# DECORRELATION FOR AUDIO OBJECT CODING

*Lars Villemoes, Toni Hirvonen, and Heiko Purnhagen*

Dolby Sweden AB, Stockholm, Sweden

## ABSTRACT

Object-based representations of audio content are increasingly used in entertainment systems to deliver immersive and personalized experiences. Efficient storage and transmission of such content can be achieved by joint object coding algorithms that convey a reduced number of downmix signals together with parametric side information that enables object reconstruction in the decoder. This paper presents an approach to improve the performance of joint object coding by adding one or more decorrelators to the decoding process. Listening test results illustrate the performance as a function of the number of decorrelators. The method is adopted as part of the Dolby AC-4 system standardized by ETSI.

***Index Terms*—** Audio coding, Object-based audio, Decorrelation, Joint object coding

## 1. INTRODUCTION

The object-based representation of immersive audio content is a powerful approach that combines intuitive content creation with optimal reproduction over a large range of playback configurations using suitable rendering systems and also enables personalized experiences [1]. In such a representation, an object comprises both the audio waveform itself as well as dynamic object metadata, conveying e.g. its spatial position. This representation enables optimal reproduction over playback systems ranging from immersive loudspeaker configurations (for example a 7.1.4 configuration with a 7.1 setup in the horizontal plane and 4 ceiling speakers), over binaural playback on headphones, to legacy configurations like 5.1 and 2-channel stereo. To enable the delivery of object-based audio content to consumer entertainment systems, an efficient representation of the immersive audio content is required to facilitate storage and transmission at low bit rates. A common approach to achieve this is to convey $N$ objects by means of a reduced number $M < N$ of downmix signals together with parametric side information that enables object reconstruction in the decoder. Such algorithms are, for example, used in the MPEG Spatial Audio Object Coding (SAOC) system [2] and the Joint Object Coding (JOC) tool [3].

Considering channel-based audio content, similar approaches that convey the audio channels using a lower number of downmix signals are well established and can be referred to as parametric spatial audio coding algorithms [4, 5]. The basic example of this approach is a Parametric Stereo (PS) system where a 2-channel stereo signal is conveyed by means of a mono downmix channel and parametric side information. Compared to earlier approaches like Intensity Stereo (IS) [6], the introduction of a decorrelator in the decoder resulted in a significant improvement of the performance [7, 8], since it improves the capabilty of the decoder to re-instantiate perceptually important cues like ambience or source width. Decorrelators are also used in parametric spatial coding systems for 5.1 surround content [9, 10] and channel-based immersive content [11].

The introduction of decorrelators in a parametric coding system for object-based content, however, turns out to be more challenging than in the channel-based case, in particular if a low number of decorrelators is desired in order to minimize the computational complexity of the decoder. Moreover, it is desirable to add decorrelation directly in the object domain, since this enables object decoding and object rendering to take place in two separate devices. The MPEG-H SAOC-3D system [12, 13], on the other hand, adds decorrelation as part of the rendering.

This paper is structured as follows. First, a general framework for joint object coding is described. Then, the approach to add decorrelators in this framework is presented. Finally, listening test results are presented and conclusions are drawn.

## 2. JOINT OBJECT CODING FRAMEWORK

A generalized Joint Object Coding (JOC) framework was presented in [3] and is briefly summarized here.

Input to the encoder is content in an object-based immersive representation, comprising waveforms of the $N$ object signals and the associated object metadata. First, an $M$-channel downmix of the immersive content is generated by a downmix renderer, governed by the spatial positions conveyed as object metadata. Based on the object and downmix signals, JOC parameters are computed that enable an approximate reconstruction of the audio objects from the downmix in the decoder. Finally, perceptual audio coding algorithms are used to convey the downmix itself at a low bit rate. The JOC parameters as well as the object metadata are included as side information in the bitstream.

In the decoder, first the downmix signals are decoded.

Next, the JOC parameters are used to generate an approximate reconstruction of the object signals. Finally, the reconstructed object signals together with the associated object metadata are processed by a renderer to generate a presentation suitable for the playback configuration available at the decoder side.

The JOC parameters for object reconstruction are computed and applied in a time- and frequency-varying manner, enabled by a perceptually motivated separation of the object and downmix signals into a set of non-uniform frequency bands and a temporal framing. The intersection of a frequency band and a temporal frame can be referred to as a time-frequency tile and JOC parameters are computed for each tile.

