FREQUENCY DOMAIN SINGULAR VALUE DECOMPOSITION FOR EFFICIENT SPATIAL AUDIO CODING

Sina Zamani, Tejaswi Nanjundaswamy, Kenneth Rose

Department of Electrical and Computer Engineering, University of California Santa Barbara, CA 93106
E-mail: {sinazmn, tejaswi, rose}@ece.ucsb.edu

ABSTRACT

Advances in virtual reality have generated substantial interest in accurately reproducing and storing spatial audio in the higher order ambisonics (HOA) representation, given its rendering flexibility. Recent standardization for HOA compression adopted a framework wherein HOA data are decomposed into principal components that are then encoded by standard audio coding, i.e., frequency domain quantization and entropy coding to exploit psychoacoustic redundancy. A noted shortcoming of this approach is the occasional mismatch in principal components across blocks, and the resulting suboptimal transitions in the data fed to the audio coder. Instead, we propose a framework where singular value decomposition (SVD) is performed after transformation to the frequency domain via the modified discrete cosine transform (MDCT). This framework not only ensures smooth transition across blocks, but also enables frequency dependent SVD for better energy compaction. Moreover, we introduce a novel noise substitution technique to compensate for suppressed ambient energy in discarded higher order ambisonics channels, which significantly enhances the perceptual quality of the reconstructed HOA signal. Objective and subjective evaluation results provide evidence for the effectiveness of the proposed framework in terms of both higher compression gains and better perceptual quality, compared to existing methods.

Index Terms— Higher Order Ambisonics, spatial audio coding, audio compression, 3D audio

1. INTRODUCTION

The ambisonics paradigm was originally developed [1, 2] as a promising technique for reproducing three dimensional sound fields. However, it was not commercially successful at the time due to poor directionality and the limited size of the "sweet spot". Later, with the advent of higher order ambisonics (HOA)[3], the approach was extended by sound field decomposition into higher order spherical components, resulting in improved localization, spatial resolution, and increased size of the sweet spot. Recent advances in virtual reality and many related applications such as streaming of real time music performances or cinematic scenes, have fueled interest in HOA to accurately reproduce spatial audio. HOA data (typically obtained from a microphone array) are high dimensional and pose significant challenges on storage and transmission for practical applications, which motivate the development of novel effective compression techniques to considerably reduce the required bit-rate.

Early approaches [4] directly encoded individual HOA channels with Advanced Audio Coding (AAC), independent of other channels. It was observed that allocating more bits to lower order components increases the sound quality in the sweet spot but decreases spatial resolution. As such approaches clearly neglect inter-channel redundancies, later work [5] employed a lossless compression technique to exploit inter-channel correlations, by choosing one of the channels as reference for predicting the other channels, followed by encoding of the prediction residue.

Recently, a new spatial audio coding standard, MPEG-H 3D Audio [6], has emerged. The HOA input is decomposed into predominant sound elements and ambient background components, using standard singular value decomposition (SVD), and each of these are coded separately via an AAC based coder, where quantization and entropy coding are performed in the frequency domain to exploit psychoacoustic redundancies. While good broadcast quality has been reported for bit-rates around 300 kbps [7], the premise of this paper is that higher compression efficiency and better perceptual quality can be achieved by employing SVD in the frequency domain. In the MPEG-H approach, there is often a mismatch of principal components across blocks, both in terms of order of components and their respective basis vectors. MPEG-H employs an elaborate matching technique in combination with an overlap-add technique to mitigate this shortcoming. However, transitions between blocks remain suboptimal and introduce inefficiencies in the core codec and degrade the perceptual quality. Our approach completely eliminates this issue as we first transform to frequency domain via MDCT which ensures smooth transition across blocks with its built-in overlap. Moreover, optimal SVD can now be adapted to different frequencies, instead of a compromise decomposition for the entire spectrum. Finally, we employ noise substitution in a novel way to compensate for ambient energy loss and further improve perceptual quality of the rendered HOA data.

