Principles of Echolocation
in Bats and Machines

Note 5

Beyond Delay-and-Sum

MVDR, MUSIC, spatial smoothing and sparse reconstruction for a single sonar ping

Jan Steckel
Cosys-Lab, University of Antwerp, and Flanders Make, Belgium jan.steckel@uantwerpen.be

Version 2026-11-18 16:41

1 Why Look Beyond Delay-and-Sum?

Delay-and-sum (DAS) beamforming is the workhorse of in-air sonar. It delays every microphone signal so that a wave from the direction of interest lines up, and adds them. It is robust, linear and cheap, but its angular resolution is set by the size of the array. Two reflectors closer than about λ/D radians, with D the aperture, merge into one blob. For the horizontal aperture of an eRTIS, about 66 mm, that is 4.8° at 60 kHz, and worse at the low end of the call. Radar and underwater acoustics have a family of methods that go beyond this limit, most famously the minimum-variance beamformer of Capon (MVDR) and the subspace method MUSIC [15, 3]. A younger family describes the scene as a few reflectors and finds them with a sparsity penalty [14].

In this note, we first set up the signal model and delay-and-sum carefully, because every later method reuses its symbols. We then look at what MVDR and MUSIC need, why a single sonar ping cannot give it to them, how spatial smoothing repairs that, and why the repair asks for a regular array. We end with sparse reconstruction, which needs neither a covariance matrix nor a regular array. Non-linear beamformers such as delay-multiply-and-sum (DMAS) are the subject of Note 6, *Multiplying Instead of Adding*.

Throughout, we use a simulation in the horizontal plane (Fig. 1): a uniform linear array of 32 elements, 2.14 mm apart (half a wavelength at 80 kHz, so the same 66 mm aperture as the eRTIS), an eRTIS-like call (a hyperbolic sweep from 75 to 20 kHz, sampled at 450 kHz), and two equally strong reflectors at 2 m that are only 5° apart, plus a third one elsewhere.

2 Direction Becomes Delay

Sound is a pressure wave that travels at $c \approx 343$ m/s. Far from the source, the wavefronts are planes, and the only thing that distinguishes one source from another is its direction. We describe that direction with the unit vector $\boldsymbol{u}$, which points from the array *towards* the source, and the position of microphone m with the vector $\boldsymbol{r}_m$, measured from a reference point, here the centre of the array (Fig. 2). If the wave carries the waveform $s(t)$, as received at the reference point, then the pressure at position $\boldsymbol{r}$ is

$$p(\boldsymbol{r}, t) = s\big(t + \boldsymbol{u}^T \boldsymbol{r}/c\big). \tag{1}$$

The sign is easy to check. A microphone with $\boldsymbol{u}^T \boldsymbol{r}_m > 0$ lies closer to the source than the reference point, by exactly the distance $\boldsymbol{u}^T \boldsymbol{r}_m$ (magenta in Fig. 2), so it hears the waveform

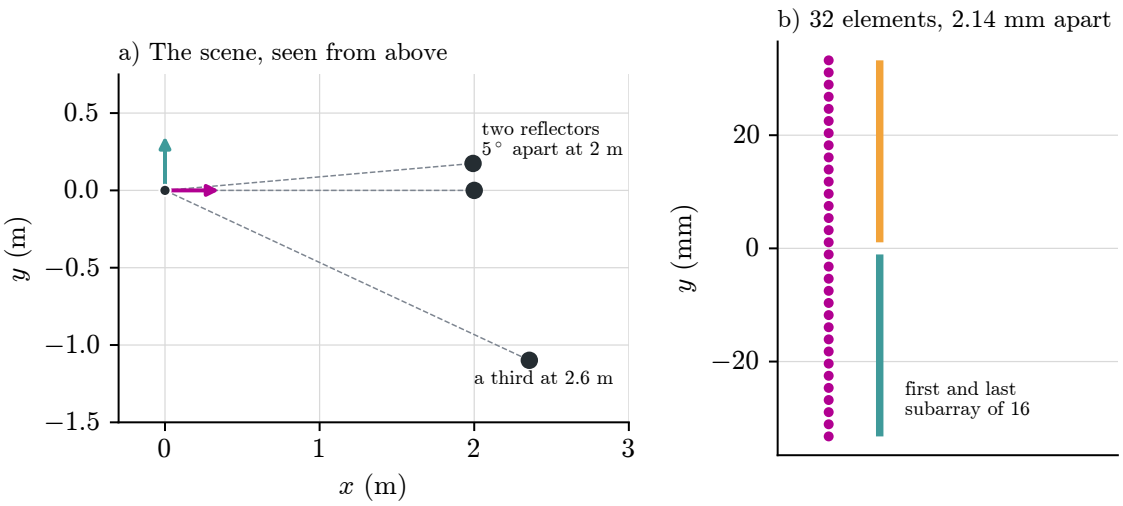


Figure 1: The simulation. a) The sonar (triad: magenta its pointing direction, teal its left) and three point reflectors, two of them 5° apart at 2 m. b) The dense uniform linear array, with the first and the last of the subarrays used for spatial smoothing (Section 7).

earlier. Microphone m therefore receives $s(t - \tau_m)$ with the delay

$$\tau_m = -\frac{\boldsymbol{u}^T \boldsymbol{r}_m}{c}, \tag{2}$$

which is negative for the microphones on the side of the source. For our line array along the y axis and a source at bearing θ (measured from the forward direction x, positive to the left), $\boldsymbol{u} = [\cos\theta, \sin\theta]^T$ and $\boldsymbol{r}_m = [0, y_m]^T$, so $\boldsymbol{u}^T \boldsymbol{r}_m = y_m \sin\theta$.