Let $\mathbf{X}$ be the time-frequency tile of the input object signals in a given filterbank or transform domain where each row of $\mathbf{X}$ contains all samples for one object signal in that tile. Define $\mathbf{Y}$ similarly for the downmix signals.

The basic object reconstruction, i.e., upmix, at the decoder without decorrelation is given by the linear combination

$$\widehat{\mathbf{X}} = \mathbf{C}\mathbf{Y}. \tag{1}$$

The coefficient matrix $\mathbf{C}$ is of size $N \times M$ and assumed here to be real valued. With the definition of the sample covariance matrices

$$\mathbf{R}_{\mathrm{uv}} = \mathrm{Re}(\mathbf{U}\mathbf{V}^*), \tag{2}$$

it is easy to show that the least squares problem for minimizing the total waveform error between $\widehat{\mathbf{X}}$ and $\mathbf{X}$ has normal equations $\mathbf{C}\mathbf{R}_{\mathrm{yy}} = \mathbf{R}_{\mathrm{xy}}$. In order to deal with the potential singularity of $\mathbf{R}_{\mathrm{yy}}$, one can use a regularized solution such as

$$\mathbf{C} = \mathbf{R}_{\mathrm{xy}}(\mathbf{R}_{\mathrm{yy}} + \varepsilon\mathbf{I})^{-1}, \tag{3}$$

where $\varepsilon \geq 0$ is a regularization constant and $\mathbf{I}$ is the identity matrix of size $M$.

For challenging material, this object reconstruction method can produce artifacts. The synthesized covariance $\widehat{\mathbf{R}}_{\mathrm{xx}} = \mathbf{C}\mathbf{R}_{\mathrm{yy}}\mathbf{C}^{\mathrm{T}}$ has at most rank $M < N$, indicating that there can be more correlation between the decoded objects than there was between the original objects, which can manifest itself as a spatial "collapse" of the rendered sound scene. Also, since faithful reproduction of the energy of any linear combinations of object signals in a time-frequency tile is equivalent to $\widehat{\mathbf{R}}_{\mathrm{xx}} = \mathbf{R}_{\mathrm{xx}}$, the linear combinations in the rendering process can give rise to timbral distortion due to frequency dependent energy errors if full covariance reconstruction is not achieved.

## 3. OBJECT DECORRELATION

### 3.1. Synthesis model

In order to better approximate the covariance structure of the $N$ object signals, a decorrelated signal contribution is added to the upmix. Examples of decorrelator implementations can be found in [5] and [7]. We assume that $K$ decorrelators are

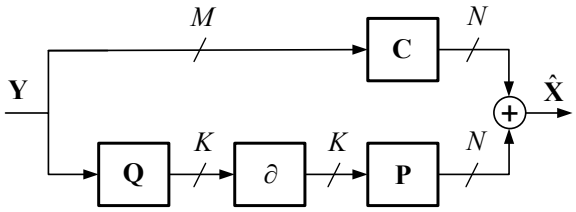


**Fig. 1**. Block diagram of the upmix with decorrelation within one time-frequency tile. The $\partial$ block performs separate decorrelation of $K$ signals and $\mathbf{C}$, $\mathbf{P}$ ,$\mathbf{Q}$ are real matrices.

running and use the notation of a vector valued decorrelator $\partial(\cdot)$ which operates with a separate decorrelator on each of its $K$ input signals. These input signals are obtained as linear combinations of the $M$ downmix signals with weights from a pre-decorrelator matrix $\mathbf{Q}$ of size $K \times M$. The $K$ separately decorrelated outputs are then upmixed and added to the $N$ object signals by means of a decorrelator upmix matrix $\mathbf{P}$ of size $N \times K$. The complete synthesis is illustrated in Fig. 1 and can be written as follows,

$$\widehat{\mathbf{X}} = \mathbf{C}\mathbf{Y} + \mathbf{P}\,\partial\,(\mathbf{Q}\mathbf{Y})\,. \tag{4}$$