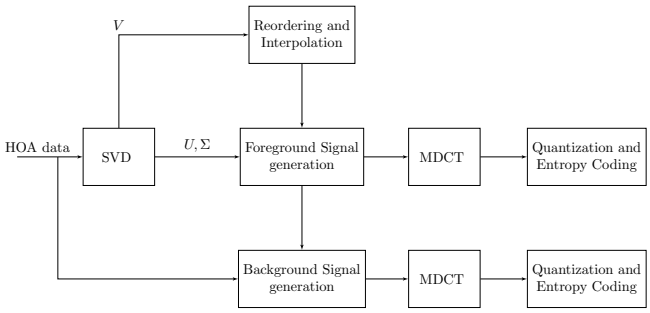

Figure 1: Overview of MPEG-H encoder

2. MPEG-H APPROACH FOR COMPRESSION OF HOA DATA

The MPEG HOA encoder [8] processes the input HOA data over frames of length $2L$ ($L = 1024$) with 50% overlap. Let the number of HOA channels be $M = (N+1)^2$, where N is the ambisonics order. For current frame f, the encoder operates on HOA data X_f, which is an $2L \times M$ matrix, and performs standard singular value decomposition (SVD),

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

where U_f is an $2L \times 2L$ unitary matrix, Σ_f is a $2L \times M$ rectangular diagonal matrix with non-zero elements on the diagonal and V_f is an $M \times M$ unitary matrix. The SVD construction ensures that predominant components, corresponding to the largest r singular values have as basis vectors the first r columns of V_f. Let V_f be truncated to the first r columns, and further be independently or differentially quantized to $\hat{V}_f$ and sent to the decoder as side information for each frame, so as to enable it to transform back the predominant components to the ambisonics domain. To keep encoder and decoder in sync, the quantized $\hat{V}_f$ is used to generate the predominant components $\tilde{Y}_f$ (now an approximation of first r columns of $U_f \Sigma_f$), as,

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

Note that the inverse term is for renormalization of the quantized basis vectors (to maintain unitarity). The next step is to code the predominant or foreground components, each corresponding to a column of $\tilde{Y}_f$, using separate instances of the core audio codec. This requires concatenating components across frames. However, since SVD arranges the basis vectors based on the singular value magnitudes, the same foreground component might change position in $\tilde{Y}$ from frame to frame depending on the magnitude of its singular value relative to others. This can result in noticeable blocking artifacts if blindly concatenated foreground components are fed to the core codec. While there are several approaches to reorder and match components with the previous frame, we employ the magnitude of correlation between column vectors of $\hat{V}_f$ and $\hat{V}_{f-1}$ as the criterion in an Hungarian matching algorithm [9], which we found to be effective.

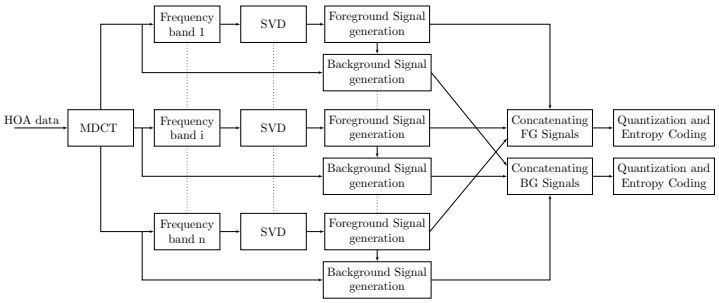

Figure 2: Overview of proposed method encoder

Even with matched components, simple concatenation across frames would introduce noticeable artifacts as a small change in the basis vector causes some mismatch at the frame boundary. Hence the encoder interpolates the column vectors of $\hat{V}$ between current frame and previous frame to ensure continuity over time. Specifically, a different transform matrix is used for each sample of the current frame, whose column vectors are obtained as,

$$\bar{v}_f^i(l) = (1 - w(l))\hat{v}_{f-1}^i + w(l)\hat{v}_f^i, \tag{3}$$

where $\hat{v}_f^i, \hat{v}_{f-1}^i$ are the ith matched column vectors of $\hat{V}$ for current and previous frames, $\bar{v}_f^i(l)$ is ith column vector for sample l in current frame and $w(l)$ is a window function, which may be the triangular or Hanning window. The interpolation should also account for the fact that the vectors might get negated from one frame to next frame by performing a sign correction when needed.

