Audio Engineering Society

Convention Express Paper 195

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

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Pre- and Post-Processing of Ambisonics Signals for Legacy Audio Codecs

Mahmoud Namazi[1,2], Toni Hirvonen[1], Sunil Bharitkar[1], and Ema Souza-Blanes[1]

[1] *Samsung Research America*
[2] *University of California, Santa Barbara*

Correspondence should be addressed to Mahmoud Namazi (mnamazi@ucsb.edu)

ABSTRACT

This work presents three variants on a core algorithm, to be used in conjunction with mono-channel legacy codecs, for the compression of Higher Order Ambisonics signals. The presented methods use spatial decorrelation and energy-compacting transforms, combined with energy-based bit allocation for decorrelated channels, in order to enhance performance. Depending on the variant, windowing may be applied to smoothen frame transition regions. Results show that the proposed methods outperform Opus' Channel Family Mapping 3 and are capable of good perceptual quality (MUSHRA > 80) at bitrates as low as 160 kb/s.

1 Introduction

Recent pushes in consumer immersive audio [1] and virtual reality [2] have revitalized interest in delivering improved spatial audio across diverse setups. Higher Order Ambisonics (HOA), a spatial audio format [3], has, in particular, drawn attention due to the desirable qualities it possesses, including the ability to easily render spatial audio to any speaker setup, as well as easily rotate the soundfield. Additionally, HOA is capable of being captured and stored at varying resolutions.

To craft immersive, interactive experiences for virtual and augmented reality, the encoding of 3D audio demands high-quality reconstruction despite the vast data volume, particularly challenging when HOA involves potentially up to 64 channels. At higher resolutions, however, HOA data can consume tens of megabits per second in terms of bandwidth. It is therefore necessary to construct new methods, tailored to HOA, capable of exploiting both the significant inter-channel and intra-channel correlations which exist.

Initial attempts at HOA compression identified trade-offs and redundancy exploitation methods critical for achieving better compression efficiency. For instance, one study [4] showed encoding HOA channels separately with the Advanced Audio Codec (AAC) high-

[2]Note: the first author was an intern at Samsung Research America during the completion of this work.

lighted the tradeoff between allocating higher bit rates for lower-order channels, enhancing sound quality in the sweet spot but at the expense of spatial resolution. Another study [5] focused on low-latency, lossless HOA signal compression, recognized substantial inter-channel correlation as a source for significant compression gains.

A large number of works have now, attempting to address the plethora of compression challenges and opportunities that come with compressing HOA data, been put forward [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. While all of these works present interesting contributions to the HOA compression literature, most suffer from the inability to use existing codecs without substantial adaptations.

Opus channel mapping family 3 [8] avoids this issue by projecting the Ambisonics signal into a pre-determined channel-based format and then employing joint channel compression. Psychoacoustic redundancies are exploited in order to maximally compress the signal.

In this work, we propose three novel variants of a pre-/post- processing method which all revolve around the common theme of applying a energy-compacting, decorrelating transform, followed by energy-based channel-by-channel bit allocation which can then be used by a legacy mono-channel codec for compression of each channel separately. First, in Section 2 we briefly cover the Higher Order Ambisonics methodology. Next, in Section 3, we describe the three methods in detail. Then, in Section 4, we compare the three variants and Opus' family channel mapping 3 through the use of MUSHRA listening tests. In Section 5, we discuss further improvements that may be gained through the use of robust SVD and future research directions, followed by conclusions in Section 6.

2 Background

Higher Order Ambisonics (HOA) is a prominent surround sound methodology [3] which leverages spherical harmonics coefficients for sound field encoding. These coefficients, in conjunction with spherical harmonics as basis functions, are amalgamated 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 $B_{mm}^{\pm 1}$ denote the HOA coefficients.

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

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 necessitates 4 channels, second-order ambisonics requires 9 channels, and third-order ambisonics demands 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.

3 Methods

3.1 SVD-4W

A block-level diagram that outlines the proposed SVD-4W pre-/post- processing method can be seen in Figure 1.