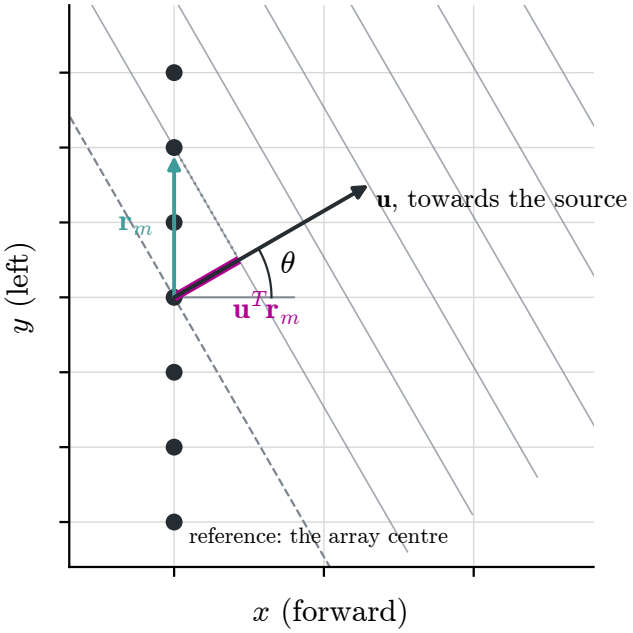


Figure 2: The signal model in the horizontal plane. A plane wave arrives from the direction $\boldsymbol{u}$ (grey lines: wavefronts, dashed: the one through the reference point). Microphone m at position $\boldsymbol{r}_m$ is closer to the source by $\boldsymbol{u}^T \boldsymbol{r}_m$ (magenta), so it hears the wave $\boldsymbol{u}^T \boldsymbol{r}_m / c$ earlier.

The plane-wave picture holds beyond about $2D^2/\lambda$, which for our 66 mm aperture at 75 kHz is 1.9 m. Our reflectors at 2 m sit right at that limit. MVDR and MUSIC below use plane-wave

steering vectors, as is usual; the sparse methods at the end use the exact geometry.

From delay to phase. If we look at one frequency f at a time, a delay is a phase shift: $s(t-\tau) \to S(f)\, e^{-j2\pi f\tau}$. Collecting the M microphone values at that frequency in a vector $\boldsymbol{x}$, a single source contributes $S(f)$ times the steering vector

$$\boldsymbol{a}(\boldsymbol{u}) = \left[e^{\,j2\pi f \boldsymbol{u}^T \boldsymbol{r}_1/c},\ \ldots,\ e^{\,j2\pi f \boldsymbol{u}^T \boldsymbol{r}_M/c}\right]^T, \tag{3}$$

and K sources plus sensor noise give the linear model

$$\boldsymbol{x} = \boldsymbol{A}\,\boldsymbol{s} + \boldsymbol{n}, \qquad \boldsymbol{A} = [\boldsymbol{a}(\boldsymbol{u}_1), \ldots, \boldsymbol{a}(\boldsymbol{u}_K)], \tag{4}$$

with $\boldsymbol{s}$ the K complex source amplitudes and $\boldsymbol{n}$ the noise. The matrix $\boldsymbol{A}$ depends only on the geometry and the frequency, never on the signals. Our calls are broadband, so we cut a short window out of every matched-filtered microphone signal (here 64 samples, about 2 cm of range), take its Fourier transform and solve one such narrowband problem per frequency bin. The bins are combined at the end.

3 Delay-and-Sum and Its Point Spread Function

Delay-and-sum compensates the delays for a look direction $\boldsymbol{u}_0$ and adds the microphones. Per bin, that is the inner product with the steering vector, and the output power is

$$y(\boldsymbol{u}_0) = \frac{1}{M}\boldsymbol{a}^{\mathsf{H}}(\boldsymbol{u}_0)\,\boldsymbol{x}, \qquad P_{\mathrm{DAS}}(\boldsymbol{u}_0) = \frac{1}{M^2}\left|\boldsymbol{a}^{\mathsf{H}}(\boldsymbol{u}_0)\,\boldsymbol{x}\right|^2. \tag{5}$$

A source in the look direction adds up M times, uncorrelated noise only $\sqrt{M}$ times, which gives an array gain of $10\log_{10} M$ dB. Scanning $\boldsymbol{u}_0$ over all directions gives the acoustic image. The image of a single point source, the beampattern $B(\boldsymbol{u}_0) = \boldsymbol{a}^{\mathsf{H}}(\boldsymbol{u}_0)\boldsymbol{a}(\boldsymbol{u})/M$, is the point spread function (PSF) of the array (Fig. 3).

For a line array, B is the discrete Fourier transform of the microphone weights, evaluated at the spatial frequency $y \sin\theta/\lambda$. Everything we know about sampling a signal in time therefore carries over to sampling a wavefield in space (Table 1). The mainlobe is about λ/D radians wide (3.8° at 75 kHz and 7.2° at 40 kHz for our array), because a longer record gives a finer frequency resolution. Uniform weights give a first sidelobe at -13 dB whatever the number of microphones, as a rectangular window does for an FFT, and a taper (shading) lowers the sidelobes at the cost of a wider mainlobe. A spacing above $\lambda/2$ undersamples the wavefield, and the result is spatial aliasing: full-height copies of the mainlobe, the grating lobes. With four times our spacing, they appear at $\pm 32°$ at 75 kHz (Fig. 3b). Over a broad band, the sidelobes at different frequencies fall at different angles and partly average out (Fig. 3a, magenta), which is what sparse random arrays exploit.

Table 1: Sampling in time and sampling in space.

	time (one microphone)	space (the array)
sample spacing	$T_s = 1/f_s$	microphone spacing d
what we resolve	frequency f	direction, through $\sin\theta/\lambda$
Nyquist	$f_s > 2f_{\max}$	$d < \lambda/2$
aliasing	a tone at a wrong frequency	a grating lobe in a wrong direction
resolution	$\Delta f \approx 1/T$ (record length)	$\Delta\theta \approx \lambda/D$ (aperture)
window	Hann on the samples	shading of the microphones

The trade-off between mainlobe width and sidelobe level cannot be beaten with fixed weights. To do better, the weights must depend on the data, and that brings us to the covariance matrix.

