SPARSE MULTICHANNEL SOURCE LOCALISATION AND SEPARATION

Ruairí de Fréin †, *Scott Rickard* †, *Barak Pearlmutter* ††

†Complex & Adaptive Systems Laboratory, University College Dublin, Ireland.

††Hamilton Institute, National University of Ireland, Maynooth, Co. Kildare, Ireland.

ABSTRACT

The DUET and DESPRIT blind source separation algorithms attempt to recover J sources from I mixtures of these sources, in the interesting case where $J > I$, with minimal information about the mixing environment or underlying source statistics. We present a semi-blind generalisation of the DUET-DESPRIT approach which allows arbitrary placement of the sensors and demixes the sources given the room impulse response. We learn a sparse representation of the mixtures on an over-complete spatial signatures dictionary. We localise and separate the constituent sources via binary masking of a power weighted histogram in location space or in attenuation-delay space. We demonstrate the robustness of this technique using synthetic room experiments.

1. INTRODUCTION

The DUET algorithm performs source separation with notable success using a pair of closely spaced microphones [6]. Extending the DUET approach to leverage observations from $I > 2$ sensors has been addressed in [7] with the proviso that the sensors are constrained to lie in a linear array with a regular-spacing. DUET and DESPRIT assume that the wave propagating to the sensor array obeys the narrow-band assumption. In this work we loosen this constraint so that observations from arbitrarily placed microphones can be leveraged along with the *DUET-pair* of microphones to perform source localisation and separation. Our technique treats channel characteristics as cues as opposed to obstacles for source localization similar in spirit to the approach in [8]. We consider the FOCUSS [5] re-weighted minimum norm methodology for sparse signal representations. Our technique is general, in that you can find a "good" "decomposition", where "good" is a task-dependent measure of your choice, e.g. not just in the L_1 norm sense, but localisation or separation quality, and where "decomposition" is arbitrary, e.g. $\mathrm{L}_1 + \lambda \mathrm{L}_2$ optimization or some more complicated conic optimization, for example, by considering a weighted $\mathrm{L}_1 + \lambda \mathrm{L}_2^2$ initialized $\mathrm{L}_{0.5} + \lambda \mathrm{L}_2^2$ objective for super-sparse resolution. We introduce our notation. A continuous time signal $s(t)$ is denoted by $s(nT) = s[n]$ where $n = 0, 1, 2, \ldots$ in the discrete time domain where T is the sampling period. A continuous time signal delayed by $\delta \in \mathbb{R}$ seconds is denoted by $s(t - \delta)$. In discrete time we define

$$s^{\delta}[n] = s(nT - \delta), \tag{1}$$

which is $s[n - \delta/T]$ if $\delta/T \in \mathbb{Z}$. A non-integer sample delay can be performed using sinc-interpolation given that the signal is bandlimited and sampled at a sufficiently high sampling rate.

$$s^{\delta}[n] = \sum_{n=-\infty}^{\infty} s[n]\mathrm{sinc}(nT - \delta). \tag{2}$$

Supported by Science Foundation Ireland Grant No. 05/Y12/1677.

In practice a finite length approximation of the sinc function leads to error in the estimate of $s^{\delta}[n]$. A non-integer delay of a bandlimited signal sampled above Nyquist rate can also be determined using a Discrete Fourier Transform (DFT). Multiplying $\mathrm{DFT}\{s[n]\} = S[k] = \sum_{n=0}^{N-1} s[n]W^{kn}$ where $W = e^{-j\frac{2\pi}{N}}$, by a linear phase term W^{kd} in the discrete frequency domain, corresponds to a circular shift of the signal by $d = \delta/T \in \mathbb{R}$ samples. Zero-padding in the time domain and consequently increasing the resolution in the frequency domain, taking the DFT and multiplying by the linear phase term gives the signal shifted in time. We define the function $\mathrm{ZP}(a, s[n], b)$ which appends a zeros to the beginning and b zeros to the end of $s[n]$. The function $\mathrm{IP}(a, s[n], b)$ removes a samples from the beginning and b samples from the end of $s[n]$.

$$s^{\delta}[n] = \mathrm{IP}\left(a, \mathrm{IDFT}\{\mathrm{DFT}\{\mathrm{ZP}(a, s[n], b)\}W^{kd}\}, b\right), \tag{3}$$

where $\mathrm{IDFT}\{S[k]\} = \frac{1}{N}\sum_{k=0}^{N-1} S[k]W^{-kn}$ is the inverse DFT. $s^{\delta}[n]$ is calculated *exactly* using the *discrete frequency domain method* (Eq. 3).

