# A FREQUENCY-DOMAIN ALGORITHM TO UPSCALE AMBISONIC SOUND SCENES

*Andrew Wabnitz, Nicolas Epain, Craig T. Jin*

Computing and Audio Research Laboratory (CARLab)
School of Electrical and Information Engineering
The University of Sydney, NSW 2006, Australia

## ABSTRACT

In this paper, a novel algorithm for upscaling ambisonic sound scenes in the frequency domain is presented. This algorithm makes use of compressed sensing techniques to calculate a set of upscaling filters. These filters are then used to increase the spherical harmonic order of a set of ambisonic sound signals to higher orders. Upscaled ambisonic sound scenes have a greater spatial resolution, which allows more loudspeakers to be used during the playback, resulting in a larger sweet spot and improved sound quality. A formal listening test was conducted to evaluate the perceptual quality of sound fields reproduced using this technique. Results show that the proposed algorithm significantly improves the perceptual fidelity of the sound field reproduction, in comparison to classical ambisonic methods.

***Index Terms***— Acoustic Signal Processing, Compressed Sensing, Signal Reconstruction, Audio Recording

## 1. INTRODUCTION

Higher order ambisonics (HOA) [1] is a popular spatial sound field reproduction technique based on spherical acoustics. In the HOA framework, the sound scene is represented as a set of signals associated with each spherical harmonic component of the sound field up to a given truncation order $L$. These signals are referred to as order-$L$ HOA signals. The order of the HOA signals directly influences the spatial resolution of the captured sound scene as well as the size of the region where the recorded sound field may accurately be reproduced with an array of loudspeakers. This region is commonly referred to as the *sweet spot*.

Typically a multichannel microphone array, such as the spherical microphone array [2], is used to acquire the HOA signals. Physical and practical limitations in the array acquisition system, such as the number of microphone sensors, limit the maximum HOA order at which the sound scene may be acquired. Hence given a certain microphone array, there are limitations on the spatial resolution and accuracy at which the sound field is reproduced.

Recent studies [3, 4, 5] have investigated the application of compressed sensing (CS) [6] techniques in order to overcome these limitations. In [5], a time-domain algorithm to upscale HOA sound scenes to higher orders was presented. Upscaling a HOA sound scene increases its spatial resolution, which results in a larger sweet spot and improved sound quality when the sound field is reproduced. The work presented in this paper aims to extend this CS-based HOA upscaling approach, with the development of a frequency-domain upscaling algorithm. This new algorithm is presented in Section 2. A formal listening test was conducted to evaluate the perceptual quality of the sound field reproduced when using the frequency- and time-domain HOA upscaling algorithms. The results of this listening test are presented in Section 3.

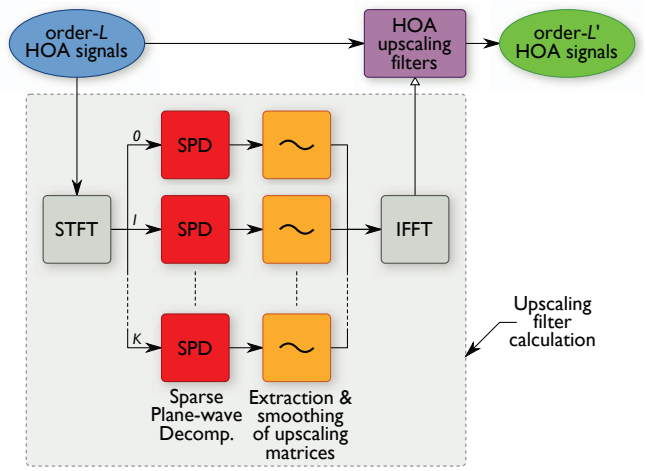


**Fig. 1**. This flowchart shows an overview of the frequency-domain HOA sound scene upscaling algorithm.

## 2. FREQUENCY DOMAIN HOA UPSCALING ALGORITHM

In this section, the proposed frequency domain HOA upscaling algorithm is described. This algorithm is summarized in Figure 1.

