A SUPER-RESOLUTION BEAMFORMING ALGORITHM FOR SPHERICAL MICROPHONE ARRAYS USING A COMPRESSED SENSING APPROACH

Ping Kun Tony Wu, Nicolas Epain, Craig Jin

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

ABSTRACT

In this paper, we present a novel beamforming algorithm that is designed for spherical microphone arrays and formulated in the spherical harmonic domain. The proposed algorithm employs sparse recovery, a compressed sensing technique, and assumes the position of the source signals are unknown. A formal listening test was conducted to evaluate the performance of the proposed algorithm and the results indicate the effectiveness of the proposed algorithm.

Index Terms— Spherical Microphone Arrays, Source Localization, Beamforming, Compressed Sensing

1. INTRODUCTION

Spherical microphone arrays provide a promising tool for the spatial analysis of complex sound fields that facilitate the transformation of microphone domain signals into the spherical harmonic domain. Working in the spherical harmonic domain has several advantages including scalability and the ability to rotate the sound scene by a simple matrix operation. Thus, spherical microphone array signal processing has become an important technique in various applications such as sound field recording [1, 2, 3], sound field analysis [4, 5, 6, 7], source localization [8, 9], speech enhancement [10], etc.

In this paper, we employ compressed sensing (CS) techniques to create a new beamforming algorithm for spherical microphone arrays. Compressed sensing is a sensing paradigm that defines the sparse inverse solutions for underdetermined systems. More information about the CS technique can be found in [11]. This work is based on our previous work [12, 13]. In [12], Wabnitz *el al.* proposed an algorithm for upscaling ambisonic sound scenes using the CS technique. It allows more loudspeakers to be used during the playback, resulting in a larger sweet spot and improves sound quality. In this work, we employ the CS technique to develop a super-resolution beamforming algorithm. The proposed beamforming algorithm is developed in the "upscaled" spherical harmonic domain. We present the results of a psychoacoustic listening test that evaluates the performance of the proposed algorithm compared with other algorithms. Section 2 describes the methods behind the proposed super-resolution beamforming technique. Section 3 describes the psychoacoustic listening test and Section 4 presents the results. In Section 5, we conclude the paper.

2. METHOD

2.1. Sparse Plane-Wave Decomposition

Consider a general spherical microphone system with L microphone signals which is modeled as:

$$x_l(t) = \sum_{n=1}^{N} g_{l,n}(t) \otimes s_n(t),\ l = 1, 2, \ldots, L\,, \qquad (1)$$

where $x_l(t)$ is the signal at the l-th microphone, $s_n(t)$ is the n-th source signal, N is the total number of the source signals, $g_{l,n}(t)$ is the impulse response describing the room transfer function from the n-th source signal location to the l-th microphone and $\otimes$ represents the convolution operation. The spherical harmonic expansion, also referred to as the higher order ambisonic (HOA) signals, of a sound field corresponding to a set of plane waves in the time-frequency domain can be expressed as a simple matrix product:

$$\mathbf{b}(m, f) = \mathbf{Y}_{\text{plw}}\mathbf{s}(m, f)\,, \qquad (2)$$

where

$$\begin{aligned}
\mathbf{b}(m, f) &= \left[b_0^0(m, f), b_1^{-1}(m, f), \ldots, b_\lambda^\lambda(m, f)\right]^\intercal\,, \\
\mathbf{Y}_{\text{plw}} &= \left[\mathbf{y}(\theta_1, \phi_1), \mathbf{y}(\theta_2, \phi_2), \ldots, \mathbf{y}(\theta_P, \phi_P)\right]\,, \\
\mathbf{y}(\theta_p, \phi_p) &= \left[Y_0^0(\theta_p, \phi_p), Y_1^{-1}(\theta_p, \phi_p), \ldots, Y_\lambda^\lambda(\theta_p, \phi_p)\right]^\intercal\,, \\
\mathbf{s}(m, f) &= \left[s_1(m, f), s_2(m, f), \ldots, s_P(m, f)\right]^\intercal\,,
\end{aligned}$$