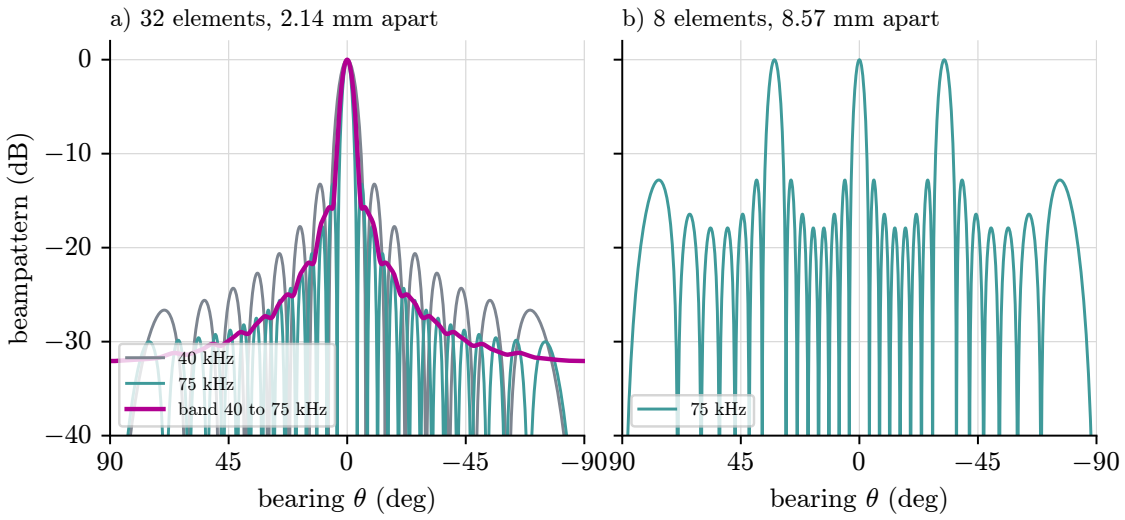


Figure 3: Delay-and-sum beampatterns, steered to broadside. a) Our dense array: the mainlobe narrows with frequency, and the band average (magenta) smooths the sidelobes. b) The same aperture with a quarter of the elements: four times the spacing, and grating lobes as high as the mainlobe.

4 The Covariance Matrix

Everything the next methods know about the scene is summarized in the spatial covariance matrix of the snapshots,

$$\boldsymbol{R} = \mathrm{E}\big[\boldsymbol{x}\,\boldsymbol{x}^{\mathsf{H}}\big] = \boldsymbol{A}\,\boldsymbol{R}_s\,\boldsymbol{A}^{\mathsf{H}} + \sigma^2\boldsymbol{I} \approx \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}_k\boldsymbol{x}_k^{\mathsf{H}}, \tag{6}$$

with $\boldsymbol{R}_s$ the covariance of the source amplitudes and σ^2 the noise power. It is an $M \times M$ Hermitian matrix. Its diagonal holds the power per microphone, and its off-diagonal elements the cross-spectra between microphone pairs. It is estimated from N snapshots, and as a rule of thumb an adaptive beamformer needs $N \geq 2M$ of them to lose less than 3 dB to estimation errors [4], which for our array means 64 snapshots. Delay-and-sum, too, is a function of $\boldsymbol{R}$: its expected output power is $P_{\mathrm{DAS}} = \boldsymbol{a}^{\mathsf{H}}\boldsymbol{R}\,\boldsymbol{a}/M^2$. Every method below is a different quadratic form in $\boldsymbol{R}$, so $\boldsymbol{R}$ is estimated once and the scan over directions is cheap.

If the K sources are uncorrelated, $\boldsymbol{R}_s$ has rank K, and the eigendecomposition of $\boldsymbol{R}$ splits in two. The K largest eigenvalues belong to the signal subspace, spanned by the steering vectors of the sources. The other $M-K$ eigenvalues all equal σ^2, and their eigenvectors span the noise subspace, which is orthogonal to every source steering vector. Counting the eigenvalues that stand out above the noise floor, for example with an information criterion [5], estimates K.

5 Two Adaptive Beamformers: MVDR and MUSIC

MVDR (Capon). Instead of fixed weights, the minimum-variance distortionless-response beamformer chooses, for every look direction, the weights $\boldsymbol{w}$ that let the look direction through unchanged, $\boldsymbol{w}^{\mathsf{H}}\boldsymbol{a} = 1$, and minimize everything else, the output power $\boldsymbol{w}^{\mathsf{H}}\boldsymbol{R}\,\boldsymbol{w}$ [1]. A Lagrange multiplier gives

$$\boldsymbol{w} = \frac{\boldsymbol{R}^{-1}\boldsymbol{a}}{\boldsymbol{a}^{\mathsf{H}}\boldsymbol{R}^{-1}\boldsymbol{a}}, \qquad P_{\mathrm{MVDR}}(\theta) = \frac{1}{\boldsymbol{a}^{\mathsf{H}}\boldsymbol{R}^{-1}\boldsymbol{a}}. \tag{7}$$

The beamformer places nulls on the other sources, which is what sharpens the image. In practice, a small multiple of the identity is added to $\boldsymbol{R}$ before the inversion (diagonal loading, $\boldsymbol{R}+\epsilon\boldsymbol{I}$). It

makes the inversion robust against estimation and calibration errors at a small cost in resolution. For very large ϵ, the weights tend to those of delay-and-sum, so ϵ is a knob between the two.

MUSIC. MUSIC uses the split into subspaces directly [2]. With $\boldsymbol{E}_n$ the $M-K$ eigenvectors of the noise subspace, it looks for the directions in which the steering vector has (almost) no component in that subspace:

$$P_{\mathrm{MUSIC}}(\theta) = \frac{1}{\left\|\boldsymbol{E}_n^{\mathsf{H}}\boldsymbol{a}(\theta)\right\|^2}. \tag{8}$$

This pseudo-spectrum has very sharp peaks, because its resolution comes from the projection and not from the aperture. Its height says nothing about the strength of a reflector, though, and MUSIC needs the number of sources K.