We perform localisation in the short-time-frequency domain where speech has improved Windowed Disjoint Orthogonality (WDO) [6] and sparsity [9]. Consequently we need to construct a time-frequency dictionary which represents the environmental attenuation and delay effects to simulate the propagation effects on the source signals in time-frequency. The Short-Time-Fourier-Transform (STFT) of $s[n]$ is,

$$\mathrm{STFT}\{s[n]\} = S[k, m] = \sum_{n=mR}^{N-1-mR} s[n]w[n - mR]W^{k(n-mR)}. \tag{4}$$

$S[k, m]$ is the STFT of a windowed signal positioned at sample mR where $w[n]$ is a window function and R is the number of hop-size samples. [k,m] are the discrete frequency and time indices respectively. The focus of this paper is to generalize the DUET and DESPRIT frame-work so that information from singleton microphones arbitrarily spaced in the environment can be leveraged to localize and separation sources. Recently submitted work [1] discusses the bias inherent in the DUET-type approach and addresses the issue of bias free spatial signature construction in time-frequency. We refer the interested reader to the companion paper [1] which discusses time-frequency spatial signature dictionary construction without windowing and wrap-around effects.

2. MIXING MODEL

In an anechoic environment, a continuous time source signal $s_j(t)$ is attenuated and delayed as it propagates the direct path to sensor x_i. The attenuation and delay effect on the j^{th} source received at the i^{th} sensor is (a_{ji}, δ_{ji}) and so $\hat{s}_{ji}(t) = a_{ji}s_j(t - \delta_{ji})$, yielding the

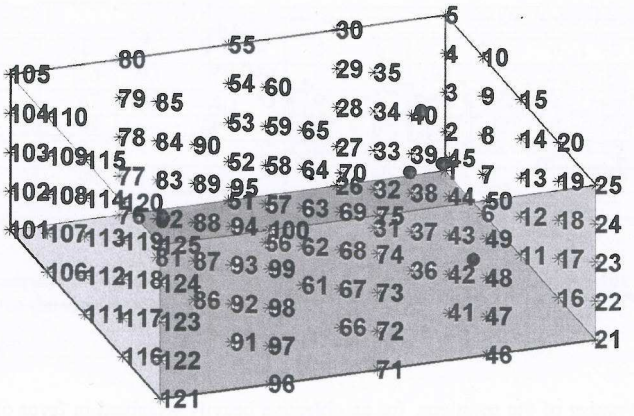


Fig. 1. Fig. 1 illustrates 2m × 2m × 2m room with grid positions every 0.5m. Sensors pairs are positioned arbitrarily (denoted by a star) so that we can perform a DUET de-mixing using each of these pairs.

mixtures,

$$x_i(t) = \sum_{j=1}^{J} \hat{s}_{ji}(t) = h_{ji}(t) \star s_j(t) \quad (5)$$

I mixture signals are observed, $x_i(t)$, at physical locations x_i, where $h_{ji}(t)$ is the continuous time transfer function from source s_j to sensor x_i. The source is constrained to lie at one of P grid points, indicated by the numeric labels in the scene illustrated in Fig. 1. Sensors pairs are placed arbitrarily in this 2m × 2m × 2m room. The grid points $p = 1 \ldots P$ are arranged with variable spatial resolution. Consider a dedicated teleconferencing room, with an arbitrary number of inexpensive microphones, where the source location, detected using a sensor array, is used to automatically identify the speaker or indicate the position of the speaker.

2.1. Time-Frequency Mixing Model

We form the vector,

$$\boldsymbol{x}[k,m] = [X_1[k,m], \ldots, X_i[k,m], \ldots, X_I[k,m]]^T \in \mathbb{C}^{I \times 1} \quad (6)$$

for each time-frequency point $[k, m]$. We construct a spatial signatures matrix for each $[k, m]$, $\boldsymbol{D}[k,m] \in \mathbb{C}^{I \times P}$,

$$\boldsymbol{D}[k,m] = \begin{array}{c} \\ \text{Sensor i} \\ \downarrow \end{array} \overbrace{\begin{pmatrix} H_{11}[k,m] & \ldots & H_{p1}[k,m] & \ldots & H_{P1}[k,m] \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ H_{1i}[k,m] & \ldots & H_{pi}[k,m] & \ldots & H_{Pi}[k,m] \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ H_{1I}[k,m] & \ldots & H_{pI}[k,m] & \ldots & H_{PI}[k,m] \end{pmatrix}}^{\longleftarrow \text{Location p} \longrightarrow} . \quad (7)$$