### 2.1. Short-Time Fourier Analysis

We start with a vector of order-$L$ time-domain HOA signals and apply a Short-Time Fourier Transform (STFT) to obtain a time-frequency tiling of the HOA signals with time index $t$ and frequency bin $k$. In other words, we have:

$$\mathbf{b}(t,k) = [b_0^0(t,k), b_1^{-1}(t,k), \ \ldots, \ b_l^m(t,k), \ \ldots, \ b_L^L(t,k)]^\top, \tag{1}$$

where $\mathbf{b}(t,k)$ is the vector of the order-$L$ HOA signals in the STFT domain, the length of the STFT analysis window is $2K$ and $b_l^m(t,k)$ is the STFT-domain HOA signal corresponding to order $l$ and degree $m$, with $m \in [-l, \ldots, l]$.

### 2.2. Sparse Plane-Wave Decomposition

We now describe the sparse plane-wave decomposition. For each frequency bin and time index, we aim to calculate a vector of $P$ plane-wave signals, $\mathbf{s}(t,k)$, such that:

$$\mathbf{b}(t,k) = \mathbf{Y}\mathbf{s}(t,k) \quad , \tag{2}$$

where

$$\mathbf{s}(t,k) = [s_1(t,k),\ s_2(t,k), \ldots,\ s_P(t,k)]^\top \tag{3}$$

and $\mathbf{Y}$ is the spatial dictionary used for the plane-wave decomposition. Matrix $\mathbf{Y}$ represents the contribution of each plane-wave source to the HOA signals and is given by:

$$\mathbf{Y} = \begin{bmatrix} Y_0^0(\theta_1,\phi_1) & Y_0^0(\theta_2,\phi_2) & \ldots & Y_0^0(\theta_P,\phi_P) \\ Y_1^{-1}(\theta_1,\phi_1) & Y_1^{-1}(\theta_2,\phi_2) & \ldots & Y_1^{-1}(\theta_P,\phi_P) \\ \vdots & \vdots & \ddots & \vdots \\ Y_L^L(\theta_1,\phi_1) & Y_L^L(\theta_2,\phi_2) & \ldots & Y_L^L(\theta_P,\phi_P) \end{bmatrix}, \tag{4}$$

with $Y_l^m(\theta,\phi)$ being the spherical harmonic function of order $l$ and degree $m$, and $(\theta_i,\phi_i)$ is the azimuth and elevation of the $i^{th}$ plane-wave source. The plane-wave directions are regularly distributed over a sphere, with the directions obtained by recursive subdivision of an icosahedron. The size of the dictionary is chosen such that the number of plane-waves is much greater than the number of order-$L$ HOA signals, *i.e.* $P \gg (L+1)^2$. The large dictionary size is chosen such that the order-$L$ HOA signals may be upscaled to a higher order $L'$. For this super-resolution operation to be successful, it is necessary for the number of plane-waves in the basis to be greater than the number of order-$L'$ HOA signals, *i.e.* $P \geq (L'+1)^2$.

With $P > (L+1)^2$, the linear system (2) is underdetermined and therefore contains infinitely many solutions. The classical method to solve (2) is to use the Moore-Penrose pseudo-inverse, giving the least-norm solution. However, this solution is not necessarily perceptually optimal [7]. This can be attributed to the fact that, out of all the possible plane-wave decompositions that mathematically explain the observed sound field, the least-norm solution is the one that distributes the sound energy the most evenly across space [3]. This is not a physically reasonable objective, given that most sound sources are discretely located in space. Moreover, it is reasonable to assume that the sound field consists of only a small number of active plane-wave sources at any instance in time. Therefore an alternative approach is one inspired by CS theory, which selects the solution that is the most sparsely distributed in space.

Applying CS formalism to (2), $\mathbf{Y}$ is referred to as the 'measurement matrix'. For HOA signals where the truncation order is quite low, this matrix is reasonably coherent and therefore is a poor candidate to satisfy the Restricted Isometry Property (RIP), which is required to guarantee perfect reconstruction [6]. Nonetheless, our aim is not perfect reconstruction, but rather applying super-resolution analysis that improves the spatial sound field reconstruction. In other words, given a sparse plane-wave sound field, the reconstruction should be better than what is provided by the least-norm solution.