For the sake of the analysis, we define a decorrelator as a 2-norm preserving operator whose output is orthogonal to all other signals under consideration. Moreover, it is assumed that the $K$ decorrelators have pairwise orthogonal outputs. It follows that the covariance of $\partial\,(\mathbf{Q}\mathbf{Y})$ is the diagonal part $\mathbf{\Lambda}$ of $\mathbf{Q}\mathbf{R}_{\mathrm{yy}}\mathbf{Q}^{\mathrm{T}}$, and the synthesized covariance is

$$\widehat{\mathbf{R}}_{\mathrm{xx}} = \mathbf{C}\mathbf{R}_{\mathrm{yy}}\mathbf{C}^{\mathrm{T}} + \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathrm{T}}, \tag{5}$$

where $\mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathrm{T}}$ is the contribution from decorrelation. If $\mathbf{\Lambda}$ is non-singular, this contribution can be matched to a target positive semidefinite covariance $\mathbf{R}_{\mathrm{w}}$ of rank at most $K$ by factorizing $\mathbf{R}_{\mathrm{w}} = \mathbf{V}\mathbf{V}^{\mathrm{T}}$ with a $N \times K$ matrix $\mathbf{V}$ and setting

$$\mathbf{P} = \mathbf{V}\mathbf{\Lambda}^{-1/2}. \tag{6}$$

### 3.2. Example encoding algorithms

Even when the number of downmix channels $M$ is given, the number of decorrelators $K$ and all the parameters in $\mathbf{C}$ and $\mathbf{P}$ can be selected in a huge number of ways, resulting in different balances between waveform and covariance reconstruction.

In the cascaded approach, $\mathbf{C}$ is selected according to (3). It is then easy to show that the covariance error $\mathbf{\Delta R} = \mathbf{R}_{\mathrm{xx}} - \mathbf{C}\mathbf{R}_{\mathrm{yy}}\mathbf{C}^{\mathrm{T}}$ is positive semidefinite. If the rank of $\mathbf{\Delta R}$ is not larger than the desired number of decorrelators $K$, then the method of the previous section can be used directly with $\mathbf{R}_{\mathrm{w}} = \mathbf{\Delta R}$ to obtain full covariance reconstruction, $\widehat{\mathbf{R}}_{\mathrm{xx}} = \mathbf{R}_{\mathrm{xx}}$. If $\mathbf{R}_{\mathrm{yy}}$ is non-singular and the downmix is

created by a fixed matrix, $\mathbf{Y} = \mathbf{DX}$, then the rank of $\mathbf{\Delta R}$ is at most $N - M$. See [10] for proof in the channel-based case.

A different mathematically appealing approach in the spirit of [14] is to aim for least squares waveform error subject to full covariance reconstruction. A closed form solution to this problem, leading to a target $\mathbf{R}_{\mathrm{w}}$ of rank at most $N-M$, is given in the appendix.

In both cases, $K = N - M$ decorrelators can be used to achieve full covariance reconstruction while maintaining optimal waveform approximation properties. For a lower number of decorrelators, it is natural to form an approximation to the target $\mathbf{R}_{\mathrm{w}}$ by using the eigenvectors corresponding to its $K$ largest eigenvalues.

When full covariance reconstruction is abandoned, a gain compensation can be used to recover at least a total energy match between $\widehat{\mathbf{X}}$ and $\mathbf{X}$ in each time-frequency tile.

### 3.3. Choice of pre-decorrelator matrix

The purely covariance-based model of decorrelation would allow us to use an arbitrary fixed pre-decorrelator matrix $\mathbf{Q}$. However, the output of an actual decorrelator resembles its input and it is undesirable to add a decorrelated version of one object to a different object. In channel-based upmix systems such as MPEG Surround [9], the fixed architecture of the downmix process makes it easy to avoid the corresponding spatial leakage, whereas the object position dependence of the downmix in our case calls for a generalized guiding principle:

> *If a decorrelator output signal contributes with a large amount to an object signal, the input to that decorrelator should receive a large contribution from an approximation of that object signal.*

A simple way to obey the principle is to let each reconstructed object be a linear combination of an approximation of the original object and a decorrelated version of that approximation. Assuming $\mathbf{CY}$ is a good approximation of the object signals, this situation is obtained by setting $\mathbf{Q} = \mathbf{C}$ and using a diagonal $\mathbf{P}$. The simplicity of this solution turns out to be a challenge for the encoding process since it can only furnish a positive diagonal decorrelator covariance contribution $\mathbf{P\Lambda P}^{\mathrm{T}}$ in (5). Another significant drawback is that it cannot accommodate the upmix examples using fewer than $N$ decorrelators which were derived in Sec. 3.2.

