Extending the DUET Blind Source Separation technique

Thomas Melia and Scott Rickard
Sparse Signal Processing Group
University College Dublin
Ireland

1 Introduction

Blind Source Separation (BSS) techniques attempt to recover N source signals from M mixtures of these sources, with limited prior knowledge of the original sources or the mixing procedure involved. The DUET BSS algorithm may be used to demix N speech sources from only 2 anechoic mixtures of the sources [1]. Limitations of DUET include

1. an inability to utilise more than 2 mixtures when available,
2. a reliance upon the W-Disjoint Orthogonal (WDO) assumption, i.e. sources do not overlap in time-frequency,
3. and a reliance upon an anechoic mixing model.

In this paper we extend the DUET technique to overcome these limitations through the use of a uniform linear sensor array. We seek to recover N source signals $s_1(t), s_2(t), \ldots, s_N(t)$ from M mixtures $x_1(t), x_2(t), \ldots, x_M(t)$ taken from a uniform linear sensor array (see Figure 1). In the frequency domain we may

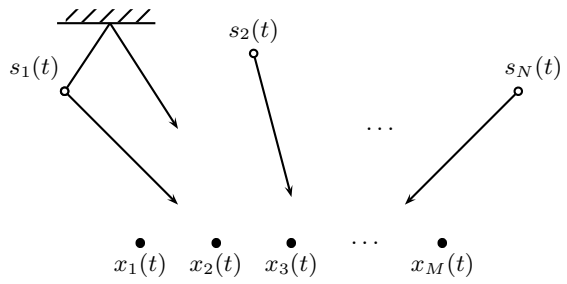

Fig. 1: A uniform linear sensor array receives M mixtures of N (far field) source signals

express the M mixtures as

$$\begin{bmatrix} X_1(\omega) \\ X_2(\omega) \\ \vdots \\ X_M(\omega) \end{bmatrix} = \mathbf{A}(\omega) \begin{bmatrix} S_1(\omega) \\ \vdots \\ S_N(\omega) \end{bmatrix} + \begin{bmatrix} V_1(\omega) \\ V_2(\omega) \\ \vdots \\ V_M(\omega) \end{bmatrix} \quad (1)$$

$$\mathbf{A}(\omega) = \begin{bmatrix} A_1(\omega) & \ldots & A_N(\omega) \\ A_1(\omega)\phi_1(\omega) & \ldots & A_N(\omega)\phi_N(\omega) \\ \vdots & & \vdots \\ A_1(\omega)\phi_1^{M-1}(\omega) & \ldots & A_N(\omega)\phi_N^{M-1}(\omega) \end{bmatrix} \quad (2)$$

where $A_n(\omega) = a_n e^{-j\omega d_n}$, a_n and d_n represent the attenuation and delay experienced by the n^{th} source signal as it propagates to the 1^{st} sensor, $\phi_n(\omega) = \alpha_n e^{-j\omega\delta_n}$, α_n and δ_n represent the attenuation and delay experienced by the n^{th} source signal as it propagates between two adjacent sensors and $V_1(\omega), V_2(\omega), \ldots, V_M(\omega)$ are independently and identically distributed noise terms. Equation (1) describes an anechoic mixing model, this model may be altered to become an echoic mixing model by adding columns to the mixing matrix (2) representing the echoic paths.

$$\mathbf{A}(\omega) = [\mathbf{A}_1(\omega)\, \mathbf{A}_2(\omega) \cdots \mathbf{A}_N(\omega)] \quad (3)$$

$$\mathbf{A}_n(\omega) = \begin{bmatrix} A_{n,1}(\omega) & \cdots & A_{n,P_n}(\omega) \\ A_{n,1}(\omega)\phi_{n,1}(\omega) & \cdots & A_{n,P_n}(\omega)\phi_{n,P_n}(\omega) \\ \vdots & & \vdots \\ A_{n,1}(\omega)\phi_{n,1}^{M-1}(\omega) & \cdots & A_{n,P_n}(\omega)\phi_{n,P_n}^{M-1}(\omega) \end{bmatrix}$$

where $A_{n,p}(\omega) = a_{n,p} e^{-j\omega d_{n,p}}$, $a_{n,p}$ and $d_{n,p}$ represent the attenuation and delay experienced by the n^{th} source signal as it propagates along its p^{th} path to the 1^{st} sensor, $\phi_{n,p}(\omega) = \alpha_{n,p} e^{-j\omega\delta_{n,p}}$, $\alpha_{n,p}$ and $\delta_{n,p}$ represent the attenuation and delay experienced by the n^{th} source signal as it propagates along its p^{th} path between two adjacent sensors and P_n is the number of propagation paths originating from the n^{th} signal which are received by the sensor array.

