} - \frac{\hat{C}_{ij}}{\sqrt{\hat{C}_{ii}\hat{C}_{jj}}} ||, \tag{3}$$

where $C = \text{cov}(\mathbf{x})$ and $\hat{C} = \text{cov}(f_\theta(\mathbf{x}))$ are the covariance matrices of the multichannel time-domain input and model output signal of n channels, respectively. Each element of the normalized matrix is then the Pearson correlation coefficient between channels i and j.

It has been found that such preservation of the convariance structure between channels is a useful target in spatial coding systems, and also account for possible linear rendering of the output afterward [9, 10, 33]. We evaluate the covariance broadband, but it is also possible to have the measure be frequency-dependent. We also experimented using non-normalized covariance instead of correlation coefficient, but found the latter more appealing for numerical stability, and the fact that the other losses seemed to also account for channel energies and level differences.

Transfer Learning from Single-Channel Models

We propose applying transfer learning to the multichannel model from previously trained mono model weights. In input and output layers, where the number of channels is increased, we replicate the original convolutional weights of the original channel to the multiple channels of the new model identically. Thus, all input and output channels of the multichannel model are processed with identical convolutional kernels at the start of the training, which then start to evolve as needed.

In practice, we construct the new model and copy the appropriate parameters from the state dictionary of the published Descript monaural model [3]. In addition to helping with RVQGAN training time consumption, transfer learning also helps if the fine-tuning stage is performed on less powerful hardware and smaller batch size. Fig. 1 illustrates the benefit of the proposed transfer learning in both faster conversion, and final error with reasonable amount of steps.

In our experiments, we do not apply transfer learning the adversarial discriminator model, only randomly initialize it as in the original work [3].

Experiment

Immersive audio material available for data-driven modeling is not as abundant as traditional formats. In this paper, we have utilized the EigenScape database [34]. It consists of eight acoustic scenes recorded spatially in 4th-order Ambisonics that are ambient in nature, but do sometimes include e.g. people speaking. We omit the additional channels and process the signals to obtain 16-channel, 3rd-order inputs at 16-bit / 44100 Hz resolution. For cross-validation, we utilize 7/8 of the samples on each scene for model training, and the keep the rest separate for validation and subjective testing. A multiple-round alternative of keeping each of the 8 individual

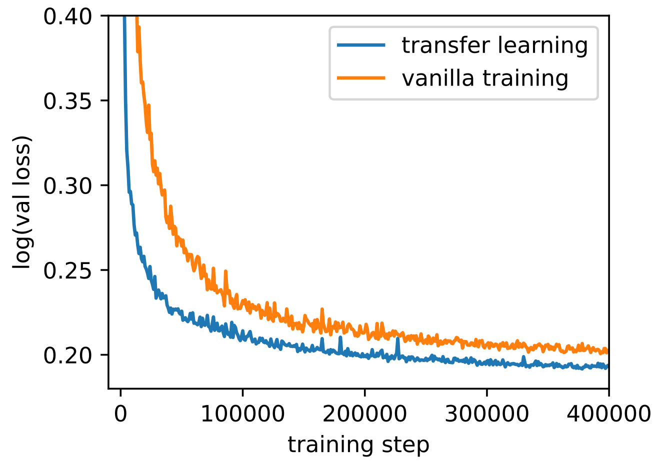

Figure 1: Example comparison between proposed transfer learning utilizing pre-trained mono model weights, and vanilla random initialization. Plot shows the development of the reconstruction multi-scale mel validation loss [3].

scenes for validation, and testing them in turn was unfortunately not feasible due to subjective testing being used. However, as we use transfer learning from a model trained with generic audio, the resulting model should retain some level of baseline generality.

The model implementation was built on the Descript codebase [3] with the aforementioned multichannel extensions. The only other change was to omit the output Tanh-nonlinearity of the decoder. The training procedure remains largely the same as in the original work, while we utilize the transfer learning scheme proposed in the previous Section . We use weighting of 1.0 for the covariance loss. The other loss weightings were kept as is (15.0 for the multi-scale mel loss, 2.0 for the feature matching loss, 1.0 for the adversarial loss and 1.0, 0.25 for the codebook and commitment losses, respectively). To illustrate a simple fine tuning procedure results, we trained for a modest number of steps (400000) and batch size (24) of 5-second samples.

