SPATIAL AUDIO CODING WITHOUT RECOURSE TO BACKGROUND SIGNAL COMPRESSION

Sina Zamani, Kenneth Rose

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

ABSTRACT

The MPEG-H 3D audio standard applies singular value decomposition (SVD) to the input higher order ambisonics data, then encodes each predominant (foreground) sound component independently using a standard core audio codec. The residual (background) signal is encoded in the ambisonic domain. This paper is motivated by the observations: i) separate coding of SVD components ignores spatial inter channel masking effects; ii) compression in both SVD and ambisonic domains is difficult to perceptually optimize; iii) Only few predominant components are encoded due to the prohibitive side information cost of specifying SVD basis vectors. The proposed coding architecture overcomes the first two concerns by performing all compression in the SVD domain with a masking threshold that is calculated jointly for all encoded components, thereby accounting for cross-component masking. The third shortcoming is circumvented by a novel method for extending a given set of SVD basis vectors at no side information cost, by computing (at both encoder and decoder) basis vectors to span the null space of the transmitted basis vectors. Experimental results provide evidence for substantial objective and subjective gains.

Index Terms— Higher order ambisonics, spatial audio coding, audio compression, 3D audio

1. INTRODUCTION

Creating interactive and immersive experiences is currently an area of significant interest, with major investment in virtual and augmented reality covering all aspects of content acquisition, storage, transmission, and display/playback. Achieving a truly immersive experience requires new formats for multimedia content, and in particular three-dimensional (3D) 360-degree audio, to represent information in a 3D space. The higher order ambisonics (HOA) paradigm [1, 2, 3, 4] is a surround sound recording and reproduction technique that captures information of a 3D sound-field in its transmission channels. The key benefit of HOA is its flexibility to enable playback with any speaker configuration ranging from headphones to complex surround sound systems, thus allowing for a diverse variety of approaches to create an immersive experience. In practical applications, HOA data can include as many as 64 channels and given the enormous amount of data consumed by 3D audio, it is critical to achieve efficient compression for networking and transmission.

The recent MPEG-H 3D audio standard [5] is the state-of-the-art for compression of HOA data. The encoder utilizes SVD to extract and encode distinct spatial audio objects, also refereed to as predominant or foreground sound components, which requires the SVD transform matrices to be encoded and sent to decoder as side information. The residual signal, not captured by the predominant components, is encoded in the ambisonics domain after it has been reduced in order, where the remaining ambisonic channels are called ambient or background components. Each foreground or background component is fed into a separate standard audio codec where it is independently encoded. Broadcast quality transmission and transparent quality transmission have been reported [6, 7] at bit-rates around 300 kbps and 500 kbps, respectively. However, one central concern with this framework, is the occasional mismatch between principal components across blocks, that could create abrupt transitions between adjacent frames. MPEG-H 3D employs an elaborate process of basis vector matching and interpolation to address this issue, however the transitions across frames remain sub optimal and degrade performance in terms of the achievable compression ratio and resulting perceptual quality. In our recent work [8], we demonstrated that considerable gains can be obtained by performing SVD in the frequency domain instead of on the original time sequence. This paradigm ensures smooth transitions between frames by leveraging the modified discrete cosine transform (MDCT) built-in overlap windows. This framework also offers the advantage of adapting the optimal SVD to different frequency bands, instead of a compromise decomposition for the entire spectrum.

Several stumbling blocks stand in the way of optimal spatial audio coding. Psychoacoustic models and distortion measures that can accurately account for human perception of 3D audio, and an effective optimization framework to find the encoding parameters that minimize such a distortion metric, are both elusive and subjects of ongoing research. Moreover, existing approaches only encode a few predominant components, mainly due to the prohibitive cost in side information. Consequently, a significant portion of the main HOA data energy and directional information leaks back to the ambient data, and MPEG-H 3D reverts to encoding the first order background data to recapture some of this signal. The difficulty of optimizing the encoding parameters is further exacerbated by the fact that the predominant sound components are encoded in the SVD domain, while the ambient data are encoded in the ambisonics domain, and it is not obvious how to properly account for masking effects and the contribution of quantization noise in each coded component to the final distortion. As a result, existing techniques default to the straightforward, though quite sub-optimal, route of encoding the predominant and ambient components independently. The main shortcomings of such approaches are that they completely neglect inter channel masking effects and fall short of realizing the full potential of SVD to decompose the HOA data into sound components thus eliminating spatial redundancies.