2 Separation in anechoic environments

2.1 Demixing N Source Signals

In order to demix the N source signals, one frequency domain based approach is to estimate a demixing matrix $\mathbf{B}(\omega)$ such that

$$\mathbf{B}(\omega)\mathbf{A}(\omega) = \begin{bmatrix} C_{1,1}(\omega) & 0 & \cdots & 0 \\ 0 & C_{2,2}(\omega) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & C_{N,N}(\omega) \end{bmatrix}$$

where $C_{1,1}(\omega), C_{2,2}(\omega), \ldots, C_{N,N}(\omega)$ are N complex numbers. In the no noise case an accurate estimate of the demixing matrix $\tilde{\mathbf{B}}(\omega)$, can be used recover the N source signals at frequency ω up to an attenuation and a delay

$$\tilde{\mathbf{B}}(\omega)\begin{bmatrix} X_1(\omega) \\ X_2(\omega) \\ \vdots \\ X_M(\omega) \end{bmatrix} \approx \begin{bmatrix} C_{1,1}(\omega)S_1(\omega) \\ \vdots \\ C_{N,N}(\omega)S_N(\omega) \end{bmatrix}.$$

Given the anechoic mixing model (1) a no noise demixing matrix estimate is of the form

$$\tilde{\mathbf{B}}(\omega) = \begin{bmatrix} 1 & \cdots & 1 \\ \tilde{\phi}_1(\omega) & \cdots & \tilde{\phi}_N(\omega) \\ \vdots & & \vdots \\ \tilde{\phi}_1^{M-1}(\omega) & \cdots & \tilde{\phi}_N^{M-1}(\omega) \end{bmatrix}^{\dagger} \tag{4}$$

where $[\cdot]^{\dagger}$ denotes the Moore-Penrose pseudo inverse (a least squares solution to the tall, non-square matrix inverse) and $\tilde{\phi}_1(\omega), \tilde{\phi}_2(\omega), \ldots, \tilde{\phi}_N(\omega)$ are N mixing parameter estimates. The parameterized formulation in (1) (and similarly in (3)) is useful as it avoids the difficulties of the permutation problem normally associated with frequency domain matrix inversion based BSS methods.

2.2 The DESPRIT technique

The DUET BSS algorithm [1] blindly estimates the mixing parameters for N speech sources in the time-frequency domain, using $M = 2$ sensors. The ESPRIT direction of arrival algorithm [2] may be used blindly estimate the mixing parameters in the time domain for N narrowband sources, using $M > N$ sensors. The DESPRIT BSS technique is a hybrid technique which blindly estimates the mixing parameters of N speech sources in the time-frequency domain, using $M \geq 2$ sensors. Under a hard WDO assumption DESPRIT can be seen as a multichannel extension of the DUET algorithm [3]. DESPRIT may also be implemented under a soft WDO assumption that allows up to $M-1$ sources to coexist at a given time-frequency point [4]. The DESPRIT mixing parameter estimation step involves the construction of a time-frequency data matrix $\mathbf{Z}(\omega, \tau)$ and its subspace decomposition via the singular value decomposition (SVD) of the time-frequency covariance matrix

$$\mathbf{R}_{\mathbf{ZZ}}(\omega, \tau) = \mathrm{E}\left\{ [\mathbf{Z}(\omega, \tau)] [\mathbf{Z}(\omega, \tau)]^H \right\}. \tag{5}$$