$(.)^\intercal$ denotes the transpose, $\mathbf{b}(m, f)$ is a $(\lambda + 1)^2 \times 1$ vector containing the STFT samples of the order-λ HOA signals for time window m and frequency bin f, $\mathbf{Y}_{\text{plw}}$ is a $(\lambda + 1)^2 \times P$ spherical harmonic matrix, truncated to order λ, with column p providing the spherical harmonic expansion for a plane-wave source located in the direction (θ_p, ϕ_p), P is the total number of columns (entries) in the matrix $\mathbf{Y}_{\text{plw}}$ (dictionary)

of possible plane-wave source directions and is typically chosen much larger than $(\lambda+1)^2$, $\mathbf{s}(m,f)$ is a $P \times 1$ vector of the plane-wave source signals. The p-th row of $\mathbf{s}(m,f)$ is non-zero if there is a signal in the direction (θ_p, ϕ_p).

Equation (2) is an under-determined system of equations. In general, there are an infinite number of solutions and the inverse problem is ill-posed. Our approach for solving this ill-posed problem is to impose sparsity on $\mathbf{s}(m,f)$, so that the resulting sound field is explained by a small number of plane-wave sources. Note that it has been found that the spatial sound field is more likely to be sparsely represented within a sub-band of frequencies than over the full bandwidth [14] and thus, the proposed algorithm operates in the time-frequency domain. Moreover, we assume that the source signals are non-moving sources so that the solutions will have a common sparsity pattern over a short time interval. We solve the following sparse recovery problem:

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{S}(m,f)\|_{1,2} \\ &\text{subject to} \quad \mathbf{B}(m,f) = \mathbf{Y}_{\text{plw}}\mathbf{S}(m,f)\,, \end{aligned} \tag{3}$$

where

$$\begin{aligned} \mathbf{B}(m,f) &= [\mathbf{b}(m\tau,f), \mathbf{b}(m\tau+1,f), \dots \\ &\quad , \mathbf{b}(m\tau+T-1,f)]\,, \\ \mathbf{S}(m,f) &= [\mathbf{s}(m\tau,f), \mathbf{s}(m\tau+1,f), \dots \\ &\quad , \mathbf{s}(m\tau+T-1,f)]\,, \end{aligned}$$

$\mathbf{B}(m,f)$ is a $(\lambda+1)^2 \times T$ matrix containing the T consecutive STFT samples of the order-λ HOA signals with τ being the increment between analysis windows, $\mathbf{S}(m,f)$ is a $P \times T$ matrix of the plane-wave source signals and $\|.\|_{1,2}$ denotes the $l_{1,2}$-norm and is defined as:

$$\|\mathbf{A}\|_{1,2} = \sum_i \sqrt{\sum_j A_{i,j}^2}\,.$$

The computational cost of the above optimization problem (3) can be high depending on the size of the matrix $\mathbf{B}(m,f)$. In order to reduce the computational complexity and the sensitivity to noise, a data dimension reduction method proposed in [15] is employed. It can been seen that recovering $\mathbf{S}(m,f)$ from $\mathbf{B}(m,f)$ is equivalent to applying a demixing matrix to the HOA signals. Thus, the optimization problem in (3) can be reformulated as:

$$\begin{aligned} &\text{minimize} \quad \|\mathbf{D}(m,f)\mathbf{B}(m,f)\|_{1,2} \\ &\text{subject to} \quad \mathbf{D}(m,f)\mathbf{Y}_{\text{plw}} = \mathbf{I}\,, \end{aligned} \tag{4}$$

where $\mathbf{D}(m,f)$ is the demixing matrix and $\mathbf{I}$ is the identity matrix. Significantly, we apply an overlap-add method which estimates the demixing matrix instead of the plane-wave signals. By applying a smoothing operation to the demixing matrix instead of the plane-wave signals, we avoid smearing the spectral characteristics of output signals which can cause speech distortion [12]. Once the demixing matrix is obtained, the plane-wave signals can be calculated by applying $\mathbf{D}(m,f)$ to the order-λ HOA signals:

$$\hat{\mathbf{s}}(m,f) = \mathbf{D}(m,f)\mathbf{b}(m,f)\,. \tag{5}$$

The estimated plane-wave signals can then be used to identify the direction of the source signals and "upscale" the order-λ HOA signal to a higher order, λ'.