A rule that adapts to a given number $K$ of decorrelators and follows the guiding principle is given by $\mathbf{Q} = \mathbf{P}^{\mathrm{T}}\mathbf{C}$. Here, each decorrelator, whose contribution to the object signals is distributed by the weights in a column of $\mathbf{P}$, will be fed by a linear combination of the estimated object signals with weights equal to these upmix weights. In order to reduce the risk of cancellations, we replace $\mathbf{P}$ with its element-wise absolute value $|\mathbf{P}|$ and settle for the design choice

$$\mathbf{Q} = |\mathbf{P}|^{\mathrm{T}}\,\mathbf{C}. \tag{7}$$

Note that this pre-decorrelator matrix $\mathbf{Q}$ can be derived from $\mathbf{C}$ and $\mathbf{P}$ in the decoder.

The simultaneous use of $\mathbf{P}$ for pre- and post-decorrelator mixing requires some thought in the adjustment of the decorrelator contribution to a given target $\mathbf{R}_{\mathrm{w}} = \mathbf{VV}^{\mathrm{T}}$ in (6). Fortunately, it is easy to show that if $\mathbf{B}$ is the diagonal part of $|\mathbf{V}|^{\mathrm{T}}\,\mathbf{CR}_{\mathrm{yy}}\mathbf{C}^{\mathrm{T}}\,|\mathbf{V}|$, then using $\mathbf{\Lambda} = \mathbf{B}^{1/2}$ in (6) is consistent with the definition of $\mathbf{\Lambda}$ as the diagonal part of $\mathbf{QR}_{\mathrm{yy}}\mathbf{Q}^{\mathrm{T}}$.

## 4. RESULTS

To study the performance of the proposed method as a function of the number $K$ of decorrelators, a MUSHRA listening test [15] was conducted for a set of 13 critical test items of object-based immersive content described in Tab. 1 of [3]. To focus on the effect of the object reconstruction method in Fig. 1, some aspects required for a practical coding system were omitted. In particular, the downmix signals were directly sent to the decoder without perceptual audio coding. The content was processed by an immersive interchange translation process [1] to generate $N = 7$ objects as input to the encoder. The encoder then generated an adaptive downmix with $M = 3$ downmix channels as described in Sec. 2.4 of [3]. Four different joint object coding configurations were compared in the test, using no decorrelation, or $K = 1$, 2, or 4 decorrelators. The cascaded approach of Sec. 3.2 was used including total energy match gain compensation. The parameters $\mathbf{C}$ and $\mathbf{P}$ sent as side information were quantized and coded. The average side information rate for the particular choice of parameter bands and temporal framing for these parameters was 8, 11, 14, and 20 kb/s for the configurations with 0, 1, 2, and 4 decorrelators, respectively.

The 7 objects reconstructed by the decoder were rendered for playback on a 7.1.4 immersive speaker configuration, and a 7.1.4 rendering of the 7 input objects was used as the open and hidden reference in the listening test and to generate the 3.5 kHz lowpass anchor. The test results of 10 subjects which passed pre- and post-screening are shown in Fig. 2.

On average over all items, the system with $K = 1$ decorrelator performs significantly better than the system without decorrelation ($K = 0$). Increasing the number of decorrelators beyond $K = 1$ did not result in any further improvement of the average performance in this experiment. The system with 4 decorrelators even tends to perform on average slightly worse that the systems with 1 or 2 decorrelators. A possible explanation is that actual decorrelators can generate artifacts. Hence, a partial reconstruction of the object covariance might sound better than a full reconstruction that requires more decorrelator contributions.

Studying the per-item results, it can be seen that items dominated by wide and ambient sounds like jungle ambience (item 6) and rainfall (item 8) show the largest improvements when decorrelators are added. Note also that for some items (like item 5, dominated by dialog with strong cave reverbera-