As a first "coarse" approach to achieve a proof of concept and demonstrate the potential benefits of circumventing the above shortcomings, this paper proposes a novel encoding architecture where only predominant components are encoded. The premise is that the capability of SVD to extract and decorrelate distinct spatial audio objects, should be fully exploited and, moreover, the setting allows

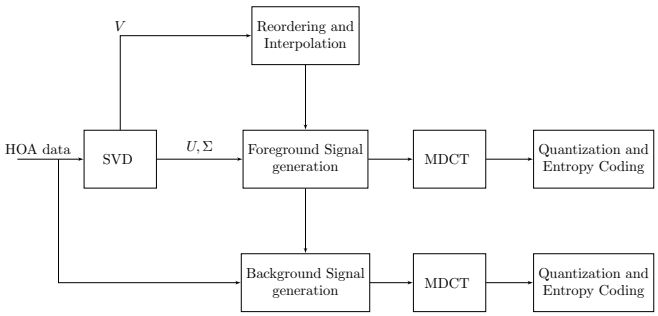

Fig. 1. Overview of the MPEG-H encoder

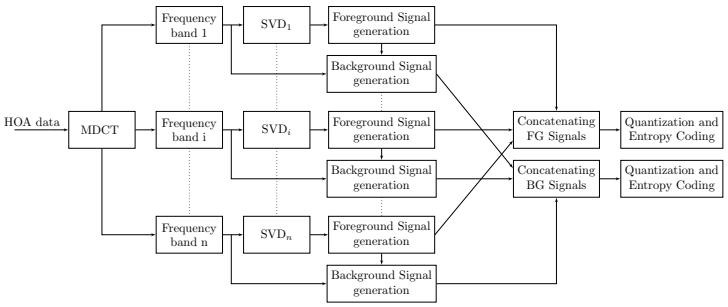

Fig. 2. Overview of frequency domain SVD as proposed in [8]

for better handling of perceptual masking effects. To show the potential gains from accounting for inter-channel masking effects, a first crude approach is proposed where the energy in a given frequency band, averaged across predominant components, is used to calculate the common quantization noise to masking threshold ratio for all encoded channels. The proposed framework, where all compression is performed in the SVD domain, offers increased adaptivity compared to existing techniques, as it encodes more predominant components. To reduce the prohibitive burden of side information, we propose a new paradigm that extends the set of predominant basis vectors with an approximate complementary set, at no side information cost. Experimental results show that the proposed framework achieves significant performance gains over conventional encoding methods.

2. BACKGROUND

2.1. HOA compression in MPEG-H 3D

The MPEG-H 3D encoder processes the input HOA data in overlapping frames of length $2L$. Let $N = (M+1)^2$ be the number of ambisonics channels, where M is the ambisonics order. Then for the data of frame f, denoted by X_f, a factorization of the following form is obtained using singular value decomposition (SVD),

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

where X_f is a $2L \times M$ matrix, U_f and V_f are unitary matrices of sizes $2L \times 2L$ and $M \times M$, respectively, and Σ_f is a rectangular diagonal matrix whose non-zero elements on the diagonal are sorted in decreasing order. Each of the N vectors in U_f (of length $2L$ samples) can be interpreted as representing normalized separated audio signals that have been decoupled from any directional information, and Σ_f stores the energy of these sound components. The spatial characteristics are captured by individual columns of V_f, or basically the basis vectors of the SVD transform.