The data matrix is constructed in the following way

$$\mathbf{Z}(\omega, \tau) = \begin{bmatrix} X_1(\omega, \tau) \\ \vdots \\ X_{M-1}(\omega, \tau) \\ X_2(\omega, \tau) \\ \vdots \\ X_M(\omega, \tau) \end{bmatrix} \tag{6}$$

and, from (1), has the form

$$\mathbf{Z}(\omega, \tau) = \begin{bmatrix} \overline{\mathbf{A}}(\omega, \tau) \\ \overline{\mathbf{A}}(\omega, \tau)\mathbf{\Phi}(\omega, \tau) \end{bmatrix} \begin{bmatrix} S_1(\omega, \tau) \\ \vdots \\ S_N(\omega, \tau) \end{bmatrix} + \begin{bmatrix} V_1(\omega, \tau) \\ \vdots \\ V_{M-1}(\omega, \tau) \\ V_2(\omega, \tau) \\ \vdots \\ V_M(\omega, \tau) \end{bmatrix}$$

where $\overline{\mathbf{A}}(\omega, \tau)$ contains the top $M-1$ rows of $\mathbf{A}(\omega, \tau)$ and $\mathbf{\Phi}(\omega, \tau)$ is a diagonal matrix with entries corresponding to the mixing parameters $\phi_1(\omega, \tau), \phi_2(\omega, \tau), \ldots, \phi_N(\omega, \tau)$. It follows that the covariance matrix $\mathbf{R}_{\mathbf{ZZ}}(\omega, \tau)$ has the form

$$\begin{bmatrix} \overline{\mathbf{A}}(\omega, \tau) \\ \overline{\mathbf{A}}(\omega, \tau)\mathbf{\Phi}(\omega, \tau) \end{bmatrix} \mathbf{R}_{\mathbf{SS}}(\omega, \tau) \begin{bmatrix} \overline{\mathbf{A}}(\omega, \tau) \\ \overline{\mathbf{A}}(\omega, \tau)\mathbf{\Phi}(\omega, \tau) \end{bmatrix}^H + \mathbf{R}_{\mathbf{VV}}(\omega, \tau) \tag{7}$$

and its SVD is of the form

$$\begin{bmatrix} \mathbf{E}_1(\omega, \tau) & \mathbf{E}_{V_1}(\omega, \tau) \\ \mathbf{E}_2(\omega, \tau) & \mathbf{E}_{V_2}(\omega, \tau) \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda}(\omega, \tau) & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}(\omega, \tau) \end{bmatrix} \begin{bmatrix} \mathbf{E}_1(\omega, \tau) & \mathbf{E}_{V_1}(\omega, \tau) \\ \mathbf{E}_2(\omega, \tau) & \mathbf{E}_{V_2}(\omega, \tau) \end{bmatrix}^H \tag{8}$$

where the number of singular values contained in $\mathbf{\Lambda}(\omega, \tau)$ is equal to the rank of $\mathbf{R}_{\mathbf{SS}}(\omega, \tau)$ and the columns of $\mathbf{E}_1(\omega, \tau)$ and $\mathbf{E}_2(\omega, \tau)$ correspond to the singular values contained in $\mathbf{\Lambda}(\omega, \tau)$. In the no noise case, there exists a non-singular matrix $\mathbf{T}$, such that

$$\begin{bmatrix} \overline{\mathbf{A}}(\omega, \tau) \\ \overline{\mathbf{A}}(\omega, \tau)\mathbf{\Phi}(\omega, \tau) \end{bmatrix} \mathbf{T} = \begin{bmatrix} \mathbf{E}_1(\omega, \tau) \\ \mathbf{E}_2(\omega, \tau) \end{bmatrix}. \tag{9}$$

For high signal to noise ratios, (9) holds approximately and so

$$\mathbf{\Phi}(\omega, \tau) \approx \mathbf{T} \left[\mathbf{E}_1(\omega, \tau)\right]^{\dagger} \left[\mathbf{E}_2(\omega, \tau)\right] \mathbf{T}^{-1}. \tag{10}$$

Thus, the $M-1$ eigenvalues from the eigenvalue decomposition of $\left[\mathbf{E}_1(\omega, \tau)\right]^{\dagger} \left[\mathbf{E}_2(\omega, \tau)\right]$ are estimates of the $M-1$ mixing parameters, $\phi_1(\omega, \tau), \phi_2(\omega, \tau), \ldots, \phi_N(\omega, \tau)$.