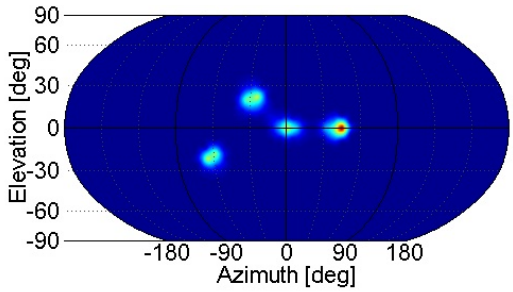

Fig. 1. A spatial spectrum for four-sources scenario is shown. The peaks correspond to the estimated position of source signals. The actual position of the source signals are $(0°, 0°)$, $(40°, 0°)$, $(-60°, -20°)$ and $(-30°, 20°)$.

2.2. Source Localization and HOA upscaling

In this section, we describe the method for localizing the sources and upscaling the HOA signals. In order to localize the sources, we calculate the power spectrum of the estimated plane-wave signals. Because the frequency band for accurate HOA encoding is limited by measurement noise at low frequencies and spatial aliasing at high frequencies, we only calculate the power spectrum of the estimated plane-wave signals within the frequency range for which there is an accurate HOA encoding. The power spectrum of the estimated plane-wave signals as a function of direction in space is referred to as the spatial spectrum and is calculated as:

$$\rho(p) = \sum_{m=1}^{M} \sum_{f=f_{\text{low}}}^{f_{\text{high}}} \|\hat{s}_p(m,f)\|^2\,,\ p = 1, 2, \dots, P\,, \tag{6}$$

where $\rho(p)$ is defined as the spatial spectrum, $\hat{s}_p(m,f)$ is the p-th column vector in $\hat{\mathbf{s}}(m,f)$ and f_{low} and f_{high} are the lower and upper cutoff frequencies index, respectively. The peaks in the spatial spectrum correspond to the location of the source signals. An example of a spatial spectrum for four sources in an anechoic scenario is shown in Figure 1. It can been seen that the peaks are easily identified and the sidelobes are suppressed.

The estimated plane-wave signals can then be used to "upscale" the order-λ HOA signals to a higher order λ', where the number of components in the higher order satisfies

$(\lambda' + 1) < P$. This can be mathematically expressed as:

$$\begin{aligned}\mathbf{b}'(m,f) &= \mathbf{Y}'_{\text{plw}}\hat{\mathbf{s}}(m,f) \\ &= \mathbf{Y}'_{\text{plw}}\mathbf{D}(m,f)\mathbf{b}(m,f) \\ &= \mathbf{P}(m,f)\mathbf{b}(m,f)\,, \end{aligned} \tag{7}$$

where

$$\begin{aligned}\mathbf{b}'(m,f) &= \left[b_0^0(m,f), b_1^{-1}(m,f), \ldots, b_{\lambda'}^{\lambda'}(m,f)\right]^{\mathsf{T}}, \\ \mathbf{Y}'_{\text{plw}} &= \left[\mathbf{y}'(\theta_1,\phi_1), \mathbf{y}'(\theta_2,\phi_2), \ldots, \mathbf{y}'(\theta_P,\phi_P)\right] \ , \\ \mathbf{y}'(\theta_p,\phi_p) &= \left[Y_0^0(\theta_p,\phi_p), Y_1^{-1}(\theta_p,\phi_p), \ldots, Y_{\lambda'}^{\lambda'}(\theta_p,\phi_p)\right]^{\mathsf{T}},\end{aligned}$$

$\mathbf{b}'(m,f)$ is the order-λ' HOA signal, $\mathbf{Y}'_{\text{plw}}$ is a $(\lambda'+1)^2 \times P$ spherical harmonic matrix, truncated to order λ' and $\mathbf{P}(m,f) = \mathbf{Y}'_{\text{plw}}\mathbf{D}(m,f)$ is defined as the upscaling HOA matrix. It should be noted that the upscaled HOA signals provide higher spatial resolution than the original HOA signals, which are order limited by the number of microphones in the array.

