Audio Engineering Society

Conference Paper 3

Presented at the AES International Conference on Machine Learning and Artificial Intelligence for Audio 2025 September 8–10, London, UK

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.

Compression of Scene-Based Higher Order Ambisonics with Multichannel RVQGAN

Toni Hirvonen and Mahmoud Namazi

Samsung Research America

Correspondence should be addressed to Toni Hirvonen (t.hirvonen@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 multichannel input. We also propose a loss function which accounts 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 ambient scene-based, third-order (16-channel) Ambisonics content with good quality at 16 kbps when trained and tested on the EigenScape database. The model presented is the first neural codec dedicated to immersive audio to the authors' knowledge and has potential applications for learning other types of content and multichannel formats.

1 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 discovering new, efficient coding methods. This is especially true in the case of Higher Order Ambisonics, a prominent surround sound encoding method [1] which has in recent times been a popular solution for immersive capture and reproduction.

At the same time, the rise of data-driven processing methods and available computational power make it feasible to explore methods beyond conventional audio coding. However, neural coding of multichannel- and immersive audio has received relatively little attention. This is partly due to lack of material available to train and test such models.

This paper presents a simple, yet general and efficient, multichannel extension to the popular RVQGAN audio coding methods suitable for the compression of Ambisonic audio. Despite being anchored in audio coding, i.e. faithfully reproducing an original reference, we also draw inspiration from music generation. The training and tests conducted with scene-based material illustrate the feasibility of the proposed method, and are aimed to inspire further applications in the future.

Our contributions include: The model presented in this paper is the first neural codec dedicated to immersive audio to the authors' knowledge. 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 in terms of spatial quality and compression efficiency.

2 Background

2.1 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 that includes a data-compressing encoder and an expanding decoder [2, 3]. Diffusion models [4, 5] as well as transformer-based alternatives [6] have also been proposed. All neural codecs, despite having very good quality at particular low-bitrate range, have issues that impede real-life use. Computational complexity, especially on the decoder side, is a predicament for real-life deployment on many devices. End-to-end neural methods can also have problems scaling towards complete transparency, and hybrid systems aim to 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].

Some attempt 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 that are often present. As an exception, Kleijn et al. [14] generated multichannel (in the experiment only stereo) as a mono channel + spatial metadata, and used generative methods for the spatial information. Two other works [15, 16] propose neural codecs which are capable of binaural speech and general speech coding for one/two speakers, respectively. However, as opposed to being more generative in nature, the present study aims to replicate and compare to the original multichannel spatial impression of non-speech material.

In the field of music generation, stereo audio output seems more common than in coding [17, 18, 19]. One possible reason is that compression and faithful reproduction of the reference signal is viewed as most suitable for content with more learnable structure than in unconstrained music, including speech (which is typically mono). Applying music generation modeling techniques to audio coding and obtaining benefits in terms of multichannel handling is not trivial. Music generation uses more relaxed measures of goodness that are not applicable to the coding paradigm. Also, as music generators typically have no audio input, not many insights can be obtained for encoding. Nevertheless, neural audio decoding and (conditional) audio generation are closely related problems.

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. Nevertheless, generative and compressive neural methods can be relevant for the coding of spatial aspects even when striving toward perceptual transparency, as discussed in Sec. 3.

We limit the practical testing of this study to ambient, scene-based background material that is typically relatively constrained in structure. Often such background element is a less critical part of the content in e.g. mixed presentations of multiple audio elements [11], and could benefit from efficient multichannel compression due to its high channel count.

However, we also postulate that the application of the proposed neural architecture is not limited to ambient background material, in cases where other types of training material is available. Such material may include more discrete moving objects, or other types of audio, such as music. As our general model is a universal approximator [20], this notion is neither unrealistic nor surprising if the model size is not limited. Nonetheless, the key contribution of this paper is to illustrate the spatial coding efficiency potential of the convolutional arithmetic, as proposed by our testing with scene-based material.

2.2 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 J_n is the n-th order Bessel function and $B_{mm}^{\pm 1}$ denotes the HOA coefficients.

