Compressive Beamforming Accelerated with the Kronecker Array Transform

Bruno Masiero[1], Vitor Nascimento[2]

[1] *University of Campinas, Brazil, Email: masiero@unicamp.br*

[2] *University of São Paulo, Brazil, Email: vitor@lps.usp.br*

Introduction

The problem of acoustic scene description with sensor arrays is to determine the number and location of (usually few) sound sources present in a (possibly noisy) sound scene from measurements of the wave field with a microphone array. Conventional beamforming is the most usual method to extract the sources' direction-of-arrival and emitted signal, even though it is characterized by low spatial resolution.

The compressive beamforming (CB) method asserts that spatially sparse signals can be recovered from arrays with reduced number of sensors by solving a convex minimization problem. However, despite the fact that the compressive sensing framework applied in CB offers computational efficiency compared to other sparsity promoting methods, its iterative algorithm is still very time consuming when compared with conventional beamforming. In the quest for a real-time implementation of CB, we present the Kronecker Array Transform (KAT) to speed up the bottleneck of the CB algorithm, namely, the matrix-vector product calculation, which requires as trade-off for considerable calculation speed up the use of a sensor array with separable geometry.

Imaging Algorithms

We consider a sensor array composed of M microphones at Cartesian coordinates $\mathbf{p}_0, \cdots, \mathbf{p}_{M-1} \in \mathbb{R}^3$ being irradiated by an arbitrary sound field which we wish to estimate. We model the sound field as the superposition of the wave fields generated by N acoustic point sources located at coordinates $\mathbf{q}_0, \cdots, \mathbf{q}_{N-1} \in \mathbb{R}^3$, where N is usually a large number in order to obtain an accurate model.

The time-domain samples of each microphone are segmented into frames of K samples, and each frame is converted to the frequency domain using the fast Fourier transform (FFT). In the presence of additive noise, the $M \times 1$ array output vector for a single frequency ω_k ($0 < k < K/2$) on a single frame can be modelled as [3]

$$\boldsymbol{x}(\omega_k) = \boldsymbol{V}(\omega_k)\, \boldsymbol{y}(\omega_k) + \boldsymbol{\eta}(\omega_k), \tag{1}$$

where $\boldsymbol{y}(\omega_k) = [y_0(\omega_k) \;\; y_1(\omega_k) \;\; \cdots \;\; y_{N-1}(\omega_k)]^T$ represents the source signals in the frequency domain, and $\boldsymbol{\eta}(\omega_k)$ represents frequency-domain noise. The array manifold matrix $\boldsymbol{V}(\omega_k) = [\boldsymbol{v}(\mathbf{q}_0, \omega_k) \;\; \boldsymbol{v}(\mathbf{q}_1, \omega_k) \;\; \cdots \;\; \boldsymbol{v}(\mathbf{q}_{N-1}, \omega_k)]$, of size $M \times N$, describes the transfer function between each source n and each sensor m at frequency ω_k.

Assuming that the point sources are in the far field, we define the look direction for source n as $\mathbf{u}_n = -\mathbf{q}_n / \|\mathbf{q}_n\|$. The array manifold vector for source n is then given by $\boldsymbol{v}(\mathbf{q}_n, \omega_k) = \left[e^{j(\omega_k/c)\mathbf{u}_n^T \mathbf{p}_0} \;\; \cdots \;\; e^{j(\omega_k/c)\mathbf{u}_n^T \mathbf{p}_{M-1}}\right]^T$. Please note that in the remaining of this manuscript we omit the dependency on ω_k in $\boldsymbol{V}(\omega_k)$, $\boldsymbol{y}(\omega_k)$ and $\boldsymbol{x}(\omega_k)$ in order to simplify notation.

Spatial filtering

There exist several techniques for estimating $\boldsymbol{y}$ from the array output vector $\boldsymbol{x}$. The authors reviewed some of these techniques in [6], using the *cross spectral matrix* (CSM) for data extraction. We now adapt those techniques for direct estimation with equation (1).

We start with the conventional beamformer (CBF) [3], that is implemented as a weighted sum of the signals captured by the microphones, i.e.,

$$\hat{y} = \boldsymbol{w}^H \boldsymbol{x}, \tag{2}$$

where $\boldsymbol{w} = [w_1 \;\; w_2 \;\; \cdots \;\; w_M]^T$ is a complex weight vector.