A naïve approach to performing frequency-domain sparse plane-wave decomposition would be to solve (2) using a single-measurement vector type convex optimization problem for each frequency bin, as described by:

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{s}(t,k)\|_1 \\ &\text{subject to} \quad \mathbf{Y}\mathbf{s}(t,k) = \mathbf{b}(t,k) \quad , \end{aligned} \tag{5}$$

where $\|.\|$ denotes the $l_1$-norm. However, this approach is vulnerable to discontinuities that may arise between neighbouring frequency bins, which results in audible artefacts when playing back the upscaled HOA sound scene over an array of loudspeakers or headphones. The approach taken in this paper to minimise this vulnerability is to solve a multiple-measurement vector (MMV) type convex optimization problem for each frequency bin, *i.e.*:

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{S}(n,k)\|_{12} \\ &\text{subject to} \quad \mathbf{Y}\mathbf{S}(n,k) = \mathbf{B}(n,k) \quad . \end{aligned} \tag{6}$$

where:

- $\mathbf{B}(n,k)$ is the matrix containing the $T$ consecutive STFT samples of the HOA signals for frequency $k$ and time window $n$:

$$\mathbf{B}(n,k) = [\mathbf{b}(n\tau{+}1,k)\,,\, \mathbf{b}(n\tau{+}2,k)\,,\, \ldots,\, \mathbf{b}(n\tau{+}T,k)] \qquad n \in \mathbb{N}\ , \tag{7}$$

  with $\tau$ being the increment between successive analysis windows (typically $\tau = 0.5T$).

- $\mathbf{S}(n,k)$ is the matrix of the plane-wave signals:

$$\mathbf{S}(n,k) = [\mathbf{s}(n\tau{+}1,k)\,,\, \mathbf{s}(n\tau{+}2,k)\,, \ldots, \mathbf{s}(n\tau{+}T,k)]\ . \tag{8}$$

- $\|.\|_{12}$ denotes the $l_{12}$-norm, which is defined as:

$$\|\mathbf{A}\|_{12} \triangleq \sum_i \sqrt{\sum_j |A_{ij}|^2} \quad . \tag{9}$$

Note that the $l_{12}$-norm promotes sparsity only in the spatial dimension for the time-windowed data. It is important that the analysis window is chosen short enough to ensure sufficient sparsity and that a small time increment, $\tau$, is used in order to avoid abrupt changes in the sound scene.

Finding the plane-wave signals is equivalent to finding a demixing matrix, $\mathbf{D}(n,k)$, such that:

$$\mathbf{S}(n,k) = \mathbf{D}(n,k)\mathbf{B}(n,k)\ . \tag{10}$$

The effect of $\mathbf{D}(n,k)$ is to demix the HOA signals into plane-wave signals. Using Eq. (10), the optimization problem (6) can be reformulated as:

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{D}(n,k)\mathbf{B}(n,k)\|_{12} \\ &\text{subject to} \quad \mathbf{Y}\mathbf{D}(n,k) = \mathbf{I} \quad , \end{aligned} \tag{11}$$

where $\mathbf{I}$ is the identity matrix. An Iteratively Reweighted Least Squares (IRLS) algorithm [8] is applied to solve Problem (11) for $\mathbf{D}(n,k)$.

The size of the optimization problem (6) increases with the number of measurement vectors, typically a few hundred to a few thousand in our implementation. However these vectors are not linearly independent, therefore the size of the measurement matrix $\mathbf{B}(n,k)$ can be reduced by expressing it in the subspace defined by its first $(L+1)^2$ singular vectors, as proposed in [9] and detailed in [5]. This data reduction lowers the computational cost of solving the optimization problem. For example, in the case of order-1 HOA signals with 1024 measurement vectors, $\mathbf{B}(n,k)$ is a $4 \times 1024$ matrix. Once expressed in the proper subspace, the size of the measurement matrix is reduced to $4 \times 4$. Note that this process does not affect the result of the optimization problem.