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

To achieve faithful reproduction of the original sound field, an infinite number of spherical harmonics would theoretically 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 [21] or using techniques such as SVD to reduce the dimensionality of the Ambisonics signal [22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32]. 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 applications may be that the material typically captured with HOA is more truncated in terms of probability distribution than unconstrained general audio. Modeling a general 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 argue that common uses of HOA would also involve a truncated distribution: spatial capture is often applied to background ambience and/or the overall scene of e.g. a live event in transmission systems such as [11].

In the former case, the HOA element can be thought as "bed" [33], where it can be complemented with separately coded discrete elements such as audio objects [34, 8, 11]. In the latter case even without the additional audio elements, the scene-based HOA capture can hypothetically be distinguished from generic music etc. material by having added ambience and reverberation, and less perceptually discrete sounds. Both of these use cases also apply to synthetic HOA in VR/AR, where again, the discrete virtual elements are more naturally handled by audio objects [34].

While the above implies that a major share of HOA content can be less difficult to compress in the traditional audio coding sense, for spatial fidelity it desirable to use at least third-order Ambisonics [35]. This in turn implies the need for efficient channel-wise compression of HOA content with 16 or more channels. Therefore, real-life application of neural methods for scene-, and bed-based audio can benefit from the low rates offered by data-driven compression, and can be viable even without scaling to high-quality transparency. To illustrate, we test our proposed method for scene-based HOA coding with commonly available material.

3 Methods

3.1 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 convolutional layer of RVQGAN to match the number of audio channels. In the case of third-order Ambisonics material, this channel count is 16. If desirable, it is possible to keep the dimensionality, and therefore the compression efficiency of the remaining model layers. The change in the first model layer is illustrated in Fig. 1. Besides this architechture modification, the present proposal implies non-trivial changes to the typical loss functions, which is discussed more in Sec. 3.3.

More formally, the standard convolutional layer used in deep neural networks (DNNs) [36] utilizes multiple channels and kernels, which can be also thought of

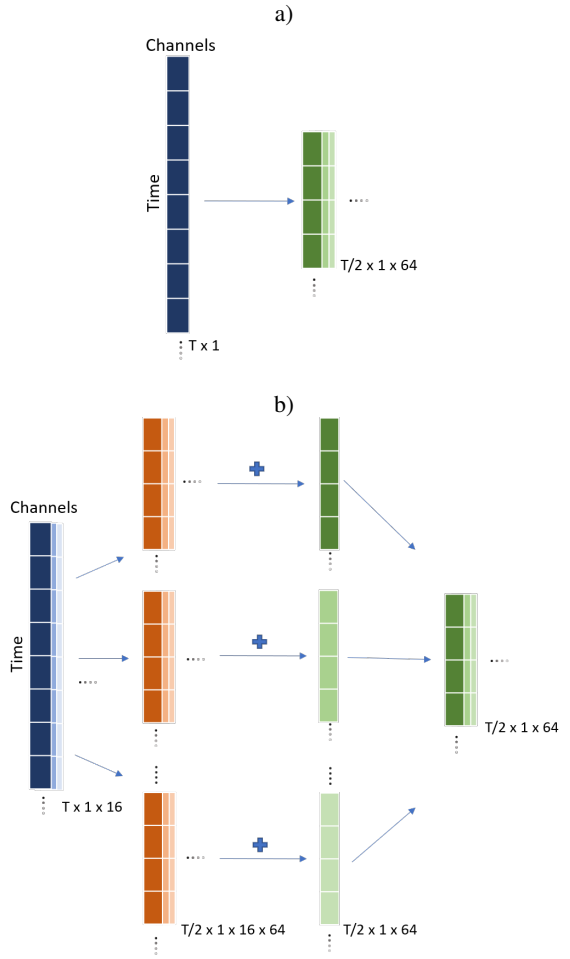


Fig. 1: Visualization of the first layer in a) the single-channel Descript [3] and b) the proposed codec for third-order Ambisonics. The dimensions of the tensors are indicated at each processing stage.

as FIR filters. Time-domain PCM audio signals are typically treated as 1-D time series signals, i.e. the kernels are one-dimensional (as in Conv1D layers). In this case, the layer output for the output channel c_{out} is

$$\begin{aligned}\text{output}(c_{out}) &= \text{bias}(c_{out}) \\ &+ \sum_{k=0}^{n_{in}-1} \text{kernel}(c_{out},k)*\text{input}(k)\end{aligned} \tag{2}$$

where * is the valid cross-correlation operator, and k denotes the input channel index. While the sequence index is omitted in Eq. 2, the layer input and output have channel count and sequence length of $[n_{in}, l_{in}]$ and $[n_{out}, l_{out}]$, respectively. Each input channel 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.