Deterministic Beamformers

There are several ways to calculate $\boldsymbol{w}$, the most straightforward manner being CBF, where the spatial filter $\boldsymbol{w}$ is chosen so that the filter output power is maximized when the array is excited by a plane wave arriving from $\boldsymbol{u}$. Thus, we aim to solve

$$\arg\max_{\boldsymbol{w}} \mathrm{E}\left\{ |\hat{y}|^2 \right\}. \tag{3}$$

Substituting (1) and (2) into (3) and assuming that the sound field is composed of a single plane wave propagating in direction $-\boldsymbol{u}$, we obtain the cost function

$$J = \left|\boldsymbol{w}^H \boldsymbol{v}(\boldsymbol{u})\right|^2 |f|^2 + \|\boldsymbol{w}\|^2 \sigma^2. \tag{4}$$

Note that, assuming spatially uncorrelated noise, $\sigma^2 = E\left\{\boldsymbol{\eta}^H \boldsymbol{\eta}\right\}$. To avoid the trivial solution $\|\boldsymbol{w}\| \to \infty$, CBF limits the filter gains by adding the restriction $\|\boldsymbol{w}\| = 1$. In this case, J is maximized when $\left|\boldsymbol{w}^H \boldsymbol{v}(\boldsymbol{u})\right|^2$ is maximum. If we now apply the Cauchy-Schwarz inequality to the this term we verify that

$$\left|\boldsymbol{w}^H \boldsymbol{v}(\boldsymbol{u})\right|^2 \leq \|\boldsymbol{w}\|^2 \|\boldsymbol{v}(\boldsymbol{u})\|^2 = 1 \cdot \left(\boldsymbol{v}(\boldsymbol{u})^H \boldsymbol{v}(\boldsymbol{u})\right), \tag{5}$$

*This work was partly supported by FAPESP (Grants #14/06066-6 and #14/04256-2) and CNPq (Grant #306268/2014-0).

thus the value of $\boldsymbol{w}$ that maximizes the filter output power at frequency ω_k when the array is excited by a plane wave arriving from $\boldsymbol{u}$ is given by

$$\boldsymbol{w}_{\mathrm{CBF}}(\boldsymbol{u}) = \frac{\boldsymbol{v}(\boldsymbol{u})}{\|\boldsymbol{v}(\boldsymbol{u})\|} = \frac{\boldsymbol{v}(\boldsymbol{u})}{\sqrt{M}}. \tag{6}$$

Equation (6) indicates that the CBF acts by applying a delay to the signals captured by each sensor, so that the signals arriving from $\boldsymbol{u}_n$ are aligned in time and thus constructively added. Note that the CBF is a deterministic method since its weights do not depend on the statistics of the incoming signal, but only on the "listening" direction and the geometry of the array.

Another very common deterministic beamformer is the *delay-and-sum* (DAS) beamformer [12]. Similarly to the CBF, the DAS seeks to compensate for the relative delay at each sensor and then averages the resulting signals, thus

$$\boldsymbol{w}_{\mathrm{DAS}}(\boldsymbol{u}) = \frac{1}{M}\boldsymbol{v}(\boldsymbol{u}) = \frac{\boldsymbol{v}(\boldsymbol{u})}{\boldsymbol{v}^H(\boldsymbol{u})\boldsymbol{v}(\boldsymbol{u})}. \tag{7}$$

Thus, both the DAS and the CBF are equivalent, except for a scalar gain.

To obtain the acoustic image, we need to estimate the sound intensity coming from each direction $\boldsymbol{u}$ in a predefined grid, to obtain a vector of estimates $\hat{\boldsymbol{y}}$. Using a fixed beamformer for each direction in the grid and assuming a perfect estimate of the signal, we have

$$\hat{\boldsymbol{y}} = \alpha\, \boldsymbol{V}^H \boldsymbol{x}, \tag{8}$$

where $\alpha = M^{-1/2}$ for CBF and $\alpha = M^{-1}$ for DAS.

Convolutional bluring

Consider a single plane wave travelling along direction $-\boldsymbol{u}_m$. Assuming a discrete scan grid, the estimated pixel corresponding to a generic look direction $\boldsymbol{u}_n$, is given by

$$\hat{y}_n = \boldsymbol{w}^H(\boldsymbol{u}_n)\boldsymbol{v}(\boldsymbol{u}_m)\, y_m \equiv P(\boldsymbol{u}_n, \boldsymbol{u}_m) y_m. \tag{9}$$

For a fixed source direction $\boldsymbol{u}_m$, the term $P(\boldsymbol{u}_n, \boldsymbol{u}_m)$, considered as a function of $\boldsymbol{u}_n$, is the array's *complex point spread function* (CPSF), which describes the array's complex response at directions $\boldsymbol{u}_n$ to an input plane wave arriving from direction $-\boldsymbol{u}_m$. $P(\boldsymbol{u}_n, \boldsymbol{u}_m)$ is defined over the entire space and can be interpreted as a spatial sampling function that should ideally be maximally sharp, that is (for our current discrete model), equal to $P(\boldsymbol{u}_n, \boldsymbol{u}_m) = \delta_{\boldsymbol{u}_n, \boldsymbol{u}_m}$, where $\delta_{\boldsymbol{u}_n, \boldsymbol{u}_m} = 1$ if $\boldsymbol{u}_n = \boldsymbol{u}_m$ and zero otherwise. However, as microphone arrays have a limited number of sensors, their typical CPSF will present a larger beamwidth and consequently a smeared acoustic image.

