Audio Engineering Society

Conference Paper

Presented at the 2022 International Conference on
Audio for Virtual and Augmented Reality
2022 August 15–17, 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.

Spatial Audio Compression with Adaptive Singular Value Decomposition Using Reconstructed Frames

Mahmoud Namazi[1], Ahmed Elshafiy[1], and Kenneth Rose[1]

[1] *University of California, Santa Barbara*

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

ABSTRACT

MPEG-H 3D Audio is the current standard for the compression of higher-order ambisonics data. It uses singular value decomposition (SVD) to spatially decorrelate higher-order ambisonics data, followed by the modified discrete cosine transform to exploit temporal decorrelation. Prominent and ambient sound components are then separately encoded (e.g., using the standard core audio codec) and sent to the decoder. Significant improvements in bitrate and audio quality have been gained in earlier work over MPEG-H by applying the SVD operation in the frequency domain rather than the ambisonics domain. In this work, we provide additional compression gains by adaptively calculating and extending the set of SVD basis vectors, at negligible increase in side information cost, using information attained from the previously reconstructed frame. Objective and subjective results provide evidence for higher compression gains when compared to existing methods.

1 Introduction

Immersive virtual 3D experiences can only be achieved by developing techniques to efficiently encode and stream multimedia content formatted to capture information in three-dimensional spaces. The recent focus on virtual reality (VR) has revived interest in higher order ambisonics (HOA) with its ability to reproduce spatial audio, with any speaker configuration, which enables its use in a widely diverse range of applications. To create immersive and interactive experiences for virtual and augmented reality applications, 3D audio must be encoded in such a way that it can be reconstructed with high quality. Due to the vast amount of data, noting that the HOA paradigm [1] may involve as many as 64 channels, effective transmission and reconstruction of high quality spatial audio poses a considerable compression challenge.

Early attempts at HOA compression made crucial observations with respect to possible tradeoffs and means for redundancy exploitation to achieve higher compression efficiency. In [2], where HOA channels were encoded seperately using the Advanced Audio Codec (AAC), the observation was made that higher bit rate allocation to encode the lower order channels, increases the sound quality in the sweet spot, at the cost of lower spatial resolution. In [3], the objective was low-latency, lossless compression of HOA signals, and an observation was made that there exists significant inter-channel correlation in HOA data, which can be exploited for large compression gains.

One attempt at meeting the aforementioned challenges in HOA compression, MPEG-H 3D Audio [4], has

become the standard for compression of spatial audio. MPEG-H utilizes SVD to decompose HOA data into foreground and background sound components. These components are then separately encoded using the standard audio codec, where psychoacoustic redundancies are exploited in the frequency domain, and sent to the decoder. The transformations are followed by the matching and interpolation of components to mitigate potential mismatch of predominant components in the transitions between consecutive frames (in terms of bitrate and/or reconstruction quality). While matching and interpolation serve to reduce mismatches of principal components between frames, transitions between frames are still suboptimal and unsatisfactory leading to reduced performance, which can be measured both objectively and subjectively.

Improvements upon MPEG-H were made in a series of papers from our lab [5, 6, 7] which introduced interchanging the order of SVD and modified discrete cosine transform (MDCT) to mitigate frame transition effects and achieve greater energy compaction, a noise substitution technique to employ in discarded HOA channels for enhanced perceptual quality, and an adaptive framework for selecting a subset of the SVD basis vectors for explicit transmission, and approximating or replacing the remaining SVD basis vectors via an analysis of the corresponding null space.

In the aforementioned contributions, previous frame data is only exploited for the differential quantization of the foreground basis vectors. In this work, we propose using the previous frame's SVD basis vectors (available at both the encoder and decoder) in order to estimate a significantly better set of extended SVD basis vectors for the current frame, at no additional side information cost. Experimental results show significant gains in terms of objective quality, at higher bitrates, and subjective results provide further validation.