Consider the HOA data X which is an $L \times N$ matrix consisting of L time samples and N Ambisonics channels.

We first decompose X into frames X_m, indexed by m, of length K with an overlapping region of length W with adjacent frames. A square-root Hann window is applied to the overlapping regions before further processing. Following this, singular value decomposition (SVD) is applied on a frame-by-frame basis. This yields

$$X_m = U_m S_m {V_m}^T$$

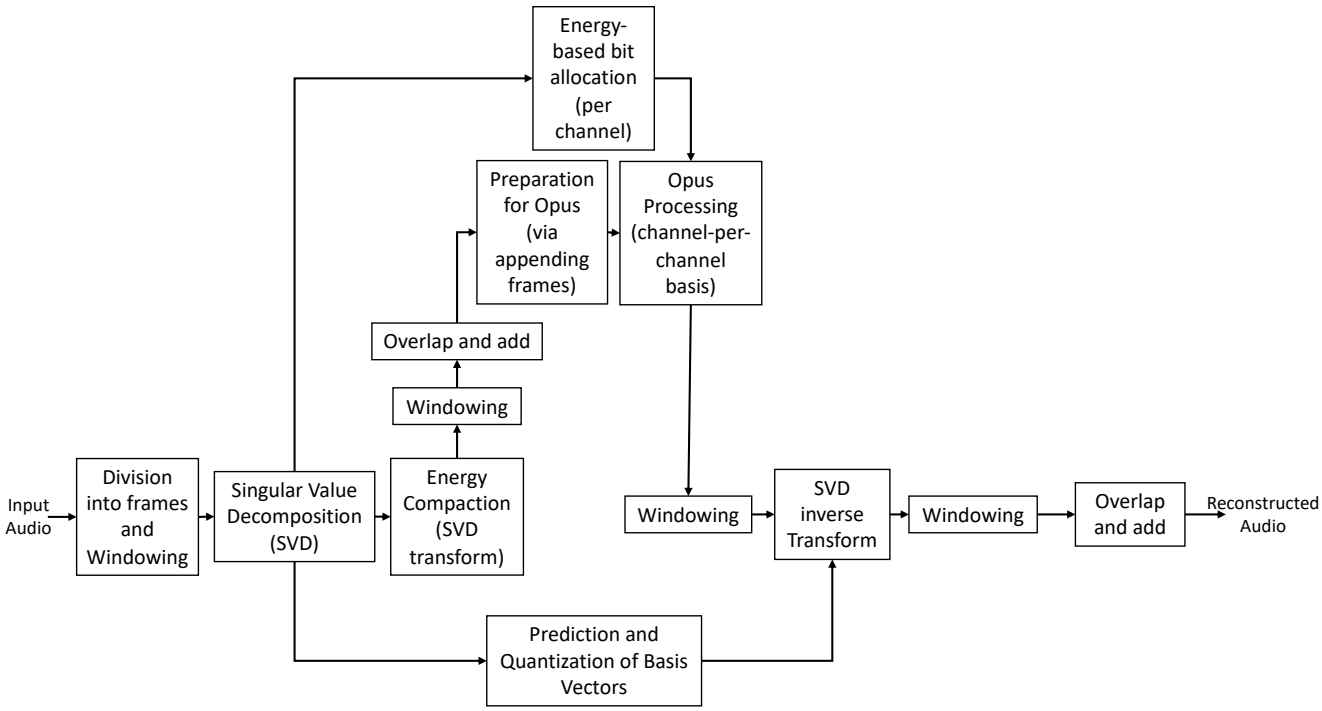


Fig. 1: Outline of proposed SVD-4W pre-/post- processing method.

The basis vectors, V_m, contain the directional information associated with the frame. We truncate V_m to its first z basis vectors, which followed by linear prediction and quantization transmit $\hat{V}_m$ to the decoder, on a frame-by-frame basis, as side information. Following this, we use $\hat{V}_m$, which will be recovered at the decoder, to apply a PCA-like transform, yielding

$$Y_m = X_m \hat{V}_m (\hat{V}_m{}^T \hat{V}_m)^{-1}$$

Note: the purpose of the inverse term is to renormalize the quantized basis vectors (in order to maintain unitarity). The square-root Hann window is then applied to the overlap regions of Y_m.