Now, considering that the sound field is composed of a superposition of N plane waves, we verify that the estimated acoustic image is given by

$$\hat{y}(\boldsymbol{u}_n) = \sum_{m=1}^{N} P(\boldsymbol{u}_n, \boldsymbol{u}_m) y(\boldsymbol{u}_m). \tag{10}$$

Equation (10) can be interpreted as a spatial convolution [6, 13], i.e., when calculating an acoustic image with conventional or optimal beamformers the result is, in fact, the convolution of the actual acoustic image with the array's CPSF. This explains why images produced by standard beamformers are commonly described as smeared or blurred.

Compressive beamforming

The beamforming techniques are robust to noise but suffers from low resolution (as discussed above) and the presence of sidelobes [8, 14, 15]. To counter these effects, a recent work proposes to casts the problem as regularized inverse problem [14].

Regularized signal reconstruction has been a topic of interest for many decades, and gained significant momentum with the popularity of compressive sensing [1, 2]. Indeed, many image reconstruction problems can be recast as convex optimization problems, which can be solved with computationally efficient iterative methods. While many of these techniques were designed for imaging applications, they have remained limited to fields such as medical image reconstruction. Therefore, most of these developments have not yet been applied to acoustic imaging.

ℓ_1-Regularized Least Squares

We assume that the acoustic field arriving at the microphone array is generated by only a few compact sources, i.e., the source distribution is *sparse*. Note that in this case the transfer matrix $\boldsymbol{V}$ will have more columns than rows, so (1) is underdetermined. Prior models of the source distribution can now be incorporated as constraints that allow the underdetermined system of equations to be solved. In this case we apply a sparsity constraint to regularize the inversion problem, as suggested in [14], casting the problem as a basis pursuit with denoising problem (BPDN)—a kind of optimization problem that has been studied in detail in the compressive sensing literature [10]—which has the form

$$\begin{aligned} \underset{\hat{\boldsymbol{y}}}{\text{minimize}} \quad & \|\hat{\boldsymbol{y}}\|_1 \\ \text{subject to} \quad & \|\boldsymbol{x} - \boldsymbol{V}\hat{\boldsymbol{y}}\|_2 \leq \sigma. \end{aligned} \tag{11}$$

The ℓ_1 constraint $\min \|\hat{\boldsymbol{y}}\|_1$ serves to regularize the problem while forcing sparsity, and can be efficiently implemented, e.g., with the SPGL1 algorithm [10].

TV-Regularized Least Squares

To address scenarios where the acoustic images are not sparse in their canonical representations, another possibility is to reconstruct acoustic images with total variation (TV) regularization.

The isotropic total variation norm is defined as

$$\|\boldsymbol{Y}\|_{\mathrm{TV}} = \sum_{i,j} \sqrt{[\nabla_x \boldsymbol{Y}]_{i,j}^2 + [\nabla_y \boldsymbol{Y}]_{i,j}^2} \tag{12}$$

where ∇_x and ∇_y are the first difference operators along the x and y dimensions with periodic boundaries, and i and j are the indices in the x and y dimensions, respectively.

The following optimization problem can then be solved

$$\underset{\hat{\boldsymbol{y}}}{\text{minimize}} \quad \|\widehat{\boldsymbol{Y}}\|_{TV} + \mu/2 \, \|\boldsymbol{x} - \boldsymbol{V}\hat{\boldsymbol{y}}\|_2^2 \,, \tag{13}$$

where $\hat{\boldsymbol{y}} = \text{vec}\{\widehat{\boldsymbol{Y}}\}$ denotes vectorization by stacking the columns of a matrix. The first term measures how much an image oscillates. Therefore, it is smallest for images with plateaus and monotonic transitions, and tends to privilege simple solutions with small amounts of noise. The second term ensures a good fit between the reconstructed image and the measured data. This formulation was first proposed for image denoising [9], and was later generalized and applied successfully to many image reconstruction problems. This method provides accurate and stable image reconstructions with guaranteed convergence, and can be efficiently implemented with, e.g., the TVAL3 algorithm [4].