In order to evaluate and compare the proposed Ambisonics neural codec, a MUSHRA [35] listening test was carried out. For quality reference, we process the HOA reference signals with Opus channel mapping family 3 coding intended for HOA content [36], at 160 kbit/s rate, which is on the lower quality region of the Opus HOA, to demonstrate how conventional coding methods with 10x bitrate compare. In total, four listening conditions were compared:

- Low anchor (LA) - 3.5kHz low-pass filtered version of the reference
- Opus' Channel Family Mapping 3 (OPUS) compression at 160 kbit/s
- Proposed neural ambisonics codec (PROP) at 16 kbit/s
- Hidden reference (H-REF)

Rendering the HOA signals to the 7.1.4 loudspeaker layout was done with the IAMF system [11] which utilizes a version of the EBU ADM Renderer [37]. Listening tests took place in a 7m (L) x 5.33m (W) x 3.05 (H) listening room equipped with a 7.1.4 playback system. The loudspeakers' and listener's positions were based on [38]. Loudspeakers were level-matched at the listener's position and meet the ITU-R specification [39].

Eight listeners completed the MUSHRA test. Participants were told to compare the four listening conditions to the known reference, and to rate the overall sound quality on the MUSHRA 100-point scale, while focusing on the audio quality and spatial impression correctness versus the hidden reference.

Results and Discussion

As can be seen from Fig. 2, the proposed neural codec operating on 16 kbit/s is able to achieve "good" quality on the MUSHRA scale, and outperform a traditional method with 10x less compression rate. While the traditional channel family mapping Opus method starts to break down at 160 kbit/s, and its quality would improve at higher rates, we opted to illustrate the quality potential of the neural methods rather than find the precise rate where the two methods would match.

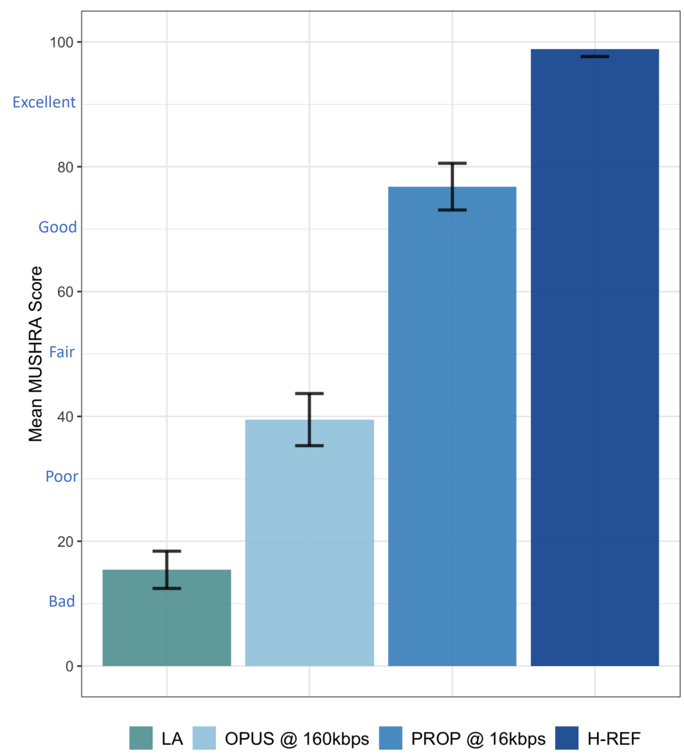

Figure 2: MUSHRA score mean and 95% confidence interval over 8 tracks and 8 listeners, with 7.1.4 immersive loudspeaker listening.

The audio material used here is from a truncated distribution of ambient scenes, and also limited in amount in order to illustrate a typical use case, as discussed in the Ambisonics-section earlier. However, the use of transfer learning from a generic model arguably retains some level of baseline generality. Also, the listening test conditions did enable critical evaluation of the sound quality similarly as in typical immersive audio testing. We also did not include in the tests any mixed presentations with additional audio elements [11], so that the HOA coding quality would be more forgiving.

Despite the promising results, the main caveats of the present method, and most neural codecs remain 1) the lack of scaling to different bitrates without retraining or model modification, and 2) the model computational complexity. For these, future work will hopefully provide solutions.

Acknowledgements

The authors would like to thank Samsung's US Audio Lab staff and especially Ema Souza Blanes for helping with the listening tests.

References

[1] Jerome Daniel, Sebastien Moreau, and Rozenn Nicol, "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging," in *Audio Engineering Society Convention 114*, March 2003.

[2] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[3] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-fidelity audio compression with improved RVQGAN," 2023, arXiv 2306.06546.

[4] Zhengpu Zhang, Jianyuan Feng, Yongjian Mao, Yehang Zhu, Junjie Shi, Xuzhou Ye, Shilei Liu, Derong Liu, and Chuanzeng Huang, "High-fidelity diffusion-based audio codec," *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 344–348, 2024.

[5] Robin San Roman, Yossi Adi, Antoine Deleforge, Romain Serizel, Gabriel Synnaeve, and Alexandre D'efossez, "From discrete tokens to high-fidelity audio using multi-band diffusion," 2023, arXiv 2308.02560.

[6] Yuzhe Gu and Enmao Diao, "ESC: Efficient speech coding with cross-scale residual vector quantized transformers," 2024, arXiv 2404.19441.