Next, Y_m is coded using a mono-channel legacy codec, such as Opus, where bitrate is allocated to channels proportionately based on the relative weight of the associated singular values in S_m.

At the decoder, Y_m is first decoded yielding $\tilde{Y}_m$. Then, the square-root Hann window is applied again at the overlap regions of $\tilde{Y}_m$. An approximation of X_m is recovered through

$$\tilde{X}_m = \tilde{Y}_m \hat{V}_m{}^T$$

Finally, the square-root Hann window is applied again at the overlap regions before the overlap-add method is used across the overlap regions to obtain the final approximation for X.

3.2 SVD-2W

A block-level diagram that outlines the proposed SVD-2W pre-/post- processing method can be seen in Figure 2.

This method is fairly similar to the SVD-4W method described in section 3.1 and was developed first. The difference between the two methods is that in SVD-2W, windowing is only applied twice. Once after the division into frames, and then a final time before overlap and add is applied, after the SVD inverse transform.

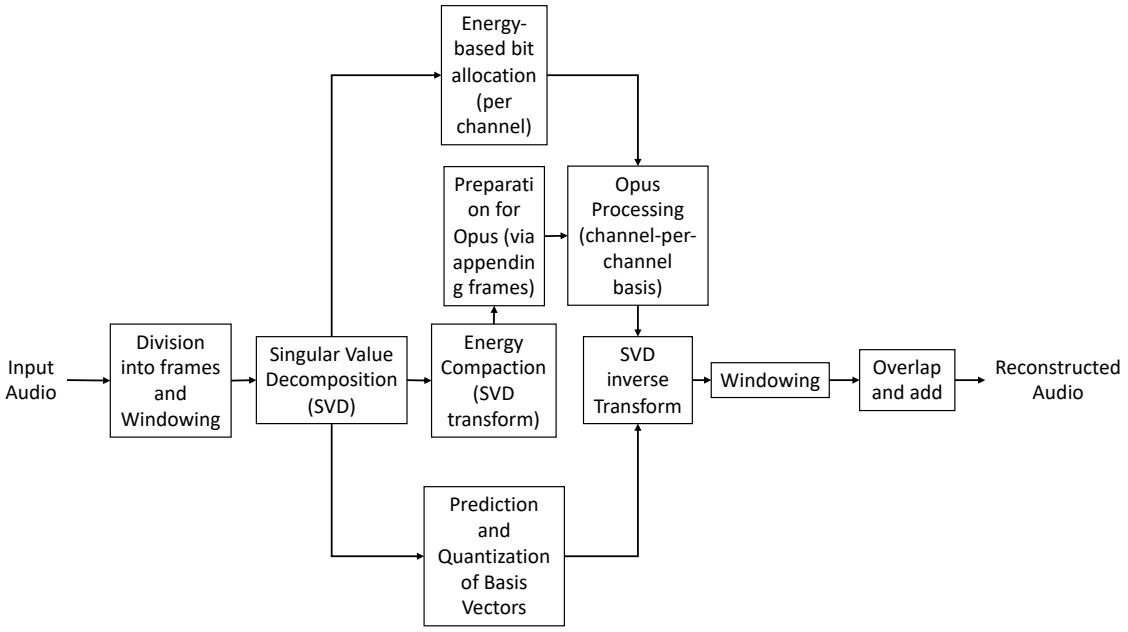


Fig. 2: Outline of proposed SVD-2W pre-/post- processing method.

SVD-2W is motivated by the desire to avoid applying the Hann window in the SVD domain, where it may cause issues due to the different sets of basis vectors used by adjacent frames. On the other hand, SVD-2W disrupts the synchronicity with respect to the original frame boundaries, where as SVD-4W maintains the original frame boundaries. This can be seen in Figure 3

However, as illustrated in Figure 3, SVD-4W desyncs the audio in terms of the frame boundaries, before the frames are fed to the legacy codec, compared to the original audio and may therefore be less desirable for particular applications.