$\boldsymbol{D}[k,m]$ gives the spatial signature, $H_{pi}[k,m]$, for every location, p, in the grid relative to each sensor, x_i, for $[k,m]$. For example, (Eq. 7) gives the dictionary constructed for a room with P potential grid locations and I sensors (Fig. 1). The J sources $[\mathrm{s}_1, \ldots \mathrm{s}_j \ldots, \mathrm{s}_J]$ are constrained to lie at a subset of the P grid points. Given the observation matrix, $\boldsymbol{x}[k,m]$, and the spatial signature matrix, $\boldsymbol{D}[k,m]$, we locate the source by learning the vector $\boldsymbol{c}[k,m] \in \mathbb{C}^{P \times 1}$ which explains the sensor observations in the most parsimonious manner given the spatial signatures dictionary,

$$\boldsymbol{x}[k,m] = \boldsymbol{D}[k,m]\boldsymbol{c}[k,m]. \quad (8)$$

We solve each subsystem $[k,m]$ independently. This approach lends itself to real-time implementation perhaps in parallel on a dedicated set of processors. Assuming WDO in time-frequency, a single source is active in $[k,m]$. The element of $\boldsymbol{c}$ with the most energy, for example $p = 25$, indexed by $\boldsymbol{c}_p[k,m]$, indicates the position of the source and appropriate transfer function.

3. SPARSE SOLUTIONS AND WINDOWED DISJOINT ORTHOGONALITY

We desire a solution to the system of equations (Eq.8) which reveals the locations p of the latent source signals $s_j[n]$. Each atom of the matrix $\boldsymbol{D}[k,m]$ describes the direct-path propagation effects between the sensors and the room grid points, a vector $\boldsymbol{c}_p[k,m]$ that satisfies (Eq.8) and that captures a large percentage of the mixture energy in a small number of elements of $\boldsymbol{c}[k,m]$ indicates the location of the source signals. We drop the time-frequency notation for convenience henceforth.

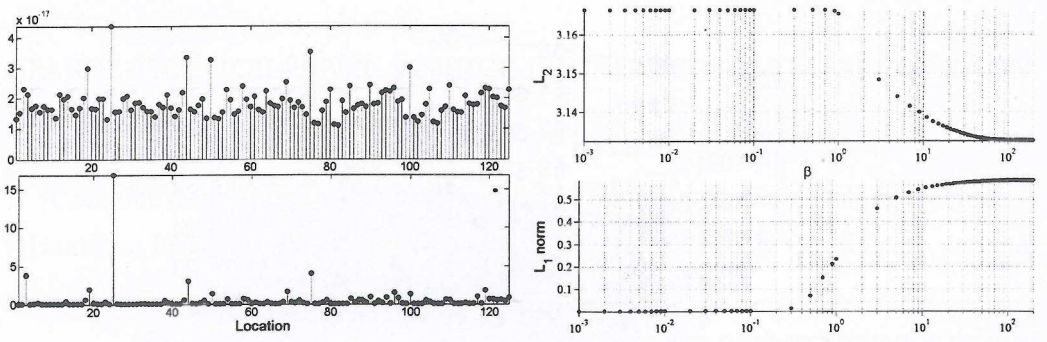
The system (Eq.8) has infinitely many solutions. One such solution is the pseudo inverse $\boldsymbol{D}^{\dagger}\boldsymbol{x} = \boldsymbol{c}$, which sheds little light on the source locations as it generally yields a dense vector $\boldsymbol{c}$. The authors show that the L_2 norm solution spreads the energy across many of the dictionary elements yielding an un-interpretable and unseparated solution in [8]. Many possible combinations of the dictionary elements can be combined to form the signal. We constrain the solution-space to reflect the assumption that as few of the atoms of $\boldsymbol{D}[k,m]$ as possible should be used to explain the mixture $\boldsymbol{x}[k,m]$. Regularization is used to enforce this prior information about the form of the solution.