In order to demix, at each time-frequency point a demixing matrix

$$\tilde{\mathbf{B}}(\omega, \tau) = \begin{bmatrix} 1 & \cdots & 1 \\ \tilde{\phi}_1(\omega, \tau) & \cdots & \tilde{\phi}_{M-1}(\omega, \tau) \\ \vdots & & \vdots \\ \tilde{\phi}_1^{M-1}(\omega, \tau) & \cdots & \tilde{\phi}_{M-1}^{M-1}(\omega, \tau) \end{bmatrix}^{\dagger} \tag{11}$$

is constructed and $\tilde{\mathbf{B}}(\omega, \tau) \left[X_1(\omega, \tau) \ldots X_M(\omega, \tau)\right]^{\mathrm{T}}$ yields $M-1$ instantaneous source estimates. However these estimates will be permuted randomly at each time-frequency point and some of them will not correspond to any source if less than $M-1$ sources are active at the given time-frequency point. To overcome these problems, we use a two-dimensional power weighted histogram of the parameter estimates

$$\tilde{\alpha}_i(\omega, \tau) = |\tilde{\phi}_i(\omega, \tau)| \text{ and } \tilde{\delta}_i(\omega, \tau) = \frac{1}{\omega} \angle \tilde{\phi}_i(\omega, \tau) \tag{12}$$

obtained each time-frequency point for $i = 1, 2, \ldots, M-1$. The histogram groups all of the mixing parameter estimates into one structure and provides the means for assigning source labels to the instantaneous demixtures. As in DUET, the histogram will contain N peaks indicating N sources and the peak centers are used as estimates of the associated source mixing parameters $\phi_1(\omega,\tau), \ldots, \phi_N(\omega,\tau)$. Since each of these instantaneous source estimates is associated with a mixing parameter estimate, the peak center estimates may be used as labels for the N sources and so, assigning a label to each of the instantaneous source estimates, the N source estimates are created. Synthesis back into the time domain follows as the final step.

3 Separation in echoic environments

3.1 Demixing N coherent sources

Assuming an echoic mixing model (3) it is still possible to demix N coherent (or fully correlated) source signals using a demixing matrix of the form

$$\tilde{\mathbf{B}}(\omega) = \left[\tilde{\mathbf{B}}_1(\omega)\, \tilde{\mathbf{B}}_2(\omega) \cdots \tilde{\mathbf{B}}_N(\omega)\right]^{\dagger} \tag{13}$$

$$\tilde{\mathbf{B}}_n(\omega) = \begin{bmatrix} 1 & \cdots & 1 \\ \tilde{\phi}_{n,1}(\omega) & \cdots & \tilde{\phi}_{n,P_n}(\omega) \\ \vdots & & \vdots \\ \tilde{\phi}_{n,1}^{M-1}(\omega) & \cdots & \tilde{\phi}_{n,P_n}^{M-1}(\omega) \end{bmatrix} .$$

The problem with this approach lies not with the actual demixing but with the estimation of the mixing parameters $\phi_{1,1}(\omega), \ldots, \phi_{1,P_1}(\omega), \ldots, \phi_{1,P_N}(\omega), \ldots, \phi_{N,P_N}(\omega)$. When the data matrix (6) is used, the DESPRIT algorithm fails to estimate echoic mixing parameters (or equivalently the mixing parameters of coherent sources) due to the rank deficiency of $\mathbf{R_{ZZ}}(\omega,\tau)$, which in turn is due to the rank deficiency of $\mathbf{R_{SS}}(\omega,\tau)$.

3.2 Mixing parameter estimation for N coherent sources

We overcome this problem by adding structure to the columns as well as the rows of the data matrix in the following way

$$\mathbf{Z}(\omega,\tau) = \begin{bmatrix} X_1(\omega,\tau) & \cdots & X_{M/2}(\omega,\tau) \\ \vdots & & \vdots \\ X_{M/2}(\omega,\tau) & \cdots & X_{M-1}(\omega,\tau) \\ X_2(\omega,\tau) & \cdots & X_{M/2+1}(\omega,\tau) \\ \vdots & & \vdots \\ X_{M/2+1}(\omega,\tau) & \cdots & X_M(\omega,\tau) \end{bmatrix} . \tag{14}$$