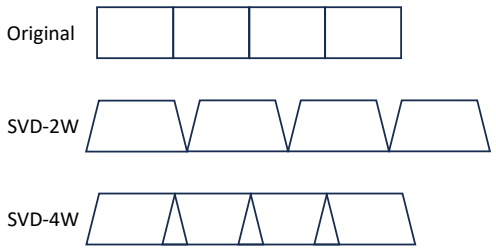


Fig. 3: Difference in frame boundaries between SVD-2W and SVD-4W methods.

3.3 INCR-SVD

A block-level diagram that outlines the proposed INCR-SVD pre-/post- processing method can be seen in Figure 4.

This method utilizes the incremental SVD method as described in [16], without any subdivision into frames, on X to yield

$$X \approx USV^T$$

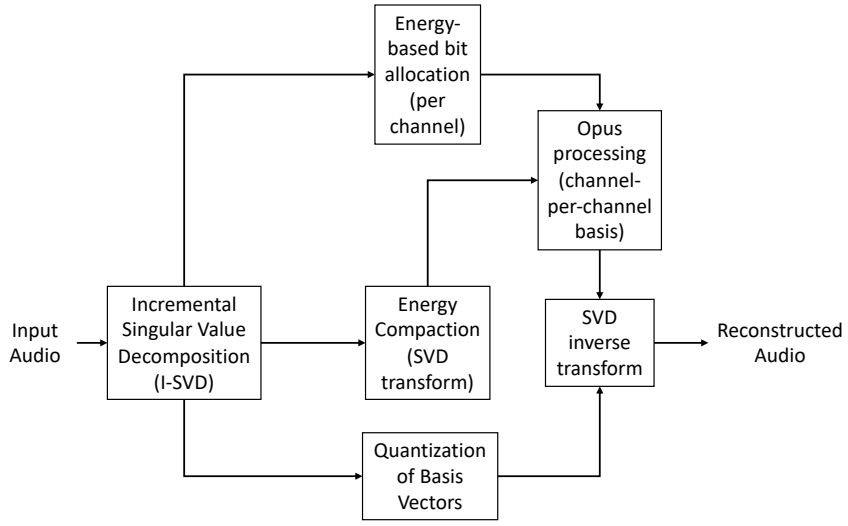


Fig. 4: Outline of proposed INCR-SVD pre-/post- processing method.

where U, S, and V are approximations for the full SVD decomposition of X. Next, V is truncated to its first z singular values, giving $\hat{V}$, which can be sent to the decoder at full precision incurring a minimal side information cost, and Y is generated

$$Y = X\hat{V}(\hat{V}^T\hat{V})^{-1}$$

Next, Y is coded using the mono-channel mode of a legacy codec, such as Opus, where bitrate is allocated to channels proportionately based on the relative weight of the associated singular values in S.

The coded signal is sent to the decoder and decoded as $\tilde{Y}$. Finally, an approximation for X is recovered through

$$\tilde{X}_m = \tilde{Y}\hat{V}_m{}^T$$

The benefit of this method is that it no longer requires the division into frames, while the downside is that the incremental SVD is not as accurate as the conventional SVD and will accrue some error.

4 Results

4.1 Listening test setup

In order to evaluate and compare the different pre-/post-processing methods, a MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) [17] listening test was conducted to compare the six listening conditions listed below:

- channel Family Mapping 3 (FM)
- the three proposed SVD pre-/post- processing methods ($SVD-4W$, $SVD-2W$, $INCR-SVD$)
- hidden reference (H_REF)
- low anchor - 3.5kHz low-pass filtered version of the reference (L_ANC).