As seen from Fig. 1 and Eq. 2, the first layer only performs an upmixing operation in traditional neural codecs with a single-channel audio input. It is not obvious how a multichannel audio input should be processed for maximum efficiency. In addition to a brute-force solution of running a mono model separately in each audio channel, an alternative method could be to capture the inter-channel structure via two-dimensional kernels (as in Conv2D 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. The present solution is rather using Conv1D layers and the standard arithmetic of Eq. 2, while increasing the input and output channels counts to match those of the source audio.

3.2 Perceptual Considerations

Justifications for the method of Sec. 3.1 being able to handle the inter-channel aspects of audio can be argued when considering human spatial perception [37]. Much of auditory localization can be explained by interaural time- and level differences, and the overall spatial envelopment is tied to interaural correlation and coherence [37]. A simple model to characterize the relationship between the audible ear signals, to which the above psychoacoustic results apply, and the codec input content is to consider them being related by a linear transformation.

If $X \in \mathbb{C}^{n\times l}$ is a (frequency-domain) multichannel input signal matrix of n channels of length l, the ear signals can be modeled as $E = AX$, where $A \in \mathbb{C}^{2\times n}$ is a (time- and frequency dependent) transform. Processes that can be reasonably approximated by this model include typical gain-based loudspeaker rendering systems for various content [38], and their combination with loudspeaker room impulse responses (RIRs). RIRs are

generally considered to represent linear time-invariant (LTI) systems. Similarly, binaural headphone renderers also typically operate as a combination of virtual loudspeakers, decorrelators and LTI head-related transfer functions [39].

Assuming such a simple LTI model, it can be seen that preserving the covariance structure of the input multichannel signal (with a reasonable time- and frequency resolution) would also result in approximating the audible interaural covariance:

$$R_E = E\overline{E} \approx AX\overline{(AX)} = AX\overline{XA} = AR_X\overline{A}, \quad (3)$$

where R_E and R_X represent the covariance matrices of the listener-audible-, and the original-format audio, and $\overline{\ }$ indicates the conjugate transpose. It is simple to see that the covariance matrix encompasses information about the inter-channel temporal and level differences, as well as correlations.

Accordingly, it has been found that such preservation of the original audio convariance structure is a highly useful target in state-of-the-art spatial coding systems, and it also accounts for possible rendering of the output afterward [9, 10, 34]. Traditional codecs also show that spatial attribute parameters can be quantized heavily, and are typically assigned less bitrate compared to waveform coding, and can be more generative in nature. Signal covariance is also discussed further in Sec. 3.3 in relation to loss functions.

With the above reasoning that preserving inter-channel properties of the original format audio constitutes in good perceptual spatial quality, the question becomes can the neural network convolution arithmetic in Eq. 2 capture and represent R_X? While it is generally difficult to derive any formal proofs related to DNNs, simple thought experiments e.g. about the operation of the model first layer seem to corroborate this. The sum of the channel-dependent kernel filter values can represent different values of inter-channel temporal (phase) and level differences, as well as correlations, given suitably well learned kernels. This informal reasoning, along with the universal approximator theorem [20] are postulated to warrant the present research.

3.3 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 changes to the multi-scale (MSD) and multi-period and waveform discriminators (MPD), as well as the multi-resolution spectrogram discriminators (MRSD) to produce a 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 the spatial quality, we consider a loss for the mismatch of the interchannel covariance structure, which was in Sec. 3.1 connected to the perceived interaural coherence, and therefore an important descriptor of perceptual spatial impression of the rendered output [37, 40, 41]. Attributes this loss aims to preserve are e.g. envelopment and diffuseness of the material. For content where these are important, the covariance loss is a simple way to emphasize them. It is also more general than the specialized loss functions proposed for stereo audio [18, 42].