Key idea 1

Delay-and-sum, MVDR and MUSIC are three quadratic forms in the same covariance matrix $\boldsymbol{R}$. MVDR places nulls on everything but the look direction, MUSIC finds the directions orthogonal to the noise subspace. Both can resolve reflectors closer than the λ/D limit of delay-and-sum, but only if $\boldsymbol{R}$ is estimated well.

6 The Sonar Problem: One Ping, Coherent Echoes

Both methods assume that $\boldsymbol{R}$ is averaged over many snapshots in which the sources are uncorrelated. In radar, in passive sonar and in a leak camera, that is easy: many snapshots arrive within a fraction of a second, and independent sources emit independent signals. In active in-air sonar, both assumptions fail.

Few snapshots. Sound is slow. A measurement out to 10 m takes 58 ms, so a sonar collects at most about 17 pings per second. In that time, a robot driving at 1 m/s moves about 6 cm, more than ten wavelengths at 80 kHz. Successive pings are therefore not new snapshots of the same scene, because the geometry, and with it every phase in $\boldsymbol{x}$, has changed.

Coherent echoes. Within one ping, all echoes are delayed copies of the same call. Two reflectors at the same range return the same waveform with a fixed relative amplitude and phase, so they are fully coherent and $\boldsymbol{R}_s$ in Eq. (6) has rank one. Averaging over time samples within the ping does not help, because every snapshot contains the same combination of the two. Figure 4 shows it: without further measures, the second eigenvalue of $\boldsymbol{R}$ lies at -160 dB, which is numerical round-off. To MUSIC, the two reflectors look like a single source in between. MVDR is worse off still: with a coherent second source, the cheapest way to minimize the output power is to cancel the look direction with the other reflector's contribution, and the target disappears (signal cancellation). Figure 5a shows the result. Without smoothing, all three methods show a single peak between the two reflectors.

7 Spatial Smoothing

If we cannot average over time, we can average over space [6, 7]. We divide the array into $M-L+1$ overlapping subarrays of L consecutive elements (Fig. 1b) and average their $L \times L$ covariance matrices:

$$\boldsymbol{R}_f = \frac{1}{M-L+1}\sum_{k=1}^{M-L+1} \boldsymbol{R}_{k:k+L-1,\ k:k+L-1}. \tag{9}$$

Seen from the subarrays, every reflector has the same steering vector up to a phase factor, but that phase factor differs between the reflectors, because their directions differ. From one subarray to the next, the relative phase of two coherent reflectors therefore changes, and averaging over the subarrays decorrelates them. Forward-backward smoothing adds the same average for the

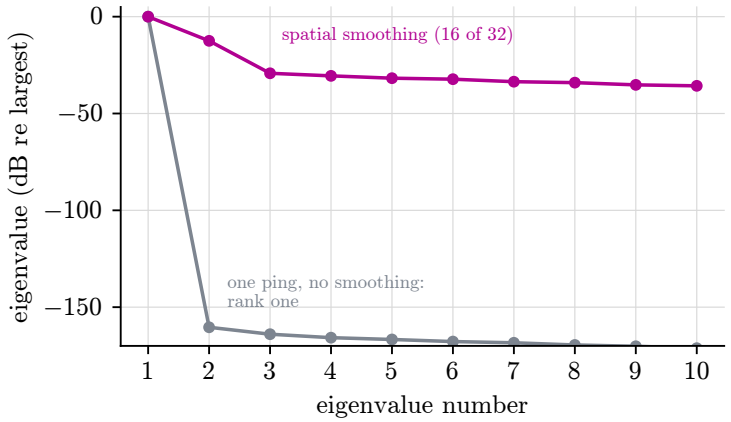


Figure 4: The ten largest eigenvalues of the covariance matrix of one ping at the range of the pair (mean over the frequency bins). Without smoothing, the matrix has rank one. With forward-backward spatial smoothing, a second eigenvalue appears for the second reflector.

reversed and conjugated array, $\boldsymbol{R}_{fb} = \frac{1}{2}\left(\boldsymbol{R}_f + \boldsymbol{J}\boldsymbol{R}_f^*\boldsymbol{J}\right)$ with $\boldsymbol{J}$ the exchange matrix, which decorrelates a further factor of two. With subarrays of 16 out of 32 elements (17 subarrays), the second eigenvalue rises to -12 dB (Fig. 4), and both methods now see two sources (Fig. 5b). With the reflectors 5° apart at 30 dB SNR, MUSIC separates them with a dip of 16.3 dB and MVDR with 4.2 dB, while delay-and-sum, even with the full aperture, shows a single peak (dip 0.0 dB). In this simulation, delay-and-sum needs about 8° to separate the two, because the lower frequencies of the call have a wider beam.