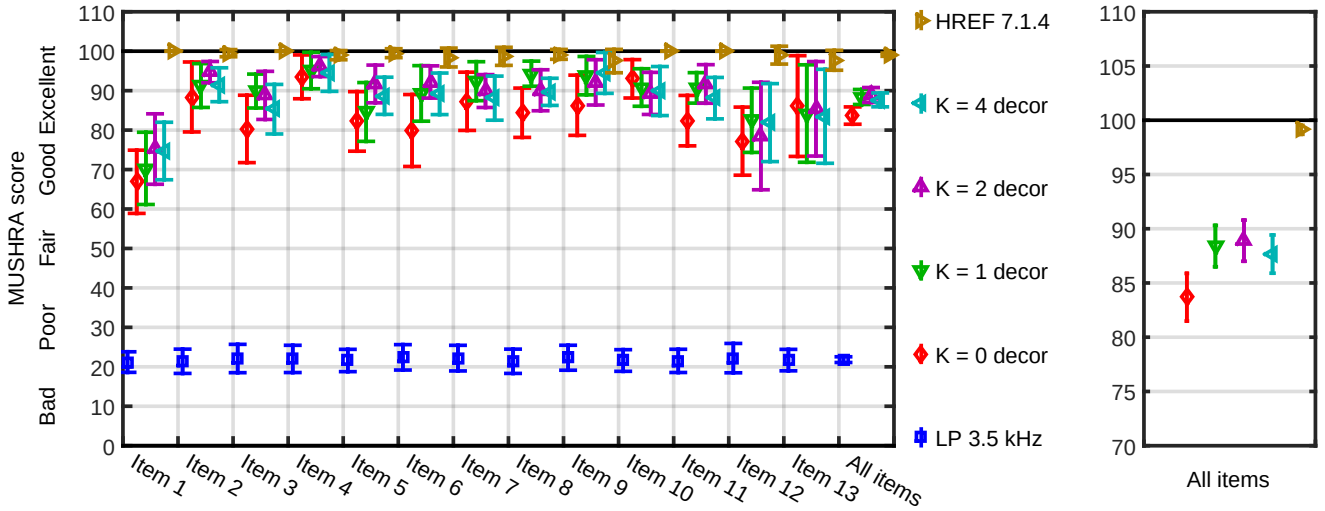


**Fig. 2**. MUSRHA listening test results for joint object coding of $N = 7$ objects using $M = 3$ uncoded downmix channels with $K = 0$, 1, 2, and 4 decorrelators for 13 critical items rendered for playback on an immersive 7.1.4 loudspeaker configuration showing mean and 95% CI for 10 subjects. The right panel shows the average over all 13 items on an enlarged scale.

tion), the system with 2 decorrelators performs clearly better than the system with 1 decorrelator.

The object decorrelation method presented here has also been tested in a real-world audio codec where, depending on target bit rate, 11 to 15 objects were conveyed using 1 to 3 decorrelators in the decoder. The MUSHRA listening test results in Sec. 4 of [3] show the performance for critical object-based immersive test items rendered for playback on an immersive 7.1.4 loudspeaker configuration. On the MUSHRA scale, excellent quality is achieved at a total rate of 384 kb/s, and even at 192 kb/s, the quality is in the upper half of the good-range.

## 5. CONCLUSIONS

This paper introduced an approach to improve the performance of joint audio object coding by adding one or more decorrelators to the decoding process. A theoretical basis was presented and listening test results indicate that already the introduction of a single decorrelator can result in a significant improvement of the subjective quality. This study presents to our knowledge the first attempt of adding decorrelation directly in the object domain. The method is part of the advanced joint object coding tool in the Dolby AC-4 system standardized by ETSI [11, 16].

## 6. APPENDIX

A claim from Sec. 3.2 is proven here. Let $\mathrm{tr}\{\mathbf{M}\}$ be the trace of a square matrix $\mathbf{M}$ and let $\mathbf{R}^{1/2}$ denote the unique positive definite square root of a positive definite matrix $\mathbf{R}$.

**Theorem 1.** *Assume that the covariances $\mathbf{R}_{\mathrm{xx}}, \mathbf{R}_{\mathrm{xy}}, \mathbf{R}_{\mathrm{yy}}$ are all of full rank. Then a solution to $\widehat{\mathbf{R}}_{\mathrm{xx}} = \mathbf{R}_{\mathrm{xx}}$ minimizing the Frobenius norm $\|\widehat{\mathbf{X}} - \mathbf{X}\|_{\mathrm{F}}^2$ is obtained by using*

$$\mathbf{C} = \mathbf{R}_{\mathrm{xx}}^{1/2}\mathbf{F}\mathbf{R}_{\mathrm{yy}}^{-1/2} \tag{8}$$

*where $\mathbf{F} = \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1/2}$ and $\mathbf{A} = \mathbf{R}_{\mathrm{xx}}^{1/2}\mathbf{R}_{\mathrm{xy}}\mathbf{R}_{\mathrm{yy}}^{-1/2}$.*