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:

$$L_{cov} = \frac{1}{2}\sum_{i=0}^{n}\sum_{j=0}^{n}\left|\left|\frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}} - \frac{\hat{C}_{ij}}{\sqrt{\hat{C}_{ii}\hat{C}_{jj}}}\right|\right|, \quad (4)$$

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

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.

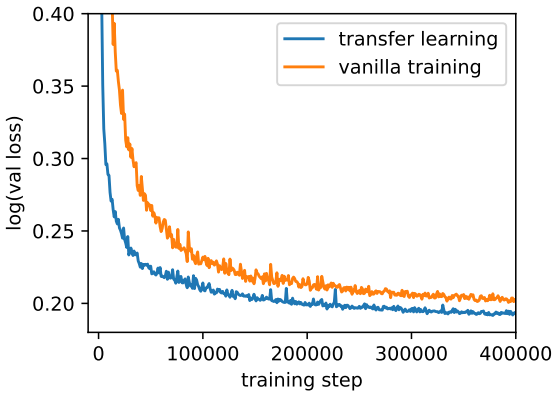


Fig. 2: Example comparison between proposed transfer learning utilizing pre-trained single-channel model weights, and vanilla random initialization. Plot shows the development of the reconstruction multi-scale mel validation loss [3].

3.4 Transfer Learning from Single-Channel Models

We propose applying transfer learning to the multichannel model from previously trained single-channel 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 evolve as needed.

In practice, we construct the new model and copy the appropriate parameters from the state dictionary of the published 16 kbps version of Descript single-channel 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. 2 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 to the adversarial discriminator model, it is only randomly initialized as in the original work [3].

4 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 [43]. It consists of eight acoustic scenes recorded spatially in fourth-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, third-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 scenes for validation, and testing them in turn was unfortunately not feasible due to subjective testing being used.

The model implementation was built on the Descript codebase [3] with the aforementioned multichannel extensions to the architecture and loss calculation (see Sec. 3). The only other changes were to omit the output Tanh-nonlinearity, and trivial modifications to make the code dependencies work with 16-channel audio. The training procedure remains largely the same as in the original work, with the exception of utilizing the transfer learning scheme proposed in Sec. 3.4. 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 (400,000) and batch size (24) of 5-second samples.

A MUSHRA [44] listening test was carried out. For quality reference, we utilize Opus channel mapping family 3 coding intended for HOA content [45], at 160 kbps 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 kbps
- Proposed neural codec (PROP) at 16 kbps
- Hidden reference (H-REF)

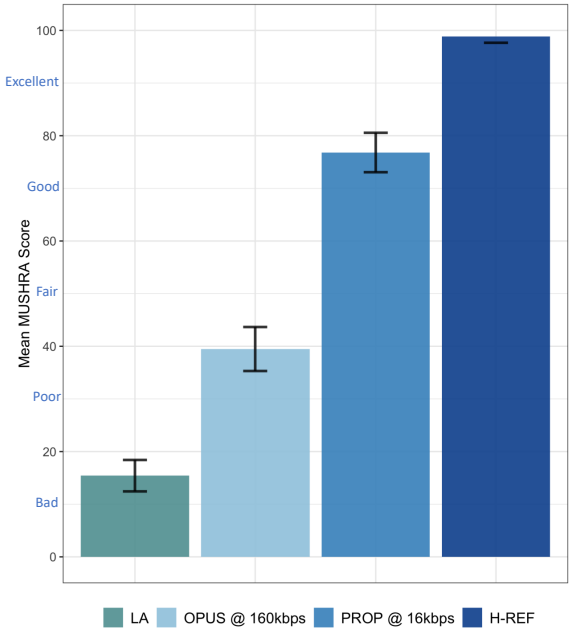


Fig. 3: MUSHRA score mean and 95% confidence interval over 8 tracks and 8 listeners, with 7.1.4 immersive loudspeaker listening.

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 [46]. 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 [47]. Loudspeakers were level-matched at the listener's position and meet the ITU-R specification [48].

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.