2.3. Super-Resolution Beamforming

Once the upscaled HOA signals and the estimated position of source signals are obtained, it is clear that we can beamform with the upscaled HOA signals. We refer to this as super-resolution beamforming. In this section, we describe two super-resolution beamformers: the super-resolution spherical beamformer (SR-SB) and the super-resolution minimum variance distortionless response (SR-MVDR) beamformer. For the SR-SB, the n-th source signal is estimated as:

$$\hat{s}_n(m,f) = \frac{\mathbf{y}'(\hat{\theta_n},\hat{\phi_n})^{\mathsf{T}}}{(\lambda'+1)^2}\mathbf{b}'(m,f), \tag{8}$$

where $(\hat{\theta_n},\hat{\phi_n})$ is the estimated direction of the n-th source signal.

For the SR-MVDR beamformer, the n-th source signal is estimated by applying a weight vector to the upscaled HOA signals:

$$\hat{s}_n(m,f) = \mathbf{w}_n(m,f)^{\mathsf{H}}\mathbf{b}'(m,f)\,, \tag{9}$$

where $(.)^{\mathsf{H}}$ denotes the Hermitian transpose. Similar to the standard MVDR beamformer, The SR-MVDR beamformer aims to minimize the output signal power subject to a distortionless constraint on the response of the beamformer in the look direction:

$$\begin{aligned}&\text{minimize} \quad \mathbf{w}_n^{\mathsf{H}}(m,f)\mathbf{R}_{\mathbf{b}'}(m,f)\mathbf{w}_n(m,f) \\ &\text{subject to} \quad \mathbf{w}_n(m,f)^{\mathsf{H}}(\mathbf{P}(m,f)\mathbf{y}(\hat{\theta_n},\hat{\phi_n})) = 1\,. \end{aligned} \tag{10}$$

where $\mathbf{R}_{\mathbf{b}'}(m,f)$ is a $\lambda' \times \lambda'$ correlation matrix for the upscaled HOA signals. The weighting vector $\mathbf{w}_n(m,f)$ can be calculated as:

$$\mathbf{w}_n(m,f) = \frac{\mathbf{R}_{b'}^{-1}(m,f)\mathbf{v}_n(m,f)}{\mathbf{v}_n(m,f)^{\mathsf{H}}\mathbf{R}_{b'}^{-1}(m,f)\mathbf{v}_n(m,f)}\,. \tag{11}$$

Table 1. The actual direction for each source signal

Source index	A	B	C	D	E	F
Azimuth (°)	0	40	-60	-30	-160	100
Elevation (°)	0	0	-20	20	0	0

where $\mathbf{v}_n(m,f) = \mathbf{P}(m,f)\mathbf{y}(\hat{\theta_n},\hat{\phi_n})$ is defined as the manifold vector for the SR-MVDR beamformer.

3. EXPERIMENT

Computer simulations were used to evaluate the performance of the super-resolution beamforming algorithms. Three different multi-sources scenarios were simulated for both anechoic and reverberant sound conditions. The multi-sources scenarios are comprised of a two-source scenario, a four-source scenario and a six-source scenario. A multichannel room acoustics simulator, MCROOMSIM, that is suitable for a spherical microphone array simulation [16], was used to simulate the anechoic and reverberant sound conditions. Both the anechoic and reverberant room had the same size of $14\text{ m} \times 10\text{ m} \times 3\text{ m}$. The average reverberation time (RT60) for the reverberation room is approximately 0.35 seconds. The spherical microphone array is located (7 m, 4 m, 1.3 m) relative to the corner of the room. The sources are positioned two metres away from the microphone array and the direction for each source relative to the microphone array is shown in Table 1. For the multi-source scenarios, the sources were located as follows: the two-source scenario uses position A and B in Table 1; the four-source scenario uses position A to D and the six-source scenario used position A to F. Source A was always set as the target signal for all scenarios.