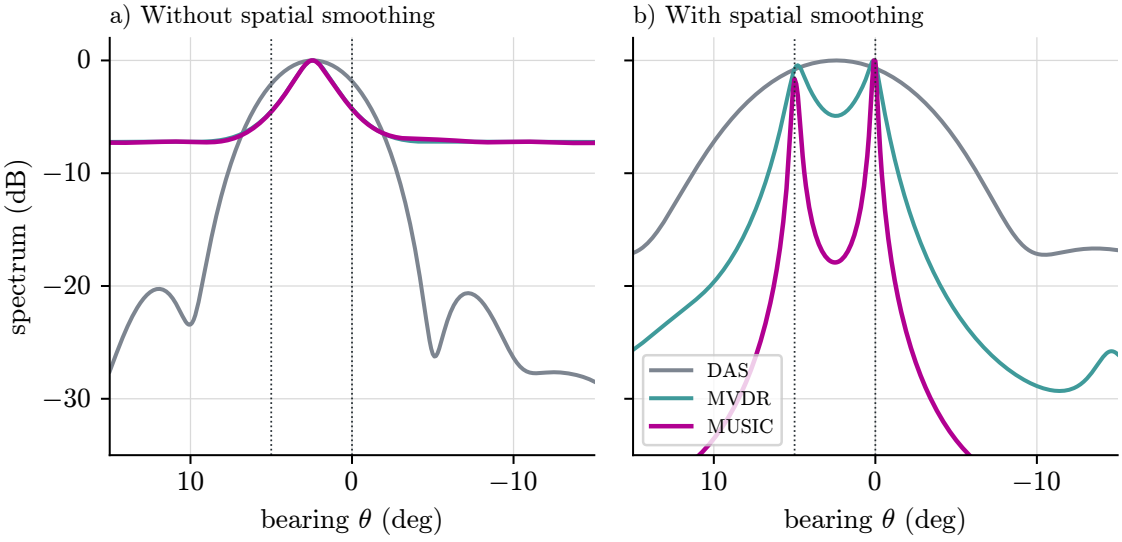


Figure 5: Bearing spectra at 2 m, averaged over the frequency bins from 40 to 75 kHz (each bin normalized to its maximum). Dotted: the true bearings, 5° apart. a) Without smoothing, one ping gives a rank-one covariance, and all methods merge the two reflectors. b) With forward-backward spatial smoothing (16 of 32 elements), MUSIC and MVDR separate them. Delay-and-sum on the smoothed covariance only gets wider, because the subarray is smaller.

Smoothing has a price: the effective aperture shrinks from M to L elements, and so does the resolution of every method. The choice of L is a compromise. More subarrays decorrelate better, fewer but larger ones keep more aperture. Forward-backward smoothing can handle up to about $2(M - L + 1)$ coherent sources, which must also be fewer than L.

Key idea 2

A single sonar ping gives coherent echoes and a rank-deficient covariance matrix, and MVDR and MUSIC fail on it. Spatial smoothing averages over subarrays instead of over time, restores the rank, and gives the methods their resolution back, at the cost of aperture.

8 Smoothing Needs a Regular Array

Spatial smoothing has a strict requirement. All subarrays must be identical copies of each other, shifted along the array, because only then is the steering vector of a reflector the same for every subarray, up to a phase. Figure 6 shows what that means in two dimensions. On a regular 8×8 array, four shifted 5×5 windows hold the same 25 microphones at the same relative positions, and stacked on one origin they coincide exactly (panel b). The same four windows on the real eRTIS layout hold 12, 10, 12, 11 microphones at unrelated positions (panel d). Their covariance matrices belong to four different arrays, and their average no longer has the structure $\boldsymbol{A}\boldsymbol{R}_s\boldsymbol{A}^{\mathsf{H}}$ of any one of them.

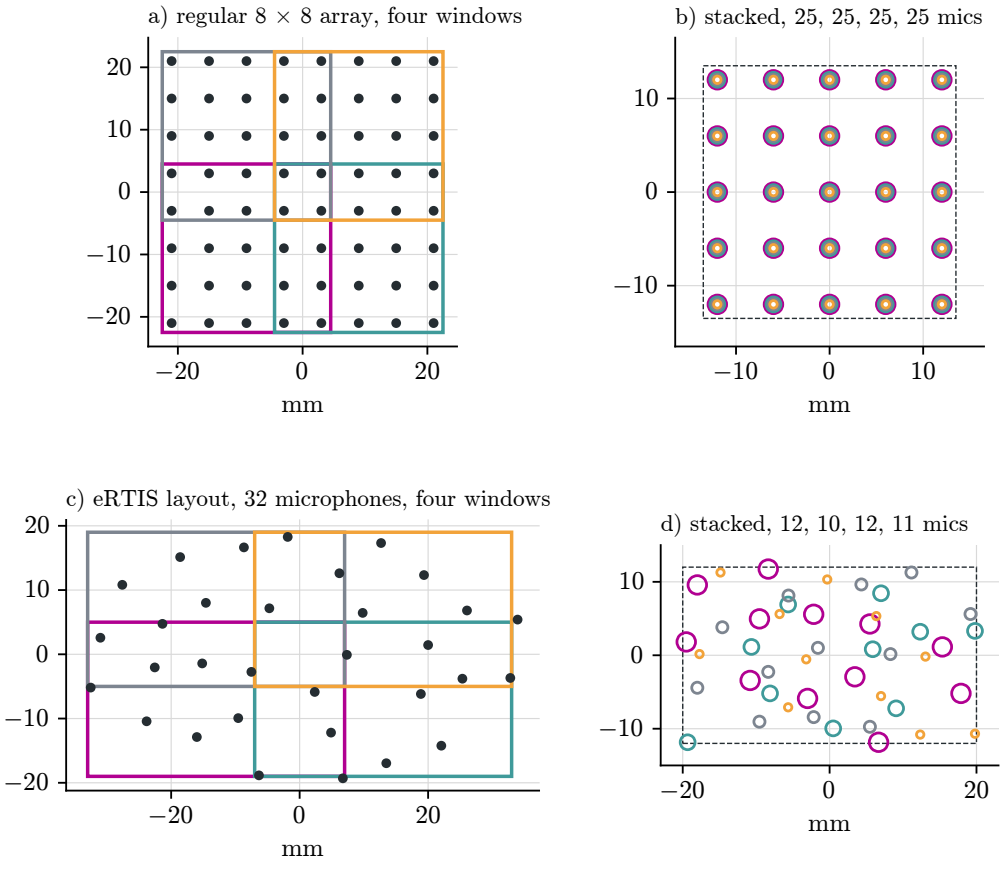


Figure 6: Spatial smoothing needs shifted copies of one subarray. a, b) On a regular 8×8 array (6 mm pitch), four shifted windows contain identical subarrays. c, d) The same idea on the 32 microphones of the eRTIS: the windows hold different numbers of microphones at unrelated positions, so there is no common steering vector.