*Furthermore, the required decorrelator contribution $\mathbf{R}_{\mathrm{w}} = \mathbf{R}_{\mathrm{xx}} - \mathbf{C}\mathbf{R}_{\mathrm{yy}}\mathbf{C}^{\mathrm{T}}$ has rank at most $N - M$.*

*Proof.* One observes that a full covariance reconstruction $\widehat{\mathbf{R}}_{\mathrm{xx}} = \mathbf{R}_{\mathrm{xx}}$ implies that the total reconstruction error can be simplified to $\|\widehat{\mathbf{X}} - \mathbf{X}\|_{\mathrm{F}}^2 = \mathrm{tr}\{(\widehat{\mathbf{X}} - \mathbf{X})(\widehat{\mathbf{X}} - \mathbf{X})^*\} = 2\,\mathrm{tr}\{\mathbf{R}_{\mathrm{xx}}\} - 2\,\mathrm{tr}\{\mathbf{C}\mathbf{R}_{\mathrm{yx}}\}$.

By defining $\widetilde{\mathbf{Y}} = \mathbf{R}_{\mathrm{yy}}^{-1/2}\mathbf{Y}$, $\widetilde{\mathbf{X}} = \mathbf{R}_{\mathrm{xx}}^{-1/2}\mathbf{X}$, and $\widetilde{\mathbf{C}} = \mathbf{R}_{\mathrm{xx}}^{-1/2}\mathbf{C}\mathbf{R}_{\mathrm{yy}}^{1/2}$, the problem transforms into maximizing $\mathrm{tr}\{\mathbf{C}\mathbf{R}_{\mathrm{yx}}\} = \mathrm{tr}\{\widetilde{\mathbf{C}}^{\mathrm{T}}\mathbf{A}\}$ under the constraint $\widetilde{\mathbf{R}}_{\mathrm{w}} = \mathbf{I} - \widetilde{\mathbf{C}}\widetilde{\mathbf{C}}^{\mathrm{T}} \geq 0$. The latter condition is equivalent to the operator 2-norm bound $\|\widetilde{\mathbf{C}}\|_2 \leq 1$. To conclude the proof of (8) it is sufficient to show that $\widetilde{\mathbf{C}} = \mathbf{F}$ solves this problem.

To see this, the central tool is Hölder's inequality for Schatten norms [17], which gives that

$$\mathrm{tr}\{\widetilde{\mathbf{C}}^{\mathrm{T}}\mathbf{A}\} \leq \|\widetilde{\mathbf{C}}\|_2\,\mathrm{tr}\{(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{1/2}\}. \tag{9}$$

With $\|\widetilde{\mathbf{C}}\|_2 \leq 1$, it follows that $\mathrm{tr}\{\widetilde{\mathbf{C}}^{\mathrm{T}}\mathbf{A}\} \leq \mathrm{tr}\{(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{1/2}\}$ and it is easy to check that this upper bound is achieved by $\widetilde{\mathbf{C}} = \mathbf{F}$, which satisfies $\|\mathbf{F}\|_2 \leq 1$ since $\mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}$.

Finally, the latter identity also implies that $\widetilde{\mathbf{R}}_{\mathrm{w}} = \mathbf{I} - \mathbf{F}\mathbf{F}^{\mathrm{T}} \geq 0$, and since $\mathbf{F}\mathbf{F}^{\mathrm{T}}\mathbf{A} = \mathbf{A}$, the $M$-dimensional range of $\mathbf{A}$ is in the null-space of $\widetilde{\mathbf{R}}_{\mathrm{w}}$, which implies the desired rank bound since $\mathbf{R}_{\mathrm{w}} = \mathbf{R}_{\mathrm{xx}}^{1/2}\widetilde{\mathbf{R}}_{\mathrm{w}}\mathbf{R}_{\mathrm{xx}}^{1/2}$. □

## 7. REFERENCES


[1] Jeffrey Riedmiller, Sripal Mehta, Nicolas Tsingos, and Prinyar Boon, "Immersive and personalized audio: A practical system for enabling interchange, distribution, and delivery of next-generation audio experiences," *Motion Imaging Journal, SMPTE*, vol. 124, no. 5, pp. 1–23, July 2015.