For each frame, V_f is truncated to the first r columns, which correspond to the largest singular values, then differentialy quantized to $\hat{V}_f$ and sent to the decoder as side information. Predominant (foreground) components are generated by projecting the original data along the quantized basis vectors (columns of $\hat{V}_f$) as,

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

Note that the inverse factor on the righthand side is for renormalization of the quantized basis vectors (to maintain unitarity). The foreground components, which approximate the first r columns $U_f \Sigma_f$, are then independently coded using MPEG's core audio coded that employs MDCT to exploit temporal correlations and psychoacoustic redundancies. However, there is no guarantee that the principal components of the current frame retain the same order as in the previous frame. Concatenating the predominant components across frames and passing them to the core audio codec without accounting for this reordering issue results in severe discontinuities across consecutive frames that can reduce the achievable compression ratio and introduce significant artifacts in the HOA reconstruction. Moreover, as foreground components capture the original data projected along the basis vectors, even small variations in $\hat{V}_f$ can exacerbate the discontinues at frame boundaries. To partially smooth these discontinues, the encoder performs an elaborate process, where first the basis vectors of adjacent frames (columns of $\hat{V}_f$ and $\hat{V}_{f-1}$) are matched for a chosen criterion (e.g., correlation) using the Hungarian algorithm [9]. Then the overlap section of these frames is projected along the two matched basis vectors, and the resulting components are temporally interpolated.

The foreground approximation of the signal is mapped back to the ambisonics domain and subtracted from X_f to produce the ambient (or background) sound components. The order of background HOA data is then reduced (usually to first order) and each channel is independently encoded using the standard core audio codec. A high level block diagram for the MPEG-H approach is shown in Fig.1.

2.2. Frequency domain SVD and perceptual noise substitution

Although the elaborate matching and interpolation procedure employed by MPEG-H 3D achieves some smoothing of discontinuities, the framework still suffers from suboptimal transitions between frames which significantly degrade the codec performance. In a recent work [8], we circumvented this fundamental shortcoming of MPEG-H 3D by performing the spatial decomposition *in the spectral domain*. In the revised framework, SVD is performed after transformation by MDCT, which ensures smooth transitions between frames, due to MDCT's built-in overlap windows. This framework also offers a major bonus: flexibility to make both the SVD and the number of components to be retained, adaptive to the needs of specific frequency bands, instead of a compromise SVD for the entire frame.

To map HOA data to the frequency domain, the encoder of [8] applies MDCT to each ambisonics channel, separately. The resulting $L \times M$ matrix, denoted by S_f, is divided into smaller frequency bands, , i.e., $S_f^T = [S_{f_1}^T S_{f_2}^T ... \ S_{f_n}^T]$, where n is the number of frequency bands with lengths $l_1, l_2, ..., l_n$, where $\sum_i l_i = L$. A different SVD decomposition can be optimized for each band , $S_{f_i} = U_{f_i} \Sigma_{f_i} V_{f_i}^T$. Transform matrices, V_{f_i}s, are truncated to first r columns, differentially quantized to $\hat{V}_{f_i}$, and sent to the decoder as side information. Similar to (2), predominant components for each band are obtained as $\tilde{Y}_{f_i} = S_{f_i} \hat{V}_{f_i} (\hat{V}_{f_i}^T \hat{V}_{f_i})^{-1}$, and are

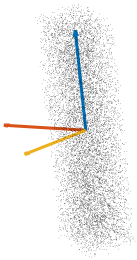

Fig. 3. A simple 3D example: the blue vector is sent to decoder, and 2 complementary vectors spanning the null space are computed.