### 2.3. HOA Upscaling

Once the demixing matrix is obtained, an upscaling matrix is calculated, whose action is to re-encode the order-$L$ HOA signals to a higher order, $L'$. The upscaling matrix for frequency bin $k$ and time analysis window $n$ is given by:

$$\mathbf{U}(n,k) = (1-\alpha)\mathbf{U}(n-1,k) + \alpha\mathbf{Y}'\mathbf{D}(n,k) \qquad (12)$$
$$0 \leq \alpha \leq 1 \quad ,$$

where $\alpha$ is the forgetting factor, $\mathbf{Y}'$ is defined similarly to (4) but now includes spherical harmonic components up to order $L'$ and $\mathbf{U}(n,k)$ is the upscaling matrix at frequency bin $k$, such that:

$$\mathbf{B}'(n,k) = \mathbf{U}(n,k)\mathbf{B}(n,k)\,, \qquad (13)$$

where $\mathbf{B}'(n,k)$ now includes spherical harmonic components up to order $L'$. The recursive nature of the upscaling matrix given in Eq. (12) acts to smooth any sharp changes that may occur between successive time windows.

To work in the time domain, the $\mathbf{U}(n,k)$ matrices are combined across frequency and the inverse Fourier transform is applied to obtain a matrix of finite impulse response filters, $\mathbf{U}^{(n)}(t)$. The order-$L$ time-domain HOA signals are then convolved with the upscaling filters to obtain the order-$L'$ time-domain HOA signals for the $n$th time frame:

$$\mathbf{b}'^{\,(n)}(t) = \mathbf{U}^{(n)}(t) \circledast \mathbf{b}^{(n)}(t) \quad , \qquad (14)$$

where $\mathbf{b}^{(n)}(t)$ and $\mathbf{b}'^{\,(n)}(t)$ are the time-domain vectors of order-$L$ and $L'$ HOA signals for the $n$th time frame and $\circledast$ denotes the convolution of a vector of signals by a matrix of filters. The upscaled HOA signals are combined across the time frames using an overlap-and-add technique.

## 3. LISTENING TEST

### 3.1. Method

A formal listening test was performed to perceptually evaluate the quality of the sound field reproduced using the frequency-domain HOA upscaling algorithm described here with the time-domain HOA upscaling algorithm described in [5]. The HOA sound scenes were upscaled from order-1 to order-4. The sound fields reproduced using these algorithms were compared to sound fields reproduced using the classical HOA approach with HOA orders 1, 2, 3 and 4. In the classical HOA approach, the loudspeaker signals are obtained using fixed decoding filters which are calculated by minimising an $l_2$-norm, as described in [10]. The listening test was performed in an anechoic room containing a spherical array of 32 loudspeakers [10]. The MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) [11] paradigm was used to conduct the test. The reference stimulus was HOA order-4, which was decoded to all 32 loudspeakers. The anchor signals were HOA order-1 signals, band-pass filtered with low and high cutoff frequencies of 500 and 4000 Hz, respectively, and decoded to all 32 loudspeakers. The upscaled HOA sound scenes were decoded to all 32 loudspeakers, while HOA orders 1, 2 and 3 were decoded to a subset consisting of 8, 12 and 24 loudspeakers, respectively. Note that these numbers of loudspeakers were chosen to optimize the quality of the playback for the different HOA orders. In the MUSHRA test, the subjects were asked to grade the quality of the rendered scene of each stimulus with how closely it matched the reference. Subjects were encouraged to move their head, while remaining seated, in order to gauge the size of the sweet spot. Each stimulus was rated between 0 and 100.

Three virtual sound scenes were used for the listening test: (1) a single talker; (2) 3 talkers consisting of two males and one female; (3) a rock-music band consisting of a singer, a piano, a bass guitar, drums and backing vocals. The sound scenes were simulated in the same reverberant room using MCROOMSIM [12], a multi-channel room acoustics simulator. This simulator has the ability to directly provide the HOA signals for the simulated reverberant scene. The average reverberation time (T30) and speech clarity index (C50) of the room were 0.32 seconds and 23 dB, respectively.