The spherical microphone array consists of two concentric arrays of 16 omnidirectional microphones. There are 16 microphones located on the surface of a rigid sphere with a radius of 3.5 cm; the other 16 microphones are located on the surface of a open sphere with a radius of 15 cm. Room impulses responses (RIRs) were obtained for different source locations around the spherical microphone array using MCROOMSIM. These RIRs were then used to filter different voice recordings and combined together to create test mixtures that simulate spherical microphone array recordings in an anechoic room and a reverberant room. In addition, a -40 dB RMS uncorrelated Gaussian white noise is added to each microphone signal in order to model the effect of measurement noise. The mixture signals were approximately 4 seconds in duration with all of the audio processed at a sampling rate of 16 kHz.

The details for the super-resolution beamforming computation are as follows. The upscaling parameters were chosen as $P = 642$, $T = 129$, $\lambda = 2$ and $\lambda' = 4$. The length of the analysis window used for estimating the demixing matrix for

the sparse plane-wave decomposition was 512 samples with a 50 percent overlap between adjacent windows. The frequency range used to compute the spatial spectrum is 350 to 3500 Hz. An iteratively reweighted least squares (IRLS) algorithm [14] was applied to solve the optimization problem (4). The forgetting factor used for smoothing between neighbouring time analysis windows was set to 0.3.

A psychoacoustic listening test was conducted to compare the super-resolution beamforming with other techniques. A MUSHRA-like [17] test paradigm was employed. The low anchor was the raw microphone signal and the reference signal was the clean source signal located at position A. The reference signal was not hidden so that listeners could clearly identify the target signal. The listening test was conducted in a sound-attenuating booth to reduce external sound interference. There were several beamforming test conditions: (1) the unprocessed raw microphone signal; (2) the SR-MVDR beamformer; (3) the SR-SB; (4) the spherical beamformer using the order-2 HOA signals (o2-SB); (5) the MVDR beamformer using the order-2 HOA signals (o2-MVDR); (6) the MVDR beamformer using the raw microphone signals (mic-MVDR) and (7) the MVDR beamformer using theoretical order-4 HOA signals (o4-MVDR). The listener's task was to rate the relative quality of the test stimuli on a scale from 0 (poor) to 100 (excellent).

4. RESULTS AND DISCUSSION

Six listeners participated in the listening test. The results of the listening test are shown in Figure 2. The average ratings for the various beamforming methods are shown for the three multi-source scenarios for both the anechoic and reverberant conditions. The average ratings were obtained by first applying a z-score to the data for each listener and then computing the average. The errorbars indicate the 95% confidence interval. For the anechoic listening condition (Figure 2a),we have the following results. The true order-4 MVDR signals scored the highest. The SR-MVDR beamformer is rated highest compared to the other beamforming methods for the two-source scenario, while the SR-SB beamformer is rated highest compared to the other beamforming methods for the six-source scenario. When there are four sources, the SR-SB and SR-MVDR beamformers perform equally well. The advantage of the super-resolution beamforming seems to come more into play under reverberant listening conditions. For the reverberant listening condition (Figure 2b), the SR-SB beamformer is rated as high as the true order-4 beamformer. Surprisingly, the SR-MVDR beamformer performs poor in reverberant sound conditions. Perhaps one explanation for this is that the manifold vector for the SR-MVDR method, $\mathbf{v}(m, f)$, varies over time and frequency and sometimes does not steer in the right direction when the target signal is silent.