An approximation of the HOA data, $\tilde{X}_f$, is generated by transforming the foreground components back to the ambisonics domain, which is then subtracted from the original data to produce the ambient (or background) HOA data. The foreground components are coded using separate instances of the core audio codec. The order of background HOA data is then reduced (from N to some t) and this lower order HOA data are also coded using the core audio codec. An illustration of the MPEG-H approach is shown in Figure 1.

3. FREQUENCY DOMAIN SVD FOR HOA DATA COMPRESSION

Clearly, the MPEG-H approach performs an elaborate process of matching and interpolating transform basis vectors of consecutive frames to improve their continuity over time and to mitigate the artifacts stemming from blockwise SVD application. We propose to circumvent this underlying and fundamental shortcoming with a framework wherein SVD is employed *after* transformation to frequency domain via MDCT, which naturally achieves the required smoothness with its built-in overlap. Moreover, this framework enables the significant flexibility to make both the SVD and the number of components to be retained, adaptive to frequency, instead of using a compromise for the varying needs of different frequency bands.

In the proposed approach, the HOA data are processed in the encoder after segmenting each HOA channel into 50% overlapped frames of length $2L$. The samples of each channel are separately transformed via MDCT after windowing to obtain the transformed data for the current frame, S_f, which is an $L \times M$ matrix. S_f is now divided into different 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 with lengths $l_1, l_2, ..., l_n$ and $\sum_i l_i = L$. For each frequency band, a different SVD is obtained, $S_{f_i} = U_{f_i} \Sigma_{f_i} V_{f_i}^T$, the top r_i components are retained (which may vary over bands), and the correspondingly truncated V_{f_i} are coded to $\hat{V}_{f_i}$ and sent to the decoder as side information. The prominent components, $\tilde{Y}_{f_i}$, are now obtained similar to (2) for each band and concatenated together to form $\tilde{Y}_f^T = [\tilde{Y}_{f_1}^T \tilde{Y}_{f_2}^T ... \ \tilde{Y}_{f_n}^T]$. Given the foreground components, an approximate $\tilde{S}_f$ is obtained and subtracted from original data S_f to obtain the residual. The columns of $\tilde{Y}_f$ are then separately quantized and entropy coded similar to the AAC codec. Finally, the residual is reduced in ambisonics order, and quantized and entropy coded similar to the AAC codec. An illustration of the proposed approach is shown in Figure 2.

3.1. Side Information Compression

To exploit the temporal correlations between transform matrices of consecutive frames, the Hungarian algorithm [9] is employed to match the column vectors of V_{f_i} matrices of consecutive frames corresponding to ith frequency band based on correlation coefficients. We used a scalar prediction coefficient (equal to correlation coefficient) for each vector. We selected approximately 10,000 frames from third order ambisonics files as training set to design a quantizer for prediction coefficients and prediction residuals using Generalized Lloyd Algorithm (GLA).

3.2. Noise Substitution

Reduction of ambient HOA data order results in suppression of ambient energy. Since coding all the channels of ambient components to retain the energy may often represent a prohibitive cost in bit-rate, we propose an approach where along with encoding reduced order ambient data with the core codec, we substitute the discarded channels with perceptual noise in frequency bins across these channels. Specifically, we measure the spectral flatness in each of the discarded channels for each of the 49 AAC frequency groups as following,

$$\text{Flatness}_f^{ij} = \frac{exp(\frac{1}{|B_f^{ij}|} \sum_{k \in B_f^{ij}} \ln B_f^{ij}[k])}{\frac{1}{|B_f^{ij}|} \sum_{k \in B_f^{ij}} B_f^{ij}[k]}, \qquad (4)$$

where B_f^{ij} are the power spectrum coefficients for channel i and frequency group j of the current frame background data. This flatness is averaged across all channels for each frequency group and if the average is larger than a threshold, randomly generated noise is added to these frequencies in the discarded ambient channels at the decoder. The average energy across all channels in this frequency group is sent to the decoder to generate the noise at the original content energy. Thus, a maximum of 49 energy values are coded similar to scalefactors in AAC and sent to the decoder for each frame.