5 Results and Discussion

As can be seen from Fig. 3, the proposed neural codec operating on 16 kbps is able to achieve "good" quality on the MUSHRA scale, and outperform a traditional method with 10x less compression rate. The traditional Opus method [45] starts to break down at 160 kbps, and its quality would improve at higher rates. However, for good quality the Opus HOA codec rate needs typically to be considerably higher, around 32 kbps per channel [49]. Here we opted to illustrate the quality potential of the neural methods with a simple comparison.

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 Sec. 2.2. 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. Although the present material was ambient background in nature, we did not test mixed presentations with additional foreground audio elements [11], which would have made the HOA coding quality less critical.

It should be noted that informal listening with few samples of cinematic music outside the EigenScape database indicated that applying the proposed model trained only on scene-based material may not be as optimal for all content. Other content types or more generic models can also require an architecture with an increased bottleneck bitrate to achieve good quality. More general training and application of the model with e.g. musical and cinematic content is left as future work.

Although not tested here, the present modeling principle is also applicable to multichannel audio formats other than third-order Ambisonics (e.g. 5.1, 7.1, 7.1.4 etc.), as long as there is sufficient training data. 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.

References

[1] 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.

[2] Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M., "SoundStream: An End-to-End Neural Audio Codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, pp. 495–507, 2021.

[3] Kumar, R., Seetharaman, P., Luebs, A., Kumar, I., and Kumar, K., "High-Fidelity Audio Compression with Improved RVQGAN," 2023, arXiv 2306.06546.

[4] Zhang, Z. et al., "High-Fidelity Diffusion-Based Audio Codec," *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 344–348, 2024.

[5] Roman, R. S., Adi, Y., Deleforge, A., Serizel, R., Synnaeve, G., and D'efossez, A., "From Discrete Tokens to High-Fidelity Audio Using Multi-Band Diffusion," 2023, arXiv 2308.02560.

[6] Gu, Y. and Diao, E., "ESC: Efficient Speech Coding with Cross-Scale Residual Vector Quantized Transformers," 2024, arXiv 2404.19441.

[7] Pedro J. Villasana T., Villemoes, L., Klejsa, J., and Hedelin, P., "Audio Decoding by Inverse Problem Solving," 2024, arXiv 2409.07858.

[8] 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.

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

[10] 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.

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

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

[13] D'efossez, A., Copet, J., Synnaeve, G., and Adi, Y., "High Fidelity Neural Audio Compression," 2022, arXiv 2210.13438.

[14] Kleijn, W. B., Chinen, M., Lim, F. S. C., and Skoglund, J., "Multi-Channel Audio Signal Generation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[15] Xu, Z., Xu, Y., Kothapally, V., Wang, H., Yang, M., and Yu, D., "SPATIALCODEC: Neural Spatial Speech Coding," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1131–1135, 2024, doi:10.1109/ICASSP48485.2024.10445853.

[16] Ratnarajah, A., Zhang, S.-X., and Yu, D., "M3-audiodec: Multi-Channel Multi-Speaker Multi-Spatial Audio Codec," *arXiv preprint arXiv:2309.07416*, 2023.

[17] Copet, J. et al., "Simple and Controllable Music Generation," 2024, arXiv 2306.05284.

[18] Evans, Z., Carr, C., Taylor, J., Hawley, S. H., and Pons, J., "Fast Timing-Conditioned Latent Audio Diffusion," 2024, arXiv 2402.04825.

[19] Evans, Z., Parker, J. D., Carr, C., Zukowski, Z., Taylor, J., and Pons, J., "Long-Form Music Generation with Latent Diffusion," 2024, arXiv 2404.10301.

[20] Zhou, D.-X., "Universality of Deep Convolutional Neural Networks," *Applied and Computational Harmonic Analysis*, 48(2), pp. 787–794, 2020.

[21] Hellerud, E., Solvang, A., and Svensson, U. P., "Spatial redundancy in Higher Order Ambisonics and its use for lowdelay lossless compression," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 269–272, 2009.

[22] Zamani, S., Nanjundaswamy, T., and Rose, K., "Frequency Domain Singular Value Decomposition for Efficient Spatial Audio Coding," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 126–130, 2017.

[23] Zamani, S. and Rose, K., "Spatial Audio Coding with Backward-Adaptive Singular Value Decomposition," in *Audio Engineering Society Convention 145*, Audio Engineering Society, 2018.