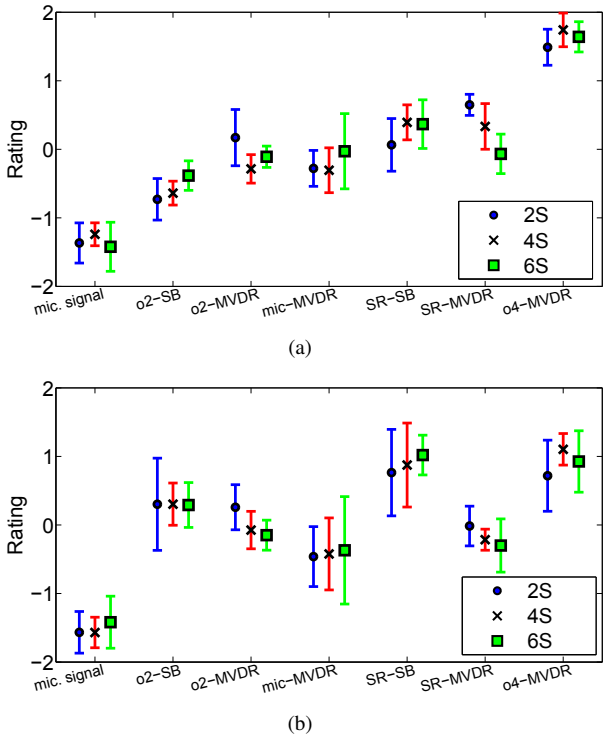

Fig. 2. The results of the listening test are shown for the two, four, and six source scenarios for the (a) anechoic and (b) reverberant sound conditions.

5. CONCLUSION

In this paper, we propose a super resolution beamforming algorithm for spherical microphone arrays. The proposed algorithm employs sparse recovery to localize and separate the source signals. Results of a formal listening indicate that the performance of the order-2 upscaled to order-4 SRSB beamformer is as good as the true order-4 MVDR beamformer. In future work, we will investigate methods to improve the upscaled MVDR beamformer.

6. REFERENCES

[1] J. Meyer and G.W. Elko, "A spherical microphone array for spatial sound recording," *J. Acoust. Soc. Am.*, vol. 111, pp. 2346 – 2346, 2002.

[2] T.D. Abhayapala and D.B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2002, vol. 2, pp. 1949–1952.

[3] I. Balmages and B. Rafaely, "Open-sphere designs for spherical microphone arrays," *IEEE Trans. on Audio,*

Speech and Language Processing, vol. 15, no. 2, pp. 727–732, 2007.

[4] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 135 – 143, 2005.

[5] M. Park and B. Rafaely, "Sound-field analysis by plane-wave decomposition using spherical microphone array," *J. Acoust. Soc. Am.*, vol. 118, no. 5, pp. 3094–3103, 2005.

[6] E.G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography*, p. 218, Academic Press, London, UK, 1999.

[7] B.N. Gover, J.G Ryan, and M.R. Stinson, "Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array," *J. Acoust. Soc. Am*, vol. 4, no. 116, October 2004.

[8] H. Sun, H. Teutsch, E. Mabande, and W. Kellermann, "Robust localization of multiple sources in reverberant environments using eb-esprit with spherical microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011.

[9] N. Epain and C. Jin, "Independent component analysis using spherical microphone arrays," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 91 – 102, 2012.

[10] Y. Peled and B. Rafaely, "Method for dereverberation and noise reduction using spherical microphone arrays," in *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010.

[11] E. Candes and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21 – 30, 2008.

[12] A. Wabnitz, N. Epain, and C. Jin, "A frequency-domain algorithm to upscale ambisonic sound scenes," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[13] P. K. T. Wu, N. Epain, and C. Jin, "A dereverberation algorithm for spherical microphone arrays using compressed sensing techniques," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2012.

[14] A. Wabnitz, *The application of compressed sensing techniques to spatial sound field reproduction*, Ph.D. thesis, University of Sydney, 2012.

[15] A. Wabnitz, N. Epain, McEwan A., and C. Jin, "Upscaling ambisonic sound scenes using compressed sensing techniques," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011.

[16] 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*, 2010.

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