The eRTIS is random on purpose. A sparse random array spreads its microphones over a large aperture without grating lobes, because it has no periodicity for them to lock on to, and the broadband call averages its sidelobes down. Why that works, and what it costs, is the topic

of Note 11, *Why Random Arrays?*; here only its consequence matters. For MVDR or MUSIC with smoothing on single pings, we need the opposite: a regular array, and a dense one, because a regular array with a spacing above $\lambda/2$ has grating lobes. At 80 kHz, half a wavelength is only 2.14 mm, so covering the two-dimensional aperture of an eRTIS (about 66×38 mm) at that density would take about 558 microphones. That is why dense high-resolution arrays for in-air sonar are small (as the 5×6 grid of the µRTIS), linear, or cross-shaped. The choice of the array is therefore also a choice of the processing.

9 Sparse Reconstruction With an L1 Penalty

Sparse reconstruction starts from a different observation. A scene seen by a sonar consists of a few strong reflectors, much fewer than the directions we can look in [8]. Instead of estimating a covariance matrix, we describe the measurement as a combination of the responses of single reflectors on a grid of possible positions, and look for the combination that uses as few of them as possible. Two versions exist, depending on which data we describe.

In the image domain. Our own first use of the idea worked on the sonar image [9], and on the acoustic image of a leak camera [10]. The energyscape $\boldsymbol{E}$, the image produced by delay-and-sum, is modelled as a weighted sum of the energyscapes of single reflectors,

$$\boldsymbol{E} = \boldsymbol{D}\,\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{10}$$

where every column of the dictionary $\boldsymbol{D}$ is the energyscape of one unit reflector at one grid position, in other words the PSF of the whole processing chain at that position, including the near field, the bandwidth and the array layout. The vector $\boldsymbol{x}$ holds one weight per grid position, and the non-zero ones mark the reflectors. With many more grid positions than independent image values, the system has infinitely many solutions, and we pick the sparsest plausible one:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\geq 0} \; \tfrac{1}{2}\big\|\boldsymbol{D}\boldsymbol{x} - \boldsymbol{E}\big\|_2^2 + \lambda\|\boldsymbol{x}\|_1. \tag{11}$$

Why the L1 norm $\|\boldsymbol{x}\|_1 = \sum_i |x_i|$? Counting the non-zero entries directly (the L0 norm) would be the natural measure of sparsity, but minimizing it means trying every combination of grid positions. The familiar L2 penalty of Tikhonov regularization keeps the problem easy but spreads the energy over many small entries, which is the blur we wanted to remove. The L1 norm is the closest convex substitute for the count: the problem stays convex, and its solutions have exact zeros [11]. The weight λ sets the balance. A larger λ gives a sparser solution, a smaller one follows the data more closely. The 2014 papers solved Eq. (11) with an interior-point method [12]. Here, we use the simpler iterative shrinkage algorithm FISTA [13]: a gradient step on the quadratic term, followed by a soft threshold that subtracts λ (scaled by the step size) from every entry and sets the negative results to zero.

The image-domain model has one weak point. An energyscape is a power, and powers of coherent echoes do not add:

$$\big|\boldsymbol{a}^{\mathsf{H}}(\boldsymbol{x}_1 + \boldsymbol{x}_2)\big|^2 = \big|\boldsymbol{a}^{\mathsf{H}}\boldsymbol{x}_1\big|^2 + \big|\boldsymbol{a}^{\mathsf{H}}\boldsymbol{x}_2\big|^2 + 2\,\mathrm{Re}\big\{(\boldsymbol{a}^{\mathsf{H}}\boldsymbol{x}_1)(\boldsymbol{a}^{\mathsf{H}}\boldsymbol{x}_2)^*\big\}. \tag{12}$$

The cross term vanishes when the two echoes do not overlap in time, which is the case for reflectors at different ranges, and it partly averages out over the band. For two reflectors at the same range it remains, and the solver explains it with extra or displaced reflectors, as the 2014 paper already observed for its measurements.

In the signal domain. The cross term disappears if we apply the same idea before squaring, to the microphone snapshots themselves [14]. Per frequency bin, Eq. (4) with the exact response of every grid position as the columns of $\boldsymbol{A}_f$ is linear in the complex amplitudes $\boldsymbol{s}_f$. The same

few grid positions should be active in all bins, so we penalize, for every grid position g, the norm of its amplitudes over the bins:

$$\min_{\{\boldsymbol{s}_f\}} \sum_f \tfrac{1}{2}\|\boldsymbol{x}_f - \boldsymbol{A}_f \boldsymbol{s}_f\|_2^2 + \lambda \sum_g \Big(\sum_f |s_{g,f}|^2 \Big)^{1/2}. \tag{13}$$

This group penalty is an L1 norm over the grid positions of an L2 norm over the bins. Nothing is averaged, so coherent echoes do no harm, and one ping is enough.

Figure 7 compares both versions on a grid of 241 bearings, 0.5° apart, with λ at 0.01 (image domain) and 0.05 (signal domain) times the value above which the solution is empty. For the equal pair (panel a), delay-and-sum shows one peak. The image-domain solution finds two reflectors, but at 4, 1°, pulled together by the cross term. The signal-domain solution puts them at 4.5, 0°. In panel b, a reflector at 0° has two weaker neighbours, one 6 dB down at 14° and one 12 dB down at −9°, which hides in the shoulder of the delay-and-sum mainlobe. The image-domain solution returns peaks at 15, 0, −3°, a phantom near the strong reflector instead of the weak one. The signal-domain solution finds all three, at 14.5, 13.5, 0, −8.5° (the reflector at 14° spreads over two grid points), at levels of −20, −10, 0, −14 dB. The levels of the weak reflectors come out too low, because the L1 penalty shrinks every amplitude by the same amount.