For both the frequency and time-domain HOA upscaling algorithms, a plane-wave dictionary size of 92 components was used. The length of the time analysis window was 1024 samples long and the sample frequency was 48 kHz. The filter bank used in the time-domain HOA upscaling algorithm consisted of 125 sub-bands. The length of the STFT analysis window in the frequency-domain HOA upscaling algorithm was 256 samples long.

### 3.2. Results

A total of twelve subjects participated in the listening test. Statistical analysis was performed on the results, with Figure 2 showing the box plots for the three sound scenes. The red plus symbols represent outlying scores. The upper and lower horizontal bars on the vertical black dashed line represent the upper and lower bounds of the data (excluding the outlying scores). The upper and lower horizontal bars of the blue box represent the upper and lower quartiles of the data and the red horizontal bar inside the blue box represents the median. The 95% confidence interval around the median line is indicated by the extent of the slanted lines extending above and below the median line, the "notch".

In all cases both the frequency and time-domain HOA upscaled sound fields were perceived much closer to the reference than the HOA order-1 reproduced sound field. In other words, the two upscaling methods significantly improved the quality of the playback. This improvement is observed in all scenes, which proves that the CS-based methods are relatively robust to the complexity of the sound scene. Furthermore, in all scenes, the performance of the upscaling algorithms is on par or better than HOA order-2 and HOA order-3. The frequency-domain algorithm appears to perform better than the time-domain algorithm in every case, although this result is not statistically significant in scene 3 given the variability in the ratings.

Finally, note that the HOA order-2 reproduced sound field was rated slightly better than the HOA order-3 sound field in scene 1, which was unexpected. This may be attributed to the fact that, in the order-3 case, the number of loudspeakers (24) used was much greater than the number of HOA signals (16).

## 4. CONCLUSION

In this paper, we have presented a novel frequency-domain algorithm for upscaling HOA sound scenes to higher orders. Upscaled HOA sound scenes have a greater spatial resolution, which allows more loudspeakers to be used during the playback, resulting in a larger sweet spot and improved sound quality. Results of a formal listening test were presented, showing that the proposed algorithm significantly improves the fidelity of the sound field reproduction when compared to classical HOA playback.

The algorithm is based on a sparse plane-wave decomposition of the sound field. The results of the listening test indicate that the proposed upscaling method is robust against the presence of multiple sources in a reverberant environment.