Listening test samples were obtained by applying these methods to a total of eight third-order Ambisonics tracks, each with a total bitrate (including side information) of 160 kb/s. For the proposed methods, we truncated the basis vectors to a total of eight right singular vectors (with quantization of each vector component to 8 bits). Each frame consisted of 4096 time-samples, with an overlap region of 128 samples, for

the two proposed methods which involve frames. After processing, files were then rendered to 7.1.4 speakers layout. The IEM plug-in implementation of the All-Round Ambisonics Decoder (AllRAD) [18] was used for rendering.

The test samples, listed in Table 1, included two ambiences (indoor scenes with no dialogue), three music extracts, one speech extract and two sport events (outdoor scenes). They were selected to capture the variety of influence the different listening conditions could have. The sections of each track chosen were those deemed most sensitive.

Except for the speech track, which was encoded into Ambisonics, the selected samples were native Ambisonics recordings, obtained from [19], [20], or provided by the Google team.

Label	Content	Dur.
Ambient1	Moving steps in a cave with water dripping	20s
Ambient2	People jumping in an indoor swimming pool	14s
Music1	Group of musicians playing in a barn	32s
Music2	Music clip with sound effects	9s
Music3	Small ensemble with vocals	26s
Speech	English speech, moving male voice	10s
Sport1	Crowd cheering/whistling in indoor stadium	34s
Sport2	Crowd clapping hands in indoor stadium	12s

Table 1: Listening test material

The samples were divided in two test sessions, and the order in which sessions were taken was distributed among assessors. A replicate was included for each trial.

A tablet interface using Max/MSP allowed the assessors to control the playback and the selection of listening conditions. A segment looping function enabled the assessors to focus on restricted sections of the audio clips. The listening test software allowed for a double-blind system, with randomization of trials order and listening conditions mapping order for each trial.

Listening tests took place in a 7m (L) x 5.33m (W) x3.05 m (H) listening room equipped with 11 loudspeakers and 4 subwoofers supporting 7.1.4 channels

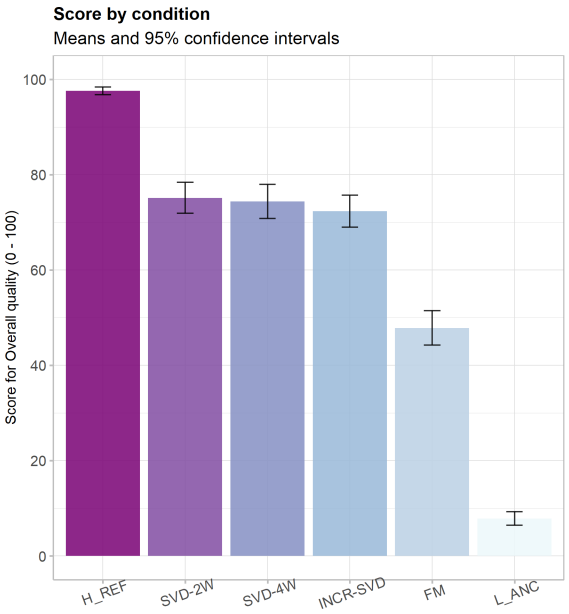


Fig. 5: Overall results

playback. The loudspeakers' and listener's positions were based on ITU-R BS.2051-2 [21] standard. Loudspeakers were level-matched at the listeners's position and met the ITU-R BS. 1116-3 [22] specification.

Fifteen listeners participated in the experiment. Assessors were instructed to compare the six listening conditions to the known reference, and to rate their overall sound quality on the MUSHRA 100-point scale.