The L_0 norm, denoted by $|\boldsymbol{c}|_0$, is the ideal sparsity measure as it counts the number of non-zero elements in $\boldsymbol{c}$. Formally the L_0 optimization problem maybe stated as

$$\min_{\boldsymbol{c}} |\boldsymbol{c}|_0 = \sum_{p=1}^{P} |\boldsymbol{c}|^0 \text{ subject to } \boldsymbol{D}\boldsymbol{c} = \boldsymbol{x}. \quad (9)$$

This solution is shown to be unique when $\boldsymbol{c}$ is sufficiently sparse and coincides with the solution of,

$$\min_{\boldsymbol{c}} |\boldsymbol{c}|_1 = \sum_{p=1}^{P} |\boldsymbol{c}|^1 \text{ subject to } \boldsymbol{D}\boldsymbol{c} = \boldsymbol{x}. \quad (10)$$



(a) (b)

Fig. 2. Fig. 2(a) compares the power of the solutions, for an objective heavily weighted in favor of the L_2^2 norm in row 1 and an objective heavily weighted in favor of a sparse objective in row 2. The source is located at position 25 on the x-axis. The sparse solution reveals the source location. Fig. 2(b) illustrates the accuracy of the estimate at position 25 as the objective becomes weighted in favor of a sparse objective. The L_2 norm distance between the true solution and estimate (in Fig. 2(b) row 1) decreases as the sparsity (which is measured using the L_1) increases (in Fig. 2(b) row 2)

in [2, 4, 3]. This is a significant result as the L_1 relaxation (10) is a convex optimization problem and a global minimum can be found for real-valued data by linear programming. The L_0 optimization problem (9) is a combinatorial optimisation problem. The gradient gives no information about the direction coefficients should move to improve the solution using a gradient descent updating procedure.

Using the L_1 norm is equivalent to assuming there is a Laplacian prior on the coefficients. Noise can be included in the objective function as the L_1 norm gives an exact solution – which includes background noise – and can generate large artefacts. The weighted mixed $\lambda L_1 + L_2^2$ objective achieves a trade-off between the level of sparsity enforced on the solution and the fidelity of the solution. Careful tuning of the weighting parameter accounts for the background noise reducing the number of artefacts due to the L_1 term.

$$\min_{\boldsymbol{c}} \lambda|\boldsymbol{c}|_1 + ||\boldsymbol{D}\boldsymbol{c} - \boldsymbol{y}||_2^2 \tag{11}$$

Fig. 2 illustrates the trade-off between sparsity in location and accurately representing the mixtures by projection onto the spatial signatures dictionary. There are 125 potential locations in the room and 1 active source. The accuracy in source estimate at the correct source location increases as the objective becomes heavily weighted in favor of a sparsity objective. Sparsity in location reveals the source location which gives an accurate estimate of the source.

3.1. A Sparse Prior

Speech is generally sparser in time-frequency than in the time or frequency domain [9]. Sparsity is defined according to the data being used in a specific application. In this work a vector is considered to be sparse if only a few of its elements are non-zero or greater than a small threshold value. The elements greater than the small threshold contain most of the power of the signal. In this work we assume sparsity in location space (Eq. 8). We posit that only one source can be active at any location in our set of source locations. Given that $P \gg J$ the sparse solution to (Eq. 8) reveals the source location. The Windowed Disjoint Orthogonality assumption posits that speech rarely overlaps in the time-frequency domain,

$$S_b[k,m]S_l[k,m] = 0 \quad \forall k, m, b \neq l. \tag{12}$$

This property, whilst only approximate is used by the DUET algorithm to partition the time-frequency scene of a mixture of speakers. We use a combination of sparsity in location and WDO in time-frequency to localise and then separate speech via time-frequency masking.

3.2. An objective function for sparse localisation

We consider a mixed L_1-L_2 objective regularized least squares iterative solver.

$$E(\boldsymbol{c}) = \arg\min_{\mathbf{c}} \frac{1}{2} \underbrace{\sum_{i=1}^{P} |\mathbf{c}_i|}_{E_1} + \lambda \underbrace{||\mathbf{D}\mathbf{c} - \mathbf{y}||_2^2}_{E_2}, \tag{13}$$

where $\boldsymbol{c} = \mathrm{Re}\{\boldsymbol{c}\} + \mathrm{Im}\{\boldsymbol{c}\}j$, and $\boldsymbol{D} = \mathrm{Re}\{\boldsymbol{D}\} + \mathrm{Im}\{\boldsymbol{D}\}j$, (where the symbols, $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$, denote the real and imaginary parts of the variable) and it is clear from the context that $j = \sqrt{-1}$. Equivalently, we represent the system (Eq. 8) as,