[24] Zamani, S. and Rose, K., "Spatial Audio Coding Without Recourse to Background Signal Compression," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 720–724, 2019.

[25] Mahé, P., Ragot, S., and Marchand, S., "First-Order Ambisonic Coding with Quaternion-Based Interpolation of PCA Rotation Matrices," in *EAA Spatial Audio Signal Processing Symposium*, pp. 7–12, Paris, France, 2019, doi:10.25836/sasp.2019.19.

[26] Xu, J., Niu, Y., Wu, X., and Qu, T., "Higher order Ambisonics Compression Method Based on Independent Component Analysis," in *Audio Engineering Society Convention 150*, Audio Engineering Society, 2021.

[27] Namazi, M., Elshafiy, A., and Rose, K., "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.

[28] Namazi, M., Elshafiy, A., and Rose, K., "On Ultra Low-Delay Compression of Higher Order Ambisonics Signals," in *2024 Data Compression Conference (DCC)*, pp. 512–521, 2024.

[29] Namazi, M. and Rose, K., "Ultra-Low Delay Lossless Compression of Higher Order Ambisonics," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 791–795, 2024.

[30] Hold, C., McCormack, L., Politis, A., and Pulkki, V., "Perceptually-Motivated Spatial Audio Codec for Higher-Order Ambisonics Compression," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1121–1125, 2024.

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

[32] Namazi, M., Hirvonen, T., Bharitkar, S., and Souza-Blanes, E., "Pre- and Post-processing of Ambisonics Signals for Legacy Audio Codecs," in *AES 156th Convention*, Audio Engineering Society, 2024.

[33] Riedmiller, J., Mehta, S., Tsingos, N., and Boon, P., "Immersive and Personalized Audio: A Practical System for Enabling Interchange, Distribution, and Delivery of Next-Generation Audio Experiences," *SMPTE Motion Imaging J*, 124(5), pp. 1–23, 2015.

[34] 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.

[35] Deutsch, T., Vollmer, L. J., and Fels, J., "An Assessment of Spatial Fidelity of Higher-Order Ambisonics Reproduction Using the SCaLAr System," in *47th Anual Conference*, Deutsche Gesellschaft für Akustik e.V. (DEGA), 2021.

[36] Dumoulin, V. and Visin, F., "A Guide to Convolution Arithmetic for Deep Learning," 2018, arXiv 1603.07285.

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

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

[39] European Broadcasting Union, "Binaural EBU ADM Renderer (BEAR) for object-based sound over headphones," *European Broadcasting Union, Geneva, Switzerland, Tech. Rep*, 2023.

[40] Blauert, J. and Lindemann, W., "Spatial Mapping of Intracranial Auditory Events for Various Degrees of Interaural Coherence," *The Journal of the Acoustical Society of America*, 79(3), pp. 806–813, 1986.

[41] Luke, R., Innes-Brown, H., Undurraga, J. A., and Mcalpine, D., "Human Cortical Processing of Interaural Coherence," *iScience*, 5(25), 2022.

[42] Steinmetz, C. J., Pons, J., Pascual, S., and Serrà, J., "Automatic Multitrack Mixing with a Differentiable Mixing Console of Neural Audio Effects," 2020, arXiv 2010.10291.

[43] Green, M. C. and Murphy, D., "EigenScape," `https://doi.org/10.5281/zenodo.1012809`, 2017.

[44] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," Recommendation BS.1534-2, International Telecommunication Union, Geneva, 2014.

[45] Skoglund, J. and Graczyk, M., "Ambisonics in an Ogg Opus Container," `https://www.rfc-editor.org/rfc/rfc8486.html`, 2018.

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

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

[48] ITU-R, "Methods for the Subjective Assessment of Small Impairments in Audio Systems," Recommendation BS.1116-3, International Telecommunication Union, Geneva, 2015.

[49] Souza-Blanes, E., Tejeda-Ocampo, C., and Bharitkar, S., "A Closer Look on Bitrate Requirements for Opus with First Order Ambisonics and AllRAD rendering to 5.1 and 7.1.4," in *AES 153th Convention*, Audio Engineering Society, 2022.