[2] Jürgen Herre, Heiko Purnhagen, Jeroen Koppens, Oliver Hellmuth, Jonas Engdegård, Johannes Hilper, Lars Villemoes, Leon Terentiv, Cornelia Falch, Andreas Hölzer, María Luis Valero, Barbara Resch, Harald Mundt, and Hyen-O Oh, "MPEG Spatial Audio Object Coding — the ISO/MPEG standard for efficient coding of interactive audio scenes," *J. Audio Eng. Soc*, vol. 60, no. 9, pp. 655–673, 2012.

[3] Heiko Purnhagen, Toni Hirvonen, Lars Villemoes, Jonas Samuelsson, and Janusz Klejsa, "Immersive audio delivery using joint object coding," in *Audio Engineering Society Convention 140*, May 2016.

[4] Christof Faller and Frank Baumgarte, "Binaural cue coding: A novel and efficient representation of spatial audio," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, May 2002.

[5] Jeroen Breebaart, Sascha Disch, Christof Faller, Jürgen Herre, Johannes Hilpert, Kristofer Kjörling, Francois Myburg, Heiko Purnhagen, and Erik Schuijers, "The reference model architecture for MPEG Spatial Audio Coding," in *Audio Engineering Society Convention 118*, May 2005.

[6] Robbert G. van der Waal and Raymond N.J. Veldhuis, "Subband coding of stereophonic digital audio signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Apr. 1991.

[7] Heiko Purnhagen, Jonas Engdegård, Jonas Rödén, and Lars Liljeryd, "Synthetic ambience in parametric stereo coding," in *Audio Engineering Society Convention 116*, May 2004.

[8] Heiko Purnhagen, "Low complexity parametric stereo coding in MPEG-4," in *Proc. Digital Audio Effects Workshop (DAFX)*, Oct. 2004.

[9] Jürgen Herre, Kristofer Kjörling, Jeroen Breebaart, Christof Faller, Sascha Disch, Heiko Purnhagen, Jeroen Koppens, Johannes Hilpert, Jonas Rödén, Werner Oomen, Karsten Linzmeier, and Kok Seng Chong, "MPEG Surround — the ISO/MPEG standard for efficient and compatible multichannel audio coding," *J. Audio Eng. Soc*, vol. 56, no. 11, pp. 932–955, 2008.

[10] Gerard Hotho, Lars Villemoes, and Jeroen Breebaart, "A backward-compatible multichannel audio codec," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 83–93, Jan. 2008.

[11] Kristofer Kjörling, Jonas Rödén, Martin Wolters, Jeff Riedmiller, Arijit Biswas, Per Ekstrand, Alexander Gröschel, Per Hedelin, Toni Hirvonen, Holger Hörich, Janusz Klejsa, Jeroen Koppens, Kurt Krauss, Heidi-Maria Lehtonen, Karsten Linzmeier, Hannes Muesch, Harald Mundt, Scott Norcross, Jens Popp, Heiko Purnhagen, Jonas Samuelsson, Michael Schug, Leif Sehlström, Robin Thesing, Lars Villemoes, and Mark Vinton, "AC-4 — the next generation audio codec," in *Audio Engineering Society Convention 140*, May 2016.

[12] Adrian Murtaza, Jürgen Herre, Jouni Paulus, Leon Terentiv, Harald Fuchs, and Sascha Disch, "ISO/MPEG-H 3D Audio: SAOC 3D decoding and rendering," in *Audio Engineering Society Convention 139*, Oct. 2015.

[13] Jürgen Herre, Johannes Hilpert, Achim Kuntz, and Jan Plogsties, "MPEG-H Audio — the new standard for universal spatial/3D audio coding," *J. Audio Eng. Soc*, vol. 62, no. 12, pp. 821–830, 2015.

[14] Juha Vilkamo, Tom Bäckström, and Achim Kuntz, "Optimized covariance domain framework for time-frequency processing of spatial audio," *J. Audio Eng. Soc*, vol. 61, no. 6, pp. 403–411, 2013.

[15] "Method for the subjective assessment of intermediate quality levels of coding systems," Recommendation ITU-R BS.1534-3, Oct. 2015.

[16] "Digital audio compression (AC-4) standard; part 2: Immersive and personalized audio," ETSI TS 103 190-2 V1.1.1, Sept. 2015.

[17] John Watrous, "Theory of quantum information, 2.3 norms of operators," 2011, `https://cs.uwaterloo.ca/~watrous/CS766/LectureNotes/02.pdf`.