4. EXPERIMENTAL RESULTS

To validate the efficacy of the proposed approach we conducted objective and subjective experiments. The experiment was on a dataset of recordings provided by Google, which consists of 6 third order ambisonics files. As the software for MPEG-H encoder is not yet publicly available, we implemented our own representative version of it, as described in Section 2, based on the published patents [10, 11] and the standard documentation [8] which serves as a baseline for comparison. Other than the explicit contributions of the new approach, the competitors are identical in terms of options enabled, etc. All side information is accounted for in the total bit-rate.

In all the experiments $r = r_i = 4, \forall i$ and $t = 1$, that is, the number of foreground and background channels are both set to 4 for all frames, which results in a total of 8 components being encoded with the core codec. In the proposed approach, we divided the frequency data into 4 uniformly sized bands and a different transform is obtained for each frequency band. While employing frequency dependent SVD always results in better compaction of energy, this does not always translate to improved RD performance for the fixed quantizers and entropy coders employed. We believe this limitation can be addressed by redesigning the quantizers and entropy coders for the new statistics. In order to obtain preliminary results we employed the "shortcut" of providing two encoding modes per frame, of using a single frequency band (mode $m_f = 0$), or using 4 frequency bands (mode $m_f = 1$), and selecting the one which minimizes the RD cost. When the mode switches between frames, the transform matrix (or matrices) of current frame are predicted from the best available previous transform matrix (or matrices).

4.1. Objective Results

Note that perceptual distortion optimization for foreground data obtained through SVD, especially in comparison to background data in ambisonics domain, is still an open problem. To obtain preliminary objective results, we simply encoded both the competing methods to minimize the bitrates for a given maximum quantization noise to mask ratio (MNMR) constraint for all bands of all channels. Investigation of the true objective perceptual distortion measure and its corresponding optimization approach is part of future work. Percentage reduction in bit-rate for the proposed method in comparison to the MPEG-H approach, obtained at different operating points is presented in first part of Ta-

	Bit-rate reduction at various operating points			Foreground contribution		Background contribution	
Sequence	∼308 kbps	∼375 kbps	∼500 kbps	Proposed	MPEG	Proposed	MPEG
2src_conv_office	7.83%	8.03%	8.37%	67.78%	59.52%	25.06%	37.09%
A Round Around-Eigen	-0.8%	0%	1.41%	56.37%	53.12%	37.27%	44.85%
doll_intro	6.64%	6.93%	7.26%	63.93%	57.24%	26.99%	38.73%
helicopter_fountain	4.58%	5.96%	7.02%	54.6%	51.33%	41.76%	47.29%
lyon	3.98%	4.72%	3.58%	59.65%	55.94%	27.65%	39.34%
Murmur2	6.50%	8.89%	10.9%	63.08%	60.53%	26.47%	37.25%
Average	4.79%	5.75%	6.44%	60.90%	56.26%	30.87%	40.76%

Table 1: Proposed framework's reduction in bit-rate and contribution of foreground and background data to total bit-rate for the two methods

ble 1. Clearly, there is a consistent improvement in performance for the proposed framework. Table 1 also presents the contribution of foreground and background data to the total bit-rate for each file (averaged over the three operating points) for the two encoding methods. Clearly, the improved energy compaction of the proposed approach results in significant reduction in bit-rate required for background, while marginally increasing the foreground bit-rate.

4.2. Subjective Results

We conducted subjective evaluations to determine the true perceptual gains using the MUSHRA listening tests [12]. This is particularly important given the above reservations about the ability of the objective measure to fully capture the perceptual quality. The test items were scored on a scale of 0 (bad) to 100 (excellent) and the tests were conducted with 8 listeners. We extracted 10s portions of each file for evaluation. The test files includes challenging scenes with speech, music and objects moving. A binaural renderer was deployed to convert the reconstructed HOA coefficients to stereo signals. Randomly ordered 4 versions of each audio sample (including a hidden reference, a 3.5 kHz low-pass filtered anchor, the encoded file using the proposed method and the encoded file using the MPEG method) were presented to the listeners. For these tests, the bit-rates (around 375 kbps) were matched for each competing file. The subjective evaluation results, including the mean and 95% confidence intervals, as presented in Figure 3 clearly demonstrates the substantially improved quality. This margin of improvement could not have been predicted from the moderate gains observed in objective results, clearly highlighting the critical need for further research in developing an appropriate objective perceptual distortion measure and corresponding optimization approach. The files used for these subjective tests are shared in [13].