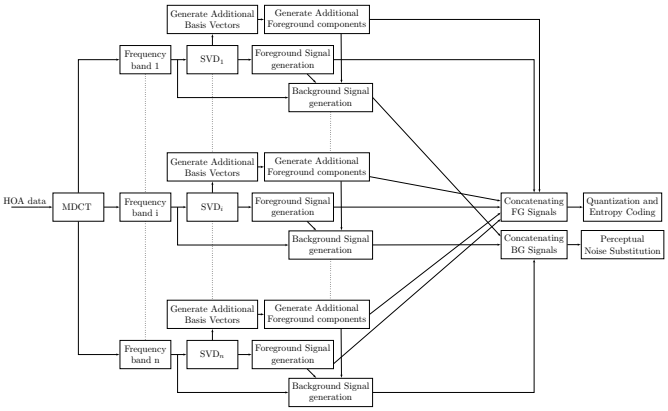

Fig. 4. Overview of the proposed encoder

concatenated to generate the foreground data for the entire frame, $\tilde{Y}_f^T = [\tilde{Y}_{f_1}^T \tilde{Y}_{f_2}^T ... \tilde{Y}_{f_n}^T]$. Predominant components are mapped back to ambisonics domain to provide an approximation of HOA data in spectral domain, $\tilde{S}_f^T = [\tilde{S}_{f_1}^T \tilde{S}_{f_2}^T ... \tilde{S}_{f_n}^T]$, where $\tilde{S}_{f_i} = \tilde{Y}_{f_i} \hat{V}_{f_i}^T$, and $\tilde{S}_f$ is subtracted from S_f to produce the background components. The predominant and ambient sound components are fed to different instances of core audio codec's quantization and entropy coding modules. The proposed approach is illustrated in Fig.2.

Finally, recall that, for rate considerations, MPEG-H 3D discards the higher order channels of the background (ambient) ambisonics data. The encoder fails to adequately compensate for the content and energy of discarded components, which causes significant suppression of ambient sound, and degradation of the perceptual quality of the overall reconstruction. In [8] we proposed a technique to mitigate this shortcoming by replacing the content of discarded channels with noise designed to be perceptually relevant. Specifically, the signal in the discarded higher order background channels is divided into the 49 critical frequency groups, and spectral flatness is calculated for each group in each channel. The flatness values of a frequency group are averaged over all channels and compared to a threshold to determine if the content is "noise-like". For each "noise-like" group, the average energy is encoded similar to scale factors, and sent to the decoder as side information. The decoder generates perceptual noise for substitution according to the energy profile specified across frequency groups [8].

3. PROPOSED APPROACH

Compression often involves an inherent tradeoff between the benefits of adaptivity and its cost in side information. MPEG-H 3D's conversion of ambisonics data to the SVD domain opens the door to effective adaptation to spatial configurations. But, in practice, such adaptivity is severely restricted to very few predominant SVD components (typically 4), due to the prohibitive cost in side information to update the SVD basis vectors. In order to compensate for this limitation, MPEG-H 3D employs an ad hoc "fix", which consists of mapping the residual of the foreground procedure back to the HOA domain for background re-encoding to capture some of the loss. We propose to instead overhaul the framework such that it maximizes adaptivity by performing all compression in the SVD domain, but at no additional cost in side information. The subterfuge is to add basis vectors that span the null space of the predominant SVD components specified to the decoder. These additional basis vectors can be computed by encoder and decoder without side information.

Specifically, consider the set of r orthogonal vectors in the truncated version of the transform matrix for frame f and frequency band i, denoted by $V_{f_i,SVD}$. The goal is to find other vectors orthogonal to this set, i.e. $g \in \mathcal{R}^M$ such that $g^T V_{f_i,SVD} = 0$. In other words, we seek vectors spanning the null space of $V_{f_i,SVD}$. A trivial example in 3D is illustrated in Fig.3 where a principal vector (blue) is sent to decoder, which allows two additional vectors spanning the null space to be computed. The original data is projected along these vectors and the p vectors corresponding to highest energy signal components are selected to extends the set of predominant vectors obtained by SVD and are placed in a matrix denoted by $V_{f_i,Null}$. The only additional side information is an index specifying which p of the basis vectors were selected. Thus the effective transform matrix is obtained by concatenating $V_{f_i,SVD}$ and $V_{f_i,Null}$ as $V_{f_i} = [V_{f_i,SVD} V_{f_i,Null}]$. The next step is to encode the $r + p$ predominant components which are obtained similar to Sec.2.2.