2 Background

2.1 MPEG-H 3D Audio Encoder

The current standard for HOA audio compression is MPEG-H 3D Audio. MPEG-H operates on sequences of $2L \times N$ data blocks, where $N = (M+1)^2$ are the ambisonics channels, M is the ambisonics order, and $2L$ is the frame length (nominal value of L is 1024 samples). The HOA data are segmented into 50% overlapped frames, where the frame of index f is denoted as X_f. SVD is then employed to spatially decorrelate the data, yielding,

$$X_f = U_f \Sigma_f V_f^T \tag{1}$$

where, by definition of the SVD transform, U_f is a $2L \times 2L$ matrix unitary matrix, Σ_f is a $2L \times N$ rectangular diagonal matrix, and V_f is a $N \times N$ unitary matrix. Each of the column vectors in U_f is interpreted as representing normalized audio data decoupled from any directional information, Σ_f is interpreted as a matrix containing the energy components of these directional sound components, and V_f is interpreted as a matrix containing the directional information associated with the columns of U_f, or essentially the basis vectors of the SVD transform.

Next, a set of r basis vectors, denoted $\overline{V}_f$, corresponding to the highest r singular values, is linearly predicted from the corresponding basis vectors of the previous frame, differentially quantized as $\hat{V}_f$, and sent to the decoder as side information. At the encoder, the prominent components (also referred to as foreground) are calculated as

$$Y_f = X_f \hat{V}_f (\hat{V}_f^T \hat{V}_f)^{-1} \tag{2}$$

which essentially approximate the first r columns of $U_f \Sigma_f$. The foreground data is then coded using the core MPEG audio codec, which employs MDCT, in order to exploit psychoacoustic redundancies and temporal correlations. The foreground signal is converted back to the ambisonics domain where it is subtracted from X_f in order to produce the background signal. This background signal is reduced in order and is then also coded using the core MPEG audio codec. A high-level block diagram outline of the MPEG-H Encoder can be seen in Fig. 1.

Due to changes in the order of the principal components (order of SVD basis vectors) between frames, significant inter-frame artifacts may appear which may cause audible audio distortions. In order to remedy this, a matching process, specifically the Hungarian algorithm [8], is employed in order to track and match the dominant components from one frame to the next, followed by interpolation.

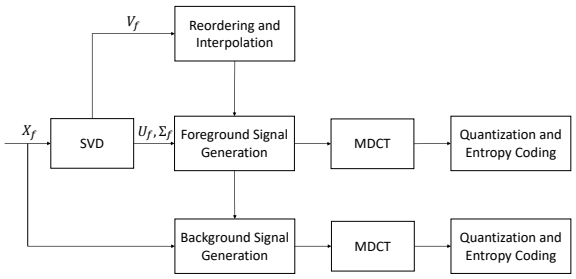

Fig. 1: Outline of the MPEG-H Encoder.

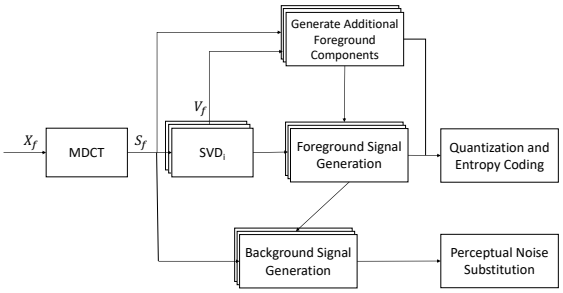

Fig. 2: Outline of the Frequency Domain SVD with Null Space encoder.

2.2 Frequency Domain SVD with Null Space Encoder

To mitigate the above mentioned inter-frame artifacts, a new approach was developed in [5] and further expanded upon in [6] and [7]. This approach employs the built-in overlap created by MDCT in order to smoothen the transition between frames and therefore significantly improved performance.