Compressive beamforming optimization problem are solved in an iterative manner and for large arrays, the calculation bottleneck lies in the matrix-vector products $\boldsymbol{V}\boldsymbol{y}$ and $\boldsymbol{V}^H\boldsymbol{x}$, similarly to the problem discussed in [8]. On the other hand, as thoroughly discussed in [11], the decomposition of a matrix in the Kronecker product of two smaller matrices has the advantage that systems of the form $\boldsymbol{M}\boldsymbol{z} \equiv (\boldsymbol{B} \otimes \boldsymbol{C})\boldsymbol{z} = \boldsymbol{r}$ can be efficiently solved. We now replicate the results presented by the authors in [5] on how to accelerate calculations of (11) and (13).

Kronecker Array Transform

A planar array is *separable* if the microphone positions form a rectangular grid [7] (all positions in the grid must be occupied). The far-field manifold matrix associated to M sensors distributed in a separable geometry and a U-space parametrized rectangular scan grid is equivalent to [7]

$$\boldsymbol{V}[m,n] = e^{j\omega_k \mathbf{u}_x^T(n)\mathbf{p}_x(m)/c} e^{j\omega_k \mathbf{u}_y^T(n)\mathbf{p}_y(m)/c} \tag{14}$$

where for simplicity we assumed the array to be horizontally oriented, p_x and p_y are the sensor coordinates in its x and y coordinates and u_x and u_y are the x and y coordinates of the U-space parametrized look direction, respectively [7].

We now define two new manifold matrices

$$\boldsymbol{V}_x[r,s] = e^{j\omega_k u_x(s) p_x(r)/c}, \tag{15a}$$

$$\boldsymbol{V}_y[g,h] = e^{j\omega_k u_y(h) p_y(g)/c}. \tag{15b}$$

The horizontal array manifold matrix $\boldsymbol{V}_x$ has size $M_x \times N_x$, and the vertical array manifold matrix $\boldsymbol{V}_y$ has size $M_y \times N_y$. M_x and M_y are the number of coordinate points in the x and y directions and N_x and N_y are the number of grid points in the x and y directions, with the restriction that $M = M_x M_y$ and $N = N_x N_y$.

We verify that $m = rM_y + s$ and $n = gN_y + h$. This allows us to rewrite (15) in relation to the indices m and n as $\boldsymbol{V}_x[\lfloor m/M_y \rfloor, \lfloor n/N_y \rfloor]$ and $\boldsymbol{V}_y[\text{mod}(m, M_y), \text{mod}(n, N_y)]$. We further verify that this is equivalent to the Kronecker product

$$\boldsymbol{V} = \boldsymbol{V}_x \otimes \boldsymbol{V}_y. \tag{16}$$

Direct Fast Transform

The bottleneck for calculating (11) and (13) is the direct matrix-vector product $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{y}$. Substituting (16) results into

$$\boldsymbol{x} = (\boldsymbol{V}_x \otimes \boldsymbol{V}_y)\, \boldsymbol{y}. \tag{17}$$

Using the well known Kronecker product identity

$$\text{vec}(\boldsymbol{B}\boldsymbol{Z}\boldsymbol{A}^T) = (\boldsymbol{A} \otimes \boldsymbol{B})\, \text{vec}\,(\boldsymbol{Z})\,, \tag{18}$$

we rewrite (17) as

$$\boldsymbol{X} = \boldsymbol{V}_y \boldsymbol{Y} \boldsymbol{V}_x^T, \tag{19}$$

where $\boldsymbol{x} = \text{vec}(\boldsymbol{X})$ and $\boldsymbol{y} = \text{vec}(\boldsymbol{Y})$. The output matrix $\boldsymbol{X} \in \mathbb{C}^{M_y \times M_x}$ contains the values of $\boldsymbol{x}$ arranged in the same geometrical disposition as the sensors in the array, with the columns of the matrix representing the vertical y-axis and the rows of the matrix representing the horizontal x-axis. The same is valid for the signal matrix $\boldsymbol{Y} \in \mathbb{C}^{N_y \times N_x}$, that contains all values of $\boldsymbol{y}$ arranged in the same geometrical disposition as the scan grid.