$$\begin{pmatrix} \mathrm{Re}\{\hat{\boldsymbol{y}}\} \\ \mathrm{Im}\{\hat{\boldsymbol{y}}\} \end{pmatrix} = \begin{pmatrix} \mathrm{Re}\{\hat{\boldsymbol{D}}\} & -\mathrm{Im}\{\hat{\boldsymbol{D}}\} \\ \mathrm{Im}\{\hat{\boldsymbol{D}}\} & \mathrm{Re}\{\hat{\boldsymbol{D}}\} \end{pmatrix} \begin{pmatrix} \mathrm{Re}\{\boldsymbol{c}\} \\ \mathrm{Im}\{\boldsymbol{c}\} \end{pmatrix} = \tilde{\boldsymbol{D}} \begin{pmatrix} \mathrm{Re}\{\boldsymbol{c}\} \\ \mathrm{Im}\{\boldsymbol{c}\} \end{pmatrix} \tag{14}$$

3.3. Update

We consider the objective function,

$$\arg\min_{\boldsymbol{c}} \hat{E(\boldsymbol{c})} = \hat{E}_1 + \lambda E_2, \text{ where} \tag{15}$$

$$\hat{E}_1 \equiv \frac{1}{2} \sum_{i=1}^{P} \boldsymbol{\alpha}_i^k |\boldsymbol{c}_i^k|^2 \text{ where } \boldsymbol{\alpha} \in \mathbb{R}^{P \times 1}. \tag{16}$$

We then solve $\frac{\partial \hat{E}(\boldsymbol{c})}{\partial \boldsymbol{c}^k} = \mathbf{0}$ for $\boldsymbol{c}^k$ (where k denotes the iteration index) yielding the linear system

$$\left(\boldsymbol{\Lambda} + \lambda \tilde{\boldsymbol{D}}^T \tilde{\boldsymbol{D}}\right) \begin{pmatrix} \mathrm{Re}\{\boldsymbol{c}\} \\ \mathrm{Im}\{\boldsymbol{c}\} \end{pmatrix} = \lambda \tilde{\boldsymbol{D}}^T \begin{pmatrix} \mathrm{Re}\{\boldsymbol{y}\} \\ \mathrm{Im}\{\boldsymbol{y}\} \end{pmatrix}, \tag{17}$$

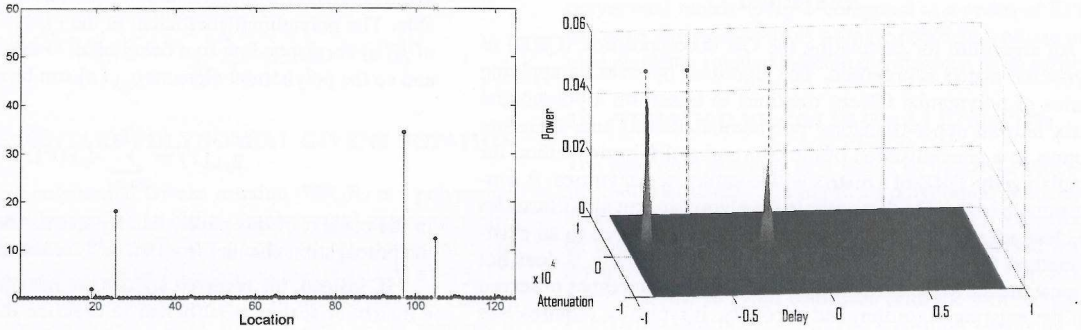
and so in terms of $\boldsymbol{y}, \boldsymbol{D}, \boldsymbol{c}$,

$$(\boldsymbol{\Lambda} + \lambda \boldsymbol{D}^H \boldsymbol{D})\boldsymbol{c} = \lambda \boldsymbol{D}^H \boldsymbol{y}, \tag{18}$$

where $(\cdot)^H$ denotes the conjugate transpose operation. We set $\boldsymbol{\alpha}_i^{k+1} = \frac{1}{|\boldsymbol{c}_i^k|}$ and iterate this procedure until we reach a fixed point.