In this method, each HOA channel signal is separately transformed by MDCT, to obtain the frame's frequency domain representation, S_f, which is an $L \times N$ matrix. S_f is then divided into frequency bands, $S_f^T = [S_{f_1}^T S_{f_2}^T ... S_{f_n}^T]$ where n is the number of frequency bands and each frequency band, indexed by i, is of length l_i, such that $\sum_i l_i = L$. SVD is applied to each frequency band, yielding the decomposition $S_{f_i} = U_{f_i} \Sigma_{f_i} V_{f_i}^T$, similar to (1), except that the signal is now already in the frequency domain. The columns of the rightmost matrix, V_{f_i}, which capture the spatial characteristics of the signal, are called the SVD basis vectors. A set of r basis vectors, denoted $\overline{V}_{f_i}$, corresponding to the r largest singular values, is linearly predicted from the corresponding basis vectors of the previous frame, differentially quantized as $\hat{V}_{f_i}$ and sent to the decoder as side information. One of the primary insights provided by [7] is that $\overline{V}_{f_i}$ can be arbitrarily extended, by both encoder and decoder, through identifying a complementary set of basis vectors that span its null space. The encoder then projects S_{f_i} on the complementary and arbitrarily generated set of basis vectors, identifies the p most prominent vector subset, $V_{f_i,\text{NULL}}$, in terms of the projection energy, and finally, sends their indices to the decoder, *at negligible side information cost*. The decoder generates the same arbitrary null space basis vectors, and extracts $V_{f_i,\text{NULL}}$ using the indices received from the encoder. This adaptive method, along with the implementation of psycho-acoustic modeling, resulted in considerable perceptual and bitrate gains in comparison to MPEG-H.

At the encoder, the prominent components (also referred to as foreground) are calculated as $Y_{f_i} = S_{f_i} \tilde{V}_{f_i} (\tilde{V}_{f_i}^T \tilde{V}_{f_i})^{-1}$, where the concatenated matrix $\tilde{V}_f = [\hat{V}_{f_i}\ V_{f_i,\text{NULL}}]$, then separately quantized and entropy coded in a manner similar to the Advanced Audio Coding (AAC) codec. The background components are obtained by subtracting the concatenated foreground data, in the ambisonics domain, from X_f. The residual is then reduced in ambisonics order, with the discarded channels replaced with perceptual noise, and finally quantized and entropy coded, before being sent to the decoder.

In [7], a simple psycho-acoustic model is introduced which uses a global masking threshold over all HOA channels for each of 49 critical frequency bands, based on the average energy in a given frequency band across all channels. If the energy in the band surpasses the global threshold, the average energy is coded in a method similar to scale factors. Note that this noise substitution model is used in place of the compression of the background signal which occurs in the MPEG-H Encoder. A high-level block diagram that outlines the Frequency Domain SVD with Null Space encoder can be seen in Fig. 2.

While [7] gave significant gains in comparison to the aforementioned previous works [5, 6] and their own

implemented version of the MPEG-H encoder, the complementary null space basis vectors were generated arbitrarily without recourse to the available information at both encoder and decoder from previous reconstructed frames. This immediately suggests that further gains can be achieved by exploiting the past frames information. In particular, the extension of the dominant SVD vectors in [7] is arbitrary and only takes into account the current frame data to select a subset of the null space vectors that were arbitrarily generated. The proposed work will exploit the previous frame information to extend the set of SVD basis vectors.

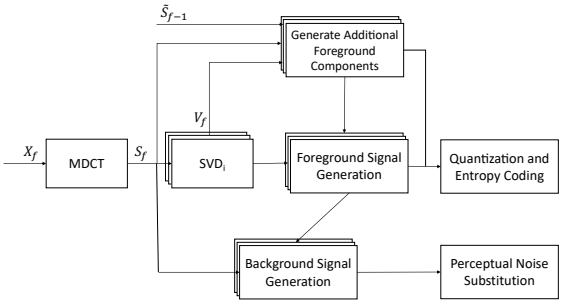

Fig. 3: Outline of the Proposed Encoder.

3 Proposed Method

This work proposes an adjustment of the adaptive framework presented in Section 2.2, wherein the reconstructed previous frame data (in the frequency domain), $\tilde{S}_{(f-1)_i}$, and the current frame's explicitly transmitted prominent basis vectors $\hat{V}_{f_i}$, which are available to both encoder and decoder, are leveraged to generate a better set of complementary basis vectors $V_{f_i,\text{NULL}}$ that span the null space, where "better" is in the sense of energy compaction. For ease of notation, and without loss of generality, the subscript i, denoting the index of frequency band, will be dropped in the subsequent discussion.