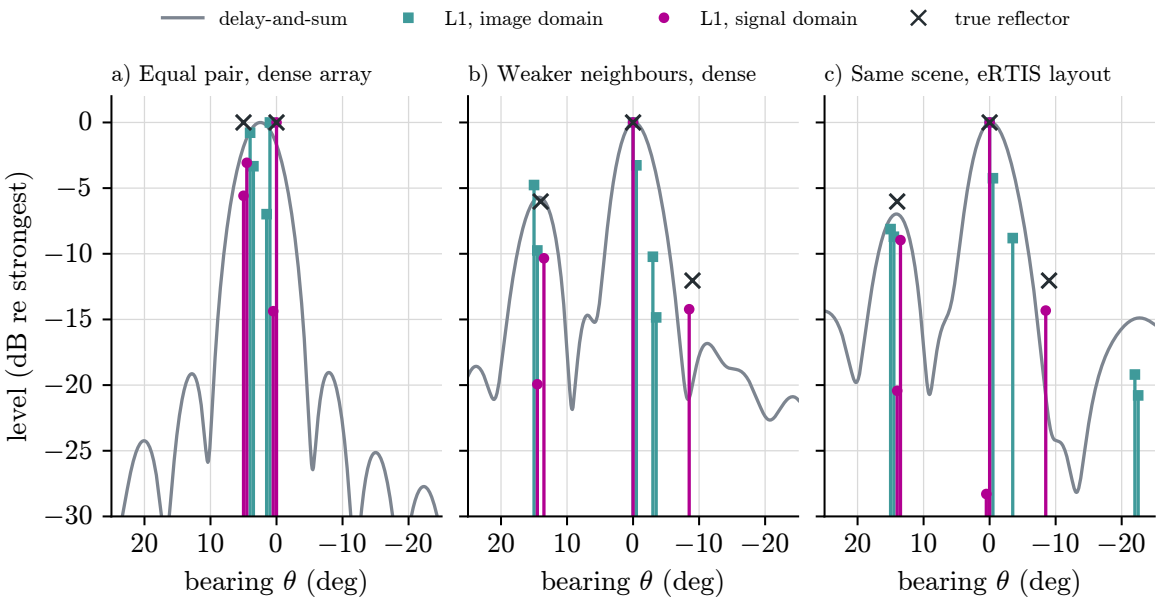


Figure 7: Sparse reconstruction from one ping at 2 m (30 dB SNR). Grey: delay-and-sum energy. Teal: L1 in the image domain (energyscape dictionary). Magenta: L1 in the signal domain (snapshots per bin, group sparsity over the bins). Crosses: the true reflectors. a) The equal pair 5° apart. b) A strong reflector with weaker neighbours at 14° (−6 dB) and −9° (−12 dB). c) The same scene on the horizontal positions of the 32 eRTIS microphones.

Panel c repeats the scene with the horizontal positions of the real eRTIS microphones, a random layout with the same aperture. Spatial smoothing is impossible on it, but the dictionary does not care where the microphones are: the signal-domain solution again finds the reflectors at 13.5, 0, −8.5°.

Key idea 3

Sparse reconstruction replaces the covariance matrix by a model of the scene as a few reflectors on a grid and an L1 penalty that keeps that number small. It works on a single ping and on any array layout, including a random one. Applied to the sonar image it is simple but suffers from the cross terms of coherent echoes; applied to the microphone snapshots it is exactly linear.

10 What It Costs

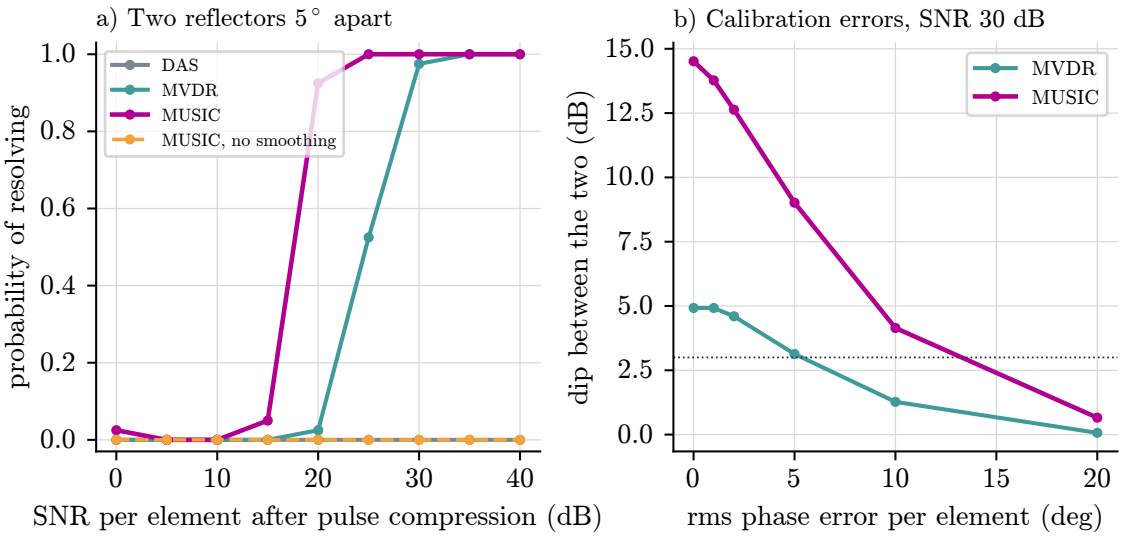


Figure 8: a) Probability that the two reflectors (5° apart) are resolved (two peaks within 1.5° of the true bearings, separated by a dip of at least 3 dB), over 40 noise realizations per SNR. b) The dip between the two for a random phase error per microphone (a calibration mismatch), mean of 20 realizations. Dotted: 3 dB.

None of these methods is free.

Signal-to-noise ratio. Their resolution depends on the SNR (Fig. 8a). In our example, MUSIC resolves the pair reliably (in 90% of the realizations) from an SNR of 20 dB per element after pulse compression, MVDR from 30 dB, and delay-and-sum not at all up to 40 dB. Without smoothing, MUSIC never resolves the pair, whatever the SNR.