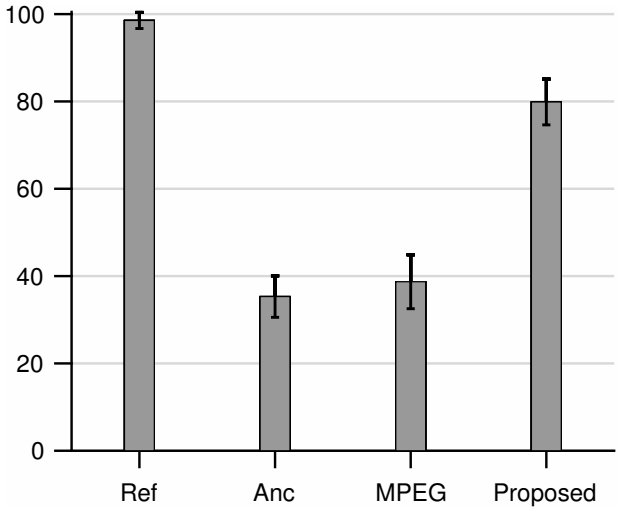

Figure 3: MUSHRA listening test results comparing the encoding techniques

5. CONCLUSION

This paper presents a new framework for compression of higher order ambisonics data by first transforming the coefficients to MDCT domain and then decomposing into principal components. Unlike the current approaches, which suffer from suboptimal transitions between frames, the proposed approach not only ensures smooth transitions, it also enables frequency dependent decomposition and selection of dominant components. Furthermore, a novel way of employing noise substitution is introduced to enhance the perceptual quality of final reconstructions. Objective and subjective results illustrate the effectiveness of the proposed approach with significant performance improvements. Future research includes optimally deciding number and size of frequency bands, frequency dependent optimization of the number of foreground and background components, redesign of quantizers and entropy coders, and investigation of better objective perceptual distortion measure.

6. ACKNOWLEDGMENT

This research was supported by Google, Inc. We are particularly grateful to Jan Skoglund and Drew Allen from Google for useful discussions and providing us with the ambisonics dataset and the binaural renderer used in Sec. 4.

7. REFERENCES

[1] M. A. Gerzon, "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[2] M. Gerzon, "Ambisonics in multichannel broadcasting and video," *Journal of the Audio Engineering Society*, vol. 33, no. 11, pp. 859–871, 1985.

[3] J. Daniel, S. Moreau, and R. Nicol, "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging," in *Audio Engineering Society Convention 114*, 2003.

[4] E. Hellerud, I. Burnett, A. Solvang, and U. P. Svensson, "Encoding higher order ambisonics with AAC," in *Audio Engineering Society Convention 124*, 2008.

[5] E. Hellerud, A. Solvang, and U. P. Svensson, "Spatial redundancy in higher order ambisonics and its use for lowdelay lossless compression," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 269–272, 2009.

[6] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D audio - the new standard for coding of immersive spatial audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779, 2015.

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

[8] *Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio*, ISO/IEC Std. ISO/IEC JTC1/SC29 23 008-3:2015, 2015.

[9] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[10] D. Sen and N. Peters, "Interpolation for decomposed representations of a sound field," Dec. 4 2014, WO2014194099 A1.

[11] D. Sen and S.-U. Ryu, "Compression of decomposed representations of a sound field," Dec. 4 2014, US20140358563 A1.

[12] *Method of Subjective Assessment of Intermediate Quality Level of Coding Systems*, ITU Std. ITU-R Recommendation, BS 1534-1, 2001.

[13] "HOA subjective listening test files." [Online]. Available: https://scl.ece.ucsb.edu/hoa-waspaa-demo