Without a good distortion measure that explicitly accounts for perceptual artifacts caused by 3D audio coding, current approaches encode all predominant and ambient sound components independently, thus neglecting inter-channel dependencies and masking effects. Leveraging the above framework where all compression is performed in the SVD domain, we propose a first "crude" framework to provide initial but strong evidence for the potential gains due to accounting for inter-channel masking effects. Specifically, we jointly calculate masking thresholds for all channels.

Let us consider the simple psychoacoustic model with a fixed signal-to-mask ratio, similar to the MPEG reference software. If we denote the energy of the i^{th} critical band by e_i, then the masking threshold for that critical band can be obtained as,

$$w_i = \begin{cases} c_i e_i & e_i > thr \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

Where c_i is a pre-defined constant and thr is a global threshold value. Unlike current approaches, we propose to use a common masking threshold for all encoded channels, calculated from the average energy of all components, in a given band. Note that while we employ a simple psychoacoustic model with a fixed signal-to-mask ratio, the underlying approach based on the average energy can be extended to more sophisticated models in a straightforward manner. Finally, there is no background signal compression, and the residual of the predominant components is converted back to the ambisonics domain, but only for the purpose of perceptual noise substitution as described in Sec. 2.2. Fig.4 depicts an overview of the proposed encoder.

4. EXPERIMENTAL RESULTS

We conducted objective and subjective tests to validate the effectiveness of the proposed approach. The following codecs are compared in our experiments:

- CMPEG: Our implementation of the MPEG-H codec as described in Sec.2.1. Due to difficulty obtaining a working current version of the MPEG-H encoder, we implemented our own representative version of it, based on published patents [10, 11] and the standard documentation [12]. Other than the explicit differentiating contributions of the competing coders, the competitors are identical in implementation, options enabled, etc.
- FSVD: Our frequency domain SVD framework proposed in [8] and described in Sec.2.2.
- PROP: The approach proposed here and described in Sec.3.

In CMPEG and FSVD the number of foreground and background components were set to 4 each, while PROP uses 4 predominant components from SVD and 4 additional components obtained by the proposed null-space technique. Thus in all competing coders a total of 8 components were encoded using the core audio codec. Two coding modes are available per frame in FSVD and PROP, one with a single band and the other with 4 frequency bands, where mode switching is performed to minimize rate. Note that the core audio encoders we used are standard compatible but not conventional, in the sense that it achieves better optimization via a trellis approach to select encoding parameters (scale factors and Huffman codebooks), as described in [13]. The test database consisted of eight third-order (16 ambisonics channels) HOA files provided by Google for UCSB research, with diverse type of audio including speech, music, singing with stationary and moving sound sources. The coders were run to minimize the maximum quantization noise to masking ratio (MNMR) criterion in all frequency bands for all encoded sound components. The coders can adjust the value of MNMR to match a given bit-rate. For both objective and subjective listening tests, HOA data were converted to stereo signals using a binaural renderer. The binaural renderer decodes HOA data to the positions of a set of loudspeakers using Max r_E [3, 4] mode-matching or L2-norm decoding techniques, and the decoded signal at each loudspeaker is convolved with the associated HRTFs for the left and right ear, and each ear's signals are added together to generate the stereo output.

4.1. Objective Results

For preliminary evaluation, we used the average quantization noise-to-mask ratio (ANMR) of final binaural reconstructions, averaged over all frames, as the distortion metric to compare the competing codecs. For a meaningful comparison in this setting, perceptual noise substitution was disabled in FSVD and PROP. The performance of the three coders is compared in Figure 5, where average distortion is plotted versus bit-rate. Distortion at a given bit-rate has been averaged over the test files, and the bit-rate range was selected to cover a wide range of reconstruction quality. It is clear that the proposed approach provides consistent coding gains, up to 4.3dB and 3.7dB over CMPEG, and FSVD, respectively.