Calibration. All of them trust a model of the array, the steering vectors or the dictionary, and any mismatch with the real array (positions, gains, phases of the microphones) costs resolution. A random phase error of 5° rms per microphone reduces the dip of MUSIC from 14.5 to 9.0 dB, and 20° wipes it out (Fig. 8b). Delay-and-sum degrades much more gracefully.

Model order and regularization. MUSIC needs the number of reflectors in the window, and choosing it too small merges reflectors while choosing it too large creates spurious peaks. Sparse reconstruction hides the same choice in λ. In a cluttered scene, the right value changes from one range cell to the next.

The grid. A sparse solution can only place reflectors on its grid. A reflector between two grid points spreads over both, and a finer grid makes neighbouring columns of the dictionary so similar that the solver can no longer choose between them.

Computation. MVDR and MUSIC need, for every range cell and every frequency bin, an inversion or an eigendecomposition of an $L \times L$ matrix. Sparse reconstruction needs hundreds to thousands of iterations, each a multiplication with the whole dictionary, and in 3D the

dictionary has a column for every range and direction. Both are orders of magnitude more work than delay-and-sum, although feasible on a GPU.

Amplitudes. The MUSIC pseudo-spectrum carries no amplitude information, and the L1 penalty biases the amplitudes of weak reflectors down (a least-squares refit on the found positions removes the bias). MVDR estimates powers, and delay-and-sum gives them without any bias.

Table 2: Four ways to find reflectors with an array, for one sonar ping.

	delay-and-sum	MVDR	MUSIC	L1, signal domain
resolution	$\approx \lambda/D$	better at high SNR	very high	very high
output	true power	power estimate	pseudo-spectrum	sparse amplitudes
coherent echoes	no problem	needs smoothing	needs smoothing	no problem
array layout	any	regular (smoothing)	regular (smoothing)	any
extra choices	none	loading ϵ	order K	λ, grid
robustness	high	moderate	low	moderate

Key idea 4

For a single ping, MVDR and MUSIC need a dense regular array and spatial smoothing; sparse reconstruction works with any layout. All of them buy resolution with SNR, calibration effort, a model choice and computation, and none of them replaces delay-and-sum as the honest estimate of power (Table 2).

Exercise

1. Two reflectors are at different ranges. Are their echoes still coherent in the window of the other? What does that mean for the need for spatial smoothing, and for the cross term of the image-domain L1 model?

2. A robot drives at 0.5 m/s and pings 15 times per second. How many wavelengths at 60 kHz does it move between pings? Could the pings ever be used as independent snapshots?

3. Show that MVDR with diagonal loading, $\boldsymbol{R}+\epsilon\boldsymbol{I}$, tends to delay-and-sum for $\epsilon \to \infty$.

4. Why does forward-backward smoothing decorrelate twice as many sources as forward smoothing alone? Consider what conjugating and reversing does to the phase of each source.

5. Take a dictionary with two columns, $[1, 0.9]^T$ and $[0.9, 1]^T$, and the data $[1, 0.9]^T$. Compare the solutions with an L2 penalty and with an L1 penalty for a few values of λ. Which one finds the single reflector?

6. Design a cross-shaped array of 64 microphones for azimuth and elevation with spatial smoothing in both directions. What aperture do you get at 80 kHz, and what is the price compared with a 32-microphone random array and sparse reconstruction?

References

[1] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," Proc. IEEE, 57(8), 1408–1418, 1969.

[2] R. O. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Trans. Antennas Propag., 34(3), 276–280, 1986.

[3] H. Krim and M. Viberg, “Two decades of array signal processing research,” IEEE Signal Process. Mag., 13(4), 67–94, 1996.

[4] I. S. Reed, J. D. Mallett and L. E. Brennan, “Rapid convergence rate in adaptive arrays,” IEEE Trans. Aerosp. Electron. Syst., 10(6), 853–863, 1974.

[5] M. Wax and T. Kailath, “Detection of signals by information theoretic criteria,” IEEE Trans. Acoust., Speech, Signal Process., 33(2), 387–392, 1985.

[6] T.-J. Shan, M. Wax and T. Kailath, “On spatial smoothing for direction-of-arrival estimation of coherent signals,” IEEE Trans. Acoust., Speech, Signal Process., 33(4), 806–811, 1985.

[7] S. U. Pillai and B. H. Kwon, “Forward/backward spatial smoothing techniques for coherent signal identification,” IEEE Trans. Acoust., Speech, Signal Process., 37(1), 8–15, 1989.

[8] A. Elfes, “Sonar-based real-world mapping and navigation,” IEEE J. Robot. Autom., 3(3), 249–265, 1987.

[9] J. Steckel and H. Peremans, “Sparse decomposition of in-air sonar images for object localization,” Proc. IEEE SENSORS, 2014.

[10] J. Steckel and H. Peremans, “Ultrasound-based air leak detection using a random microphone array and sparse representations,” Proc. IEEE SENSORS, 1026–1029, 2014.

[11] R. Tibshirani, “Regression shrinkage and selection via the lasso,” J. R. Stat. Soc. B, 58(1), 267–288, 1996.

[12] S.-J. Kim, K. Koh, M. Lustig, S. Boyd and D. Gorinevsky, “An interior-point method for large-scale ℓ_1-regularized least squares,” IEEE J. Sel. Topics Signal Process., 1(4), 606–617, 2007.

[13] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” SIAM J. Imaging Sci., 2(1), 183–202, 2009.

[14] D. Malioutov, M. Çetin and A. S. Willsky, “A sparse signal reconstruction perspective for source localization with sensor arrays,” IEEE Trans. Signal Process., 53(8), 3010–3022, 2005.

[15] H. L. Van Trees, *Optimum Array Processing*, Wiley, 2002.