Source (p)	PSR estimate mask					WDO estimate mask					WDO (0dB) Ideal					Loc %				
	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
$s_1(25)$	.95	.89	.85	.60	.86	.93	.88	.84	.60	.83	.93	.92	.91	.87	.85	31	25	14	8	6
$s_2(97)$	.98	.90	.85	.69	.67	.97	.89	.83	.68	.65	.97	.92	.91	.82	.75	59	47	26	14	11
$s_3(105)$	–	.80	.69	.58	.67	–	.79	.67	.57	.65	–	.87	.83	.79	.69	–	17	9	5	5
$s_4(5)$	–	–	.91	.86	.70	–	–	.87	.85	.68	–	–	.82	.79	.73	–	–	42	24	3
$s_5(62)$	–	–	–	.94	.90	–	–	–	.92	.89	–	–	–	.90	.86	–	–	–	43	54
$s_6(112)$	–	–	–	–	.77	–	–	–	–	.74	–	–	–	–	.78	–	–	–	–	18

Fig. 3. Results for localisation and separation experiments. PSR, WDO and the percentage of the mixture power in teh correct source positions are used to verify the performance of the technique.



(a) (b)

Fig. 4. Localisation of 3 sources from a mixture of 10 pair-wise sensors using spatial signatures (Fig. 4(a)) and the performance of DUET using just one pair of sensors (Fig. 4(b))

4. EXPERIMENTS

We perform source localisation and then separation using 10 sensors and generate synthetic mixtures of length 10 seconds using the spatial signatures from the room Fig. 1. Each mixture contains $2, 3, 4, 5$ or 6 sources from the TIMIT database. We tune the parameterized objective (Eq.13) and perform source localisation initially. We use an FFT of size $N = 2048$ which allows for accurate spatial signatures for sources propagating up to 11m, constructed using the approach in [1]. We arrange the microphones pair-wise in arbitrary locations in the room. We learn a sparse solution to (Eq. 8) for each time-frequency bin $[k, m]$. We generate a power-weighted histogram in location space. We use the percentage power in each location to measure the efficacy of the localisation step. Table 3 lists the percentage power of the mixture located at the correct position of each source s_j. The percentage power for each source is dependent on the original source signal power and propagation effects. This explains the low percentage source signal powers obtained by sources s_3 and s_4 in the 6 source mixture case. Fig.4(a) illustrates the signal power in each location, $25, 97, 105$ for a mixture of 3 sources. Taking the ratio of the mixtures from a pair of sensors can be used to generate the DUET power-weight histogram in Fig. 4(b). Three pins are used to indicate the correct source position in attenuation-delay space. A clearer estimate is found using 10 channels in Fig. 4(a) compared to the DUET estimate in Fig. 4(b). Leveraging the information from 10 channels gives a clearer indication of the source locations. We perform separation by creating a binary mask using the solution from each time-frequency point as an indicator function. We measure separation performance by comparing the binary mask for each channel with the ideal 0 dB mask. Table 3 lists the results for mixtures of 2 to 6 sources. The Perservered Signal Power Ratio (PSR) and W-Disjoint Orthogonality measure are defined in [6].

Table 3 verifies the separation and localisation performance the technique. The PSR and WDO scores obtained are comparable with the ideal (WDO 0dB) score. Although the technique is no longer blind, there are many real-world scenarios where source separation using calibrated spatial signatures is a feasible solution due to frequent use of the room for localisation and separation, for example a dedicated teleconferencing room.

5. REFERENCES

[1] R. de Frein, Scott Rickard, and Barak Pearlmutter. Constructing time-frequency dictionaries for source separation via time-frequency masking and source localisation. 2009. Submitted.

[2] David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences of the United States of America*, 100(5):2197–2202.

[3] David L. Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

[4] M. Elad and A.M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *Information Theory, IEEE Transactions on*, 48(9):2558–2567, Sep 2002.

[5] I. F. Gorodnitsky and B. D. Rao. Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45(3):600–616, 1997.

[6] Alexander Jourjine, Scott Rickard, and Ozgiir Yilmaz. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 2985–2988.

[7] Tom Melia. *Underdetermined Blind Source Separation in Echoic Environments Using Linear Arrays and Sparse Representations*. PhD thesis, University College Dublin, 2007.

[8] Barak A. Pearlmutter and Anthony M. Zador. Monaural source separation using spectral cues. *Independent Component Analysis, Granada, Spain*, pages 478–485, September 2004.

[9] Scott Rickard. Sparse sources are separated sources. In *Proceedings of the 16th Annual European Signal Processing Conference*, Florence, Italy, September 2006.