Adjoint Fast Transform

To speed up the calculation of (8) we can apply (16) to the adjoint matrix-vector product $\widehat{\boldsymbol{y}} = \boldsymbol{V}^H\boldsymbol{x}$, resulting into

$$\widehat{\boldsymbol{y}} = (\boldsymbol{V}_x \otimes \boldsymbol{V}_y)^H \boldsymbol{x}, \tag{20}$$

which can also be rewritten in a fast transform form using identity (18), such that

$$\widehat{\boldsymbol{Y}} = \boldsymbol{V}_y^H \boldsymbol{X} \boldsymbol{V}_x^*, \tag{21}$$

which is the fast implementation of $\boldsymbol{V}^H$ (note that it has the same computational cost as the direct transform).

Acceleration with the KAT

We now discuss why the forms (19) and (21) are said to be a fast transform of the direct matrix-vector product $\boldsymbol{V}\boldsymbol{y}$ and $\boldsymbol{V}^H\boldsymbol{x}$, respectively. We can readily verify that calculation of the direct product $\boldsymbol{V}\boldsymbol{y}$ requires $M_x M_y N_x N_y$ complex multiply-and-accumulate (MAC) operations. On the other hand, using (19) the required number of operations reduces to $M_x N_x N_y + M_x M_y N_y$ complex MACs when $\boldsymbol{Y}\boldsymbol{V}_x^T$ is first computed, or to $M_y N_x N_y + M_x M_y N_x$ complex MACs when $\boldsymbol{V}_y\boldsymbol{Y}$ is computed first.

If we assume that $N_x = N_y = \sqrt{N}$ and $M_x = M_y = \sqrt{M}$, and additionally, that the number of microphones contained in the array is substantially smaller than the number of scan points, i.e. $M \ll N$, than a rough estimate of the acceleration provided by the KAT lies in the order

of $\sqrt{M}$, which is in agreement with the acceleration estimated in [11]. Further acceleration might be achieved using the NFFT and NNFFT algorithms together with the KAT, as discussed in [6]. However, for the sake of brevity, we will refrain from this discussion here.

Conclusion

Conventional beamforming (CBF) is the most usual method to solve the problem of acoustic scene description, i.e., to extract the sources' direction-of-arrival and emitted signal, even though CBF is characterized by low spatial resolution. New algorithms based on the compressive sensing framework have been proposed to improve microphone array resolution, with the trade-off of increased calculation time. In this manuscript we presented the Kronecker Array Transform (KAT), capable of speeding up calculations with the compressive beamforming framework.

Literatur

[1] E. J. Candès and M. B. Wakin. An Introduction To Compressive Sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[2] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, apr 2006.

[3] D. H. Johnson and D. E. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Prentice Hall, Englewood-Cliffs N.J., 1993.

[4] C. Li. *An Efficient Algorithm For Total Variation Regularization with Applications to the Single Pixel Camera*. Master thesis, Rice University, 2009.

[5] B. S. Masiero and V. H. Nascimento. Kronecker Array Transform. *IEEE Signal Processing Letters*, submitted in 11/02/2016.

[6] V. H. Nascimento, B. S. Masiero, and F. P. Ribeiro. Acoustic Imaging Using the Kronecker Array Transform. In R. F. Coelho, V. H. Nascimento, R. L. de Queiroz, J. M. T. Romano, and C. C. Cavalcante, editors, *Signals and Images: Advances and Results in Speech, Estimation, Compression, Recognition, Filtering, and Processing*, pages 153–178. CRC Press, 2015.

[7] F. P. Ribeiro and V. H. Nascimento. Fast transforms for acoustic imaging—part I: Theory. *IEEE Transactions on Image Processing*, 20(8):2229–2240, 2011.

[8] F. P. Ribeiro and V. H. Nascimento. Fast transforms for acoustic imaging—part II: Applications. *IEEE Transactions on Image Processing*, 20(8):2241–2247, 2011.

[9] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 1992.

[10] E. van den Berg and M. P. Friedlander. Probing the Pareto Frontier for Basis Pursuit Solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008.

[11] C. Van Loan and N. Pitsianis. Approximation with Kronecker Products. In *Linear Algebra for Large Scale and Real Time Applications*, number 1991, pages 293–314. 1993.

[12] H. L. van Trees. *Optimum Array Processing: Part IV of Detection, Estimation and Modulation Theory*. John Wiley & Sons, 2002.

[13] E. G. Williams. *Fourier Acoustics: sound radiation and nearfield acoustical holography*. Academic Press, 1999.

[14] A. Xenaki, P. Gerstoft, and K. Mosegaard. Compressive beamforming. *The Journal of the Acoustical Society of America*, 136(1):260–271, 2014.

[15] T. Yardibi, J. Li, P. Stoica, N. S. Zawodny, and L. N. Cattafesta. A covariance fitting approach for correlated acoustic source mapping. *The Journal of the Acoustical Society of America*, 127(5):2920–2931, 2010.