[7] Pedro J. Villasana T., Lars Villemoes, Janusz Klejsa, and Per Hedelin, "Audio decoding by inverse problem solving," 2024, arXiv 2409.07858.

[8] Charles Robinson, Nicholas Tsingos, and Shripal Mehta, "Scalable format and tools to extend the possibilities of cinema audio," *SMPTE Motion Imaging Journal*, vol. 121, no. 8, November 2012.

[9] "Digital Audio Compression (AC-4) Standard," Standard, European Telecommunications Standards Institute, Geneva, CH, 2018.

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

[11] "Immersive Audio Model and Formats," `https://aomedia.org/iamf/`, 2024.

[12] "Advanced sound system for programme production," Standard, International Telecommunication Union, Geneva, CH, 2022.

[13] Alexandre D'efossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," 2022, arXiv 2210.13438.

[14] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, "Simple and controllable music generation," 2024, arXiv 2306.05284.

[15] Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons, "Fast timing-conditioned latent audio diffusion," 2024, arXiv 2402.04825.

[16] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons, "Long-form music generation with latent diffusion," 2024, arXiv 2404.10301.

[17] Erik Hellerud, Audun Solvang, and U. Peter Svensson, "Spatial redundancy in higher order ambisonics and its use for lowdelay lossless compression," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 269–272.

[18] Sina Zamani, Tejaswi Nanjundaswamy, and Kenneth Rose, "Frequency domain singular value decomposition for efficient spatial audio coding," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 126–130.

[19] Sina Zamani and Kenneth Rose, "Spatial audio coding with backward-adaptive singular value decomposition," in *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

[20] Sina Zamani and Kenneth Rose, "Spatial audio coding without recourse to background signal compression," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 720–724.

[21] Pierre Mahé, Stephane Ragot, and Sylvain Marchand, "First-order ambisonic coding with quaternion-based interpolation of PCA rotation matrices," in *EAA Spatial Audio Signal Processing Symposium*, Paris, France, Sept. 2019, pp. 7–12.

[22] Jiahao Xu, Yadong Niu, Xihong Wu, and Tianshu Qu, "Higher order ambisonics compression method based on independent component analysis," in *Audio Engineering Society Convention 150*. Audio Engineering Society, 2021.

[23] Mahmoud Namazi, Ahmed Elshafiy, and Kenneth Rose, "Spatial audio compression with adaptive singular value decomposition using reconstructed frames," in *Audio Engineering Society Conference: 2022 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2022.

[24] Mahmoud Namazi, Ahmed Elshafiy, and Kenneth Rose, "On ultra low-delay compression of higher order ambisonics signals," in *2024 Data Compression Conference (DCC)*, 2024, pp. 512–521.

[25] Mahmoud Namazi and Kenneth Rose, "Ultra-low delay lossless compression of higher order ambisonics," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 791–795.

[26] Christoph Hold, Leo McCormack, Archontis Politis, and Ville Pulkki, "Perceptually-motivated spatial audio codec for higher-order ambisonics compression," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1121–1125.

[27] Mahmoud Namazi, *Advancements in Higher Order Ambisonics Compression and Loss Concealment Techniques*, Ph.D. thesis, UC Santa Barbara, 2024.

[28] Vincent Dumoulin and Francesco Visin, "A guide to convolution arithmetic for deep learning," 2018, arXiv 1603.07285.

[29] Jens Blauert, "Spatial hearing: The psychophysics of human sound localization (revised edition)," 1997.

[30] Jens Blauert and Werner Lindemann, "Spatial mapping of intracranial auditory events for various degrees of interaural coherence," *The Journal of the Acoustical Society of America*, vol. 79, no. 3, pp. 806–813, March 1986.

[31] Robert Luke, Hamish Innes-Brown, Jaime A Undurraga, and David Mcalpine, "Human cortical processing of interaural coherence," *iScience*, vol. 5, no. 25, March 2022.

[32] Christian J. Steinmetz, Jordi Pons, Santiago Pascual, and Joan Serrà, "Automatic multitrack mixing with a differentiable mixing console of neural audio effects," 2020, arXiv 2010.10291.

[33] Jonas Engdegård 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.

[34] Marc Ciufo Green and Damian Murphy, "Eigenscape," `https://doi.org/10.5281/zenodo.1012809`, 2017.

[35] ITU-R, "Method for the subjective assessment of intermediate quality level of audio systems," Recommendation BS.1534-2, International Telecommunication Union, Geneva, 2014.

[36] Jan Skoglund and Michael Graczyk, "Ambisonics in an Ogg Opus container," `https://www.rfc-editor.org/rfc/rfc8486.html`, 2018.

[37] ITU-R, "Audio definition model renderer for advanced sound systems," Recommendation BS.2127-1, International Telecommunication Union, Geneva, 2023.

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

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