4.2. Subjective Results

A MUSHRA [14] listening test was conducted to evaluate the perceptual gains of the proposed codec over competing methods. Ten seconds of each of the eight audio sequences were converted to stereo audio files using the binaural renderer. The following 5 versions of each of the audio items were presented in a random order to 9 listeners: the hidden reference (ref), a 3.5 kHz low passed anchor (anc), and files encoded with CMPEG, FSVD and PROP. The subjects were asked to rate each file based on audio quality on a scale of 0 to 100, where 0 corresponds to the bottom of the scale (bad quality). The bit-rates were matched at about 200 Kbps. The averaged scores over all audio items and the 95% confidence intervals are depicted in Fig.6, where the proposed scheme is demonstrated to outperform its competitors. Despite all the matching and interpolation techniques implemented in CMPEG, the reconstructed data still suffers from blocking artifacts, which could be the potential reason behind the slightly better ratings of anchor compared to CMPEG.

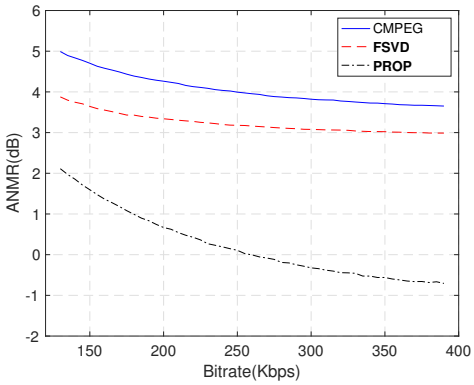

Fig. 5. Average distortion versus bit-rate of the competing coders, evaluated and averaged over the dataset

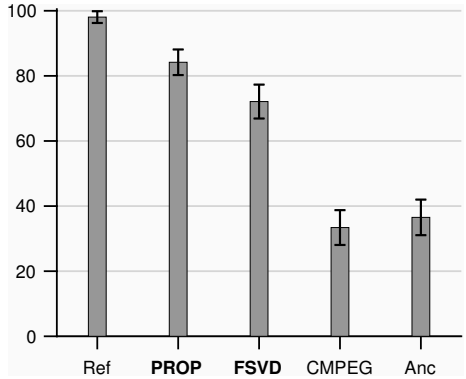

Fig. 6. MUSHRA listening test results

5. CONCLUSION

A new encoding framework for compression of HOA data is presented, where a null-space basis vector extension technique enables all compression to be performed in the SVD domain, and a jointly computed common masking threshold accounts for effects of spatial masking across components. Significant gains over existing approaches demonstrate the effectiveness of the proposed framework.

6. ACKNOWLEDGMENT

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

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.A 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] J. Daniel, J.-B. Rault, and J.-D. Polack, "Ambisonics encoding of other audio formats for multiple listening conditions," in *Audio Engineering Society Convention 105*, Sep 1998.

[5] 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.

[6] R. L. Bleidt, D. Sen, A. Niedermeier, B. Czelhan, S. Fg, S. Disch, J. Herre, J. Hilpert, M. Neuendorf, H. Fuchs, J. Issing, A. Murtaza, A. Kuntz, M. Kratschmer, F. Kch, R. Fg, B. Schubert, S. Dick, G. Fuchs, F. Schuh, E. Burdiel, N. Peters, and M. Y. Kim, "Development of the mpeg-h tv audio system for atsc 3.0," *IEEE Transactions on Broadcasting*, vol. 63, no. 1, pp. 202–236, March 2017.

[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] S. Zamani, T. Nanjundaswamy, and K. Rose, "Frequency domain singular value decomposition for efficient spatial audio coding," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2017.

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

[13] A. Aggarwal, S. L. Regunathan, and K. Rose, "A trellis-based optimal parameter value selection for audio coding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 623–633, March 2006.

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