4.2 Listening test results

The data was screened for outliers and tested for ANOVA assumptions. The data was found to be heteroscedastic (Levene's test: $p = 6.36 * 10^{-60}$), and the residuals of the linear model were not following a normal distribution (Shapiro's test: $p = 2.46 * 10^{-9}$). A Friedman's test was performed on the dataset as a non-parametric alternative, with the following factors and p-values:

- Condition (6 levels) : $\chi^2(5) = 48.1$, $p = 3.40 * 10^{-9}$
- Sample (8 levels) : $\chi^2(7) = 20.8$, $p = 4.10 * 10^{-3}$

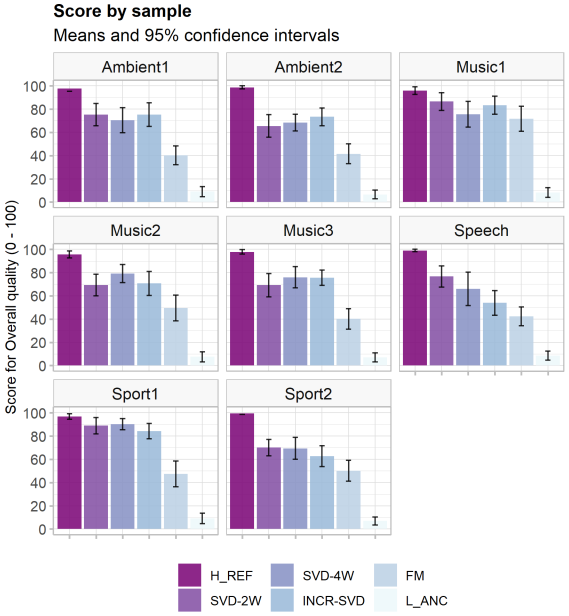


Fig. 6: Results by sample

- Interaction between Condition and Sample: $\chi^2(47) = 397.1$, $p = 6.20 * 10^{-57}$

Wilcoxon's signed rank test was used to perform post hoc paired comparisons between each conditions. Overall, the three proposed methods (SVD-2W, SVD-4W, INCR-SVD) were rated significantly lower than the hidden reference, with p-values inferior to $p = 10^{-33}$, and higher than the low anchor, with p-values inferior to $p = 10^{-60}$. No statistically significant difference was found between these three methods. The FM condition was rated significantly lower than the proposed methods and significantly higher than the low anchor.

The analysis of variance showed that there is a significant effect of the Sample factor on the results, along with an interaction between the Condition and Sample factors. This results from different trends for the tracks Music1, Speech, and Sport2. The FM condition was not rated significantly lower than proposed methods for Music1 and Sport2 tracks. In the Speech track, SVD-2W performed slightly better, and was rated significantly higher than the INCR-SVD condition.

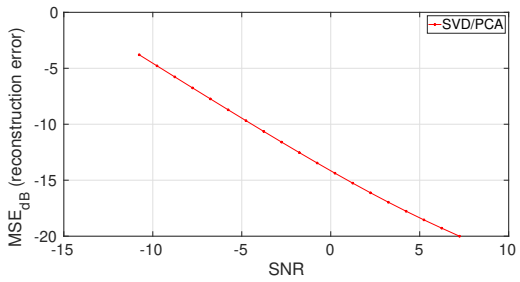


Fig. 7: SVD reconstruction performance (with three basis vectors) of B-format train signal in the presence of babble noise

5 Discussion and Future Work

Towards developing an end-to-end noise-robust encoding of Ambisonics microphone signals using SVD, a better understanding of the performance of SVD-based reconstruction in the presence of noise will be required. For example, Fig. 7 shows the SVD-based reconstruction performance as a function of SNR, using babble noise mixed in all channels of a B-format train sound recording. Three basis vectors were used to reconstruct the B-format signals in a single frame of train signal and noise (shown in Fig. 8). Reconstruction performance (viz. mean square error) was measured by comparing the reconstructed signals, using three-basis vectors, with the initially captured train signal. The performance is degraded at low SNR, and techniques to improve performance will be explored with alternative forms of basis-vector-based reconstruction in the presence of noise. Future research directions involve further examining the use of robust SVD for Ambisonics compression, as well as optimizing the number of background and foreground components.

6 Conclusion

This paper introduces an innovative framework for the pre-/post- processing of Higher Order Ambisonics for use with legacy codecs. The aim is to spatially decorrelate Ambisonics data and allocate bitrate, in an energy-based way, to a legacy coder on an SVD-domain channel-per-channel basis. Listening tests indicate advancements compared to Opus' Channel Family Mapping 3.