In the proposed approach, first a low-rank approximation of $\tilde{S}_{f-1}$ is computed in order to remove the background components. Next, in order to best approximate the spatial directions that correspond to the non-dominant SVD basis vectors in frame f, The encoder projects the low-rank approximation of $\tilde{S}_{f-1}$ on $\hat{V}_f$ then subtracts this projected data from $\tilde{S}_{f-1}$. This subsequently yields the residual, R_{f-1}, i.e., the $f-1$ frame's signal portion that only resides in the null space of $\hat{V}_f$. Next, SVD is performed on this residual, $R_{f-1} = U'_{f-1}\Sigma'_{f-1}V'^{T}_{f-1}$, which reveals the spatial directionality (in terms of the SVD basis vectors V'_{f-1}) of the residual data R_{f-1}. Consequently, V'_{f-1} contains candidate vectors that can best represent the null space of $\hat{V}_f$, based on the reconstructed previous frame. Hence, S_f is projected along V'^{T}_{f-1}, and indices for the p vectors of the highest energy components are sent to the decoder, using similar decision operations as in [7]. Since no additional data is sent to the decoder, beyond the aforementioned indices, there is no increase in side-information cost. A high-level block diagram that outlines the proposed encoder can be seen in Fig. 3.

The proposed method, similar to the changes made between MPEG-H and the frequency-domain SVD methods, further promotes smoothening of the transitions between frames. Since previous frame data is now incorporated into the calculation of the complementary basis vectors, in the frequency domain, thereby further reducing inter-frame discontinuities and artifacts.

4 Experimental Results

In order to validate the efficacy of the proposed approach, both objective measures and subjective listening tests are used. The experiments are conducted on a data set of recordings provided by Google, which consists of 16 third order ambisonics files. These files consist of a variety of sounds, with both moving and stationary sources, including: speech, music, and singing. The following codecs were compared:

- CMPEG: This is our version of the MPEG-H encoder. Due to the unavailability of an official version of the MPEG-H encoder, we implemented our own version based on available patents [9, 10] and documentation [11].[1]
- FSVD-NS: The Frequency Domain SVD with Null Space framework, proposed in [7], which calculates the SVD basis vectors by arbitrarily generating a set that spans the null space of the current frame's dominant basis vectors.

[1] Note that all competing methods build on this same implementation, i.e., they only differ in that they also implement the enhancements described earlier

- PROP: The approach proposed in this work in section 2.2, which uses both current frame and previous frame data in order to extend the current frame's set of SVD basis vectors.

In the CMPEG codec, four foreground and four background components were used, while in the FSVD-NS and PROP codecs, four foreground components were used and four additional components were calculated by each of the proposed methods for extending the SVD basis vectors (in other words, no background components). Therefore, all methods encode a total of eight components using the core audio codec. Two coding modes were used for FSVD-NS and PROP, one with a single frequency band and another with four frequency bands, the mode which minimized the rate was chosen for each frame. The employed core audio encoders use a trellis approach to select optimal scale factors and Huffman codebooks [12]. The core audio coders attempt to minimize the maximum noise to mask ratio (MNMR) for all channels and adjust the MNMR to match a particular bit-rate. In all tests, the data was converted to stereo using a binaural renderer. The renderer decodes the data from the HOA format to a set of speakers. The signals are convolved with HRTFs for each ear and then added together to get the stereo output.

4.1 Objective Results

The average quantization noise to mask ratio (ANMR) of the final binaural reconstruction, averaged across all frames, was chosen as the distortion metric in order to compare the FSVD-NS and PROP encoders. In [7], the FSVD-NS encoder was shown to considerably outperform CMPEG, with gains in the order of 4 dB (see Figure 5 in [7]). Hence the comparison here focuses on the additional gains over FSVD-NS. The ANMR-rate curves have been averaged across all test files for the two competing codecs. The results in Figure 4 show that the proposed codec outperforms the FSVD-NS codec with similar performance at bit rates below 230 Kpbs, but at higher bit rates the gains increase substantially. At the bit rate of 430 Kbps (the rate used in the subjective tests) the objective gain achieved was of about 1.7 dB.