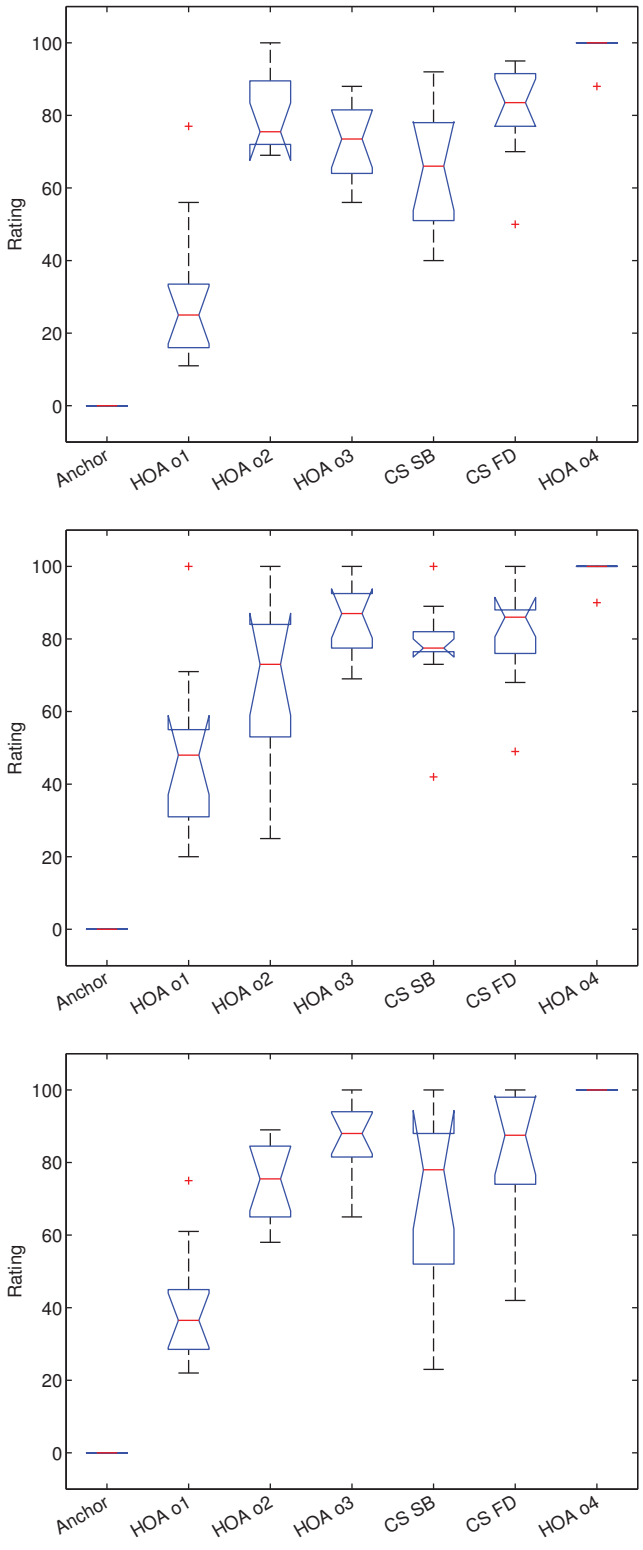


**Fig. 2**. Box plots showing the perceived quality for each spatial sound field reproduction method. The top, middle and bottom plots are the results for scenes 1, 2 and 3, respectively. Refer to the text for an explanation of the various elements in the box plot.

Performing the sparse plane-wave decomposition in the frequency domain opens new prospects for the application of CS techniques to spatial audio analysis and synthesis using microphone arrays. That is to say, having a frequency-dependent plane-wave dictionary allows one to work directly in the microphone domain, as opposed to the spherical harmonic domain. This has the potential to reduce the effect of spatial aliasing when using microphone arrays and the authors intend to explore this idea in future work.

## 5. REFERENCES


[1] J. Daniel, *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*, Ph.D. thesis, Université Paris 6, Paris, France, 2000.

[2] A. Parthy, C. Jin, and A. van Schaik, "Evaluation of a concentric rigid and open spherical microphone array for sound reproduction," in *Proceedings of Ambisonics Symposium 2009*, Graz, Austria, June 2009.

[3] N. Epain, C. Jin, and A. van Schaik, "The application of compressive sampling to the analysis and synthesis of spatial sound fields," in *Proceedings of the AES 127th Convention*, New York, NY, U.S.A., October 2009.

[4] A. Wabnitz, N. Epain, A. van Schaik, and C. Jin, "Time domain reconstruction of spatial sound fields using compressed sensing," in *Proceedings of ICASSP*, Prague, Czech Republic, May 2011.

[5] A. Wabnitz, N. Epain, A. McEwan, and C. Jin, "Upscaling ambisonic sound scenes using compressed sensing techniques," in *Proceedings of WASPAA*, New Paltz, NY, USA, October 2011.

[6] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.

[7] A. Solvang, "Spectral impairment of two-dimensional higher order ambisonics," *J. Audio Eng. Soc.*, vol. 56, no. 4, pp. 267–279, 2008.

[8] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Comm. Pure Appl. Math.*, vol. 63, no. 1, pp. 1–38, 2010.

[9] D. Malioutov, M. Çetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, 2005.

[10] N. Epain, P. Guillon, A. Kan, R. Kosobrodov, D. Sun, C. Jin, and A. van Schaik, "Objective evaluation of a three-dimensional sound field reproduction system," in *International Congress on Acoustics*, Sydney, Australia, August 2010.

[11] ITU-R BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," 2003.

[12] A. Wabnitz, N. Epain, C. Jin, and A. van Schaik, "Room acoustics simulation for multichannel microphone arrays," in *Proceedings of the International Symposium on Room Acoustics*, Melbourne, Australia, August 2010.