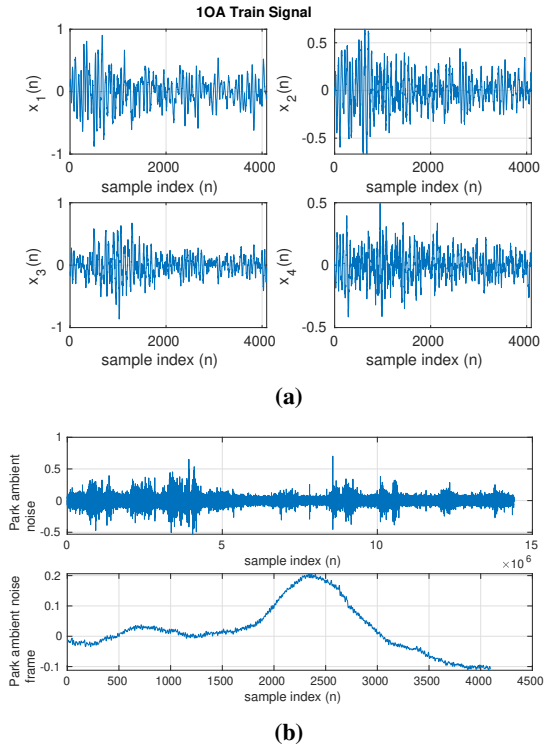


Fig. 8: (a) Frame of Ambisonics (B-format captured) train signals, (b) babble noise (top), babble noise frame (bottom) used for performance analysis (results shown in Fig. 7).

7 Acknowledgements

Samsung Electronics and Samsung Research America supported this work. The authors would like to thank Samsung's US Audio Lab staff, who helped with all aspects of this research, offered insightful suggestions, and contributed to this work.

References

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

[2] "Meta XR Audio SDK Features," `https://developer.oculus.com/documentation/unity/meta-xr-audio-sdk-features/`, 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] Burnett, I., Hellerud, E., Solvang, A., and Svensson, U. P., "Encoding Higher Order Ambisonics with AAC," in *Audio Engineering Society Convention 124*, 2008.

[5] 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.

[6] Peters, N., Sen, D., Kim, M.-Y., Wuebbolt, O., and Weiss, S. M., "Scene-based audio implemented with higher order ambisonics (HOA)," in *SMPTE 2015 Annual Technical Conference and Exhibition*, pp. 1–13, SMPTE, 2015.

[7] 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, doi:10.1109/WASPAA.2017.8170008.

[8] Skoglund, J. and Graczyk, M., "Ambisonics in an Ogg Opus Container," Technical report, 2018.

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

[10] Mahé, P., Ragot, S., and Marchand, S., "First-Order Ambisonic Coding with PCA Matrixing and Quaternion-Based Interpolation," in *DAFx19*, pp. 284–291, 2019.

[11] 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, doi:10.1109/ICASSP.2019.8682315.

[12] 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*, 2021.

[13] 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*, 2022.

[14] 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, doi: 10.1109/ICASSP48485.2024.10447577.

[15] 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, doi: 10.1109/ICASSP48485.2024.10448157.

[16] Kühl, N., Fischer, H., Hinze, M., and Rung, T., "An incremental singular value decomposition approach for large-scale spatially parallel distributed but temporally serial data – applied to technical flows," *Computer Physics Communications*, 296, p. 109022, 2024, ISSN 0010-4655, doi:https://doi.org/10.1016/j.cpc.2023.109022.

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

[18] Zotter, F. and Frank, M., "All-Round Ambisonic Panning and Decoding," *J. Audio Eng. Soc*, 60(10), pp. 807–820, 2012.

[19] "https://www.munkprod.dk/home-ny/," *Munk Productions*, 2021.

[20] "https://mhacoustics.com/demos," *mhacoustics*, 2021.

[21] ITU-R BS.2051-2, "Advanced Sound System for Programme Production," Standard, International Telecommunication Union, Geneva, CH, 2018.

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