4.2 Subjective Results

A preliminary MUSHRA listening test was performed, in order to validate the objective results and examine true perceptual gains in audio quality. Listeners were provided with five, randomly-ordered, versions of each of the 16 audio track, all matched at 430 Kbps. These included versions of the tracks encoded using the three different audio codecs (CMPEG, FSVD-NS, and PROP), a hidden reference (REF), and a 3.5 kHz low pass filtered version of the reference (ANC). They were asked, after hearing the labeled reference, to label each track with a numerical score between 0 and 100, with higher scores referring to higher audio quality (most resembling the labeled reference). The scores averaged across all audio files and the 95% confidence interval are shown in Figure 5. The results show that the proposed method, on average, outperformed all of the competing codecs, albeit with some overlap in confidence interval with the FSVD-NS compression codec.

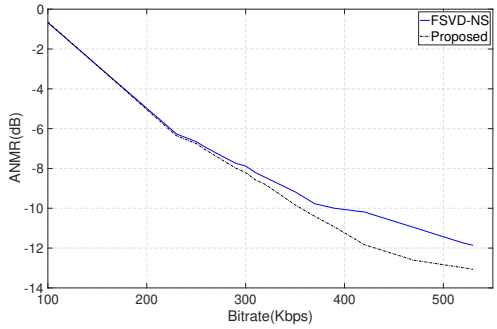

Fig. 4: Objective test results.

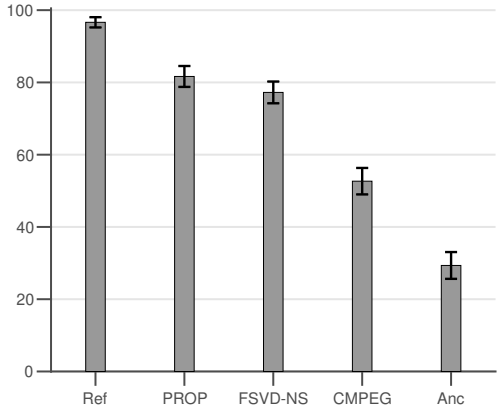

Fig. 5: MUSHRA listening test results.

5 Conclusion

This work presents a new adaptive framework, which takes into account both the previous frame reconstructed data and the current frame data, to obtain a more relevant set of basis vector spanning the null space, for extending the available set of SVD dominant basis vectors. Objective and subjective measures show improvements over the previous method where the SVD basis vectors were extended in an arbitrary fashion, which itself offered considerable gains over MPEG-H. Future research includes the optimization of the number of foreground and background components and the exploration of using additional past frames in extending the set of SVD basis vectors.

6 Acknowledgment

The authors thank Google, Inc., and especially Jan Skoglund and Drew Allen, for providing the ambisonics dataset and the binaural renderer used in this work.

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*, Audio Engineering Society, 2003.

[2] Hellerud, E., Burnett, I., Solvang, A., and Svensson, U. P., "Encoding higher order ambisonics with AAC," 2008.

[3] 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, IEEE, 2009.

[4] Herre, J., Hilpert, J., Kuntz, A., and Plogsties, J., "MPEG-H 3D audio—The new standard for coding of immersive spatial audio," *IEEE Journal of selected topics in signal processing*, 9(5), pp. 770–779, 2015.

[5] 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, IEEE, 2017.

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

[7] 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, IEEE, 2019.

[8] Kuhn, H. W., "The Hungarian method for the assignment problem," *Naval research logistics quarterly*, 2(1-2), pp. 83–97, 1955.

[9] Sen, D. and Ryu, S.-U., "Compression of decomposed representations of a sound field," 2014, uS20140358563A1.

[10] Sen, D. and Peters, N. G., "Interpolation for decomposed representations of a sound field," 2014, wO-2014194099-A1.

[11] "Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 3: 3D audio," 2015.

[12] Aggarwal, A., Regunathan, S. L., and Rose, K., "A trellis-based optimal parameter value selection for audio coding," *IEEE transactions on audio, speech, and language processing*, 14(2), pp. 623–633, 2006.