In the no noise case, the spatial covariance matrix $\mathbf{R_{ZZ}}(\omega,\tau)$ now has the form

$$\begin{bmatrix} \overline{\mathbf{A}}(\omega,\tau) \\ \overline{\mathbf{A}}(\omega,\tau)\mathbf{\Phi}(\omega,\tau) \end{bmatrix} \mathbf{\Psi_S}(\omega,\tau) \mathbf{\Psi}_S^H(\omega,\tau) \begin{bmatrix} \overline{\mathbf{A}}(\omega,\tau) \\ \overline{\mathbf{A}}(\omega,\tau)\mathbf{\Phi}(\omega,\tau) \end{bmatrix}^H \tag{15}$$

where $\overline{\mathbf{A}}(\omega,\tau)$ now contains the top $M/2$ rows of $\mathbf{A}(\omega,\tau)$, $\mathbf{\Phi}(\omega,\tau)$ is a diagonal matrix with entries corresponding to the mixing parameters $\phi_1(\omega,\tau), \phi_2(\omega,\tau), \ldots, \phi_N(\omega,\tau)$ and

$$\mathbf{\Psi}_S(\omega,\tau) = \begin{bmatrix} S_1(\omega,\tau) & \cdots & \phi_1^{M/2-1}(\omega,\tau) S_{M/2}(\omega,\tau) \\ S_1(\omega,\tau) & \cdots & \phi_2^{M/2-1}(\omega,\tau) S_{M/2}(\omega,\tau) \\ \vdots & & \vdots \\ S_1(\omega,\tau) & \cdots & \phi_{M/2}^{M/2-1}(\omega,\tau) S_{M/2}(\omega,\tau) \end{bmatrix} . \tag{16}$$

Now even for the case of N completely coherent sources, $\mathbf{\Psi}_S(\omega,\tau)$ will always be of rank $M/2$, provided that the original mixing matrix $\mathbf{A}(\omega,\tau)$ is of full rank. It follows that $\mathbf{R_{ZZ}}(\omega,\tau)$ will be of rank $M/2$ and for any N coherent sources $M/2$ estimates of the mixing parameters $\phi_1(\omega,\tau), \ldots, \phi_N(\omega,\tau)$ may be estimated at each time-frequency point when $M \geq 2N$. Equivalently $M/2$ estimates of the echoic mixing parameters $\phi_{1,1}(\omega,\tau), \ldots, \phi_{1,P_1}(\omega,\tau), \ldots, \phi_{N,1}(\omega,\tau), \ldots, \phi_{N,P_N}(\omega,\tau)$ may be estimated at each time-frequency point when

$$M \geq 2 \max\{P_1, P_2, \ldots, P_N\} .$$

This technique was inspired by the parameter estimation step used by Unitary ESPRIT [5], which may be used to estimate the angles of arrival of $N = 2$ coherent narrowband sources.

4 The echoic DESPRIT algorithm

Step 1

A uniform linear array receives M possibly echoic mixtures $x_1(t), \ldots, x_M(t)$ of N speech signals. These M signals are transformed into the time-frequency domain using a discrete windowed Fourier transform

$$X_m(\omega,\tau) = \sum_{k=0}^{K-1} W(kT-\tau) x_m(kT) e^{-j\omega kT} ,$$

where $m = 1, \ldots, M$, $W(t)$ is a window function and T is the sampling period.

Step 2

At each time-frequency point the data matrix $\mathbf{Z}(\omega,\tau)$ is constructed according to (14) and the covariance matrix $\mathbf{R_{ZZ}}(\omega,\tau)$ is approximated as

$$\tilde{\mathbf{R}}_{\mathbf{ZZ}}(\omega,\tau) = [\mathbf{Z}(\omega,\tau)]\,[\mathbf{Z}(\omega,\tau)]^H .$$

The SVD is computed to give

$$\tilde{\mathbf{R}}_{\mathbf{ZZ}}(\omega,\tau) = \mathbf{U}(\omega,\tau)\mathbf{D}(\omega,\tau)\mathbf{U}^H(\omega,\tau)$$

and

$$\begin{bmatrix}\tilde{\mathbf{E}}_1(\omega,\tau)\\ \tilde{\mathbf{E}}_2(\omega,\tau)\end{bmatrix} = \text{first } M/2 \text{ columns}\left\{\mathbf{U}(\omega,\tau)\right\} ,$$

the $M/2$ eigenvalues of $\left[\tilde{\mathbf{E}}_1(\omega,\tau)\right]^{\dagger}\left[\tilde{\mathbf{E}}_2(\omega,\tau)\right]$ are then used as $M/2$ mixing parameter estimates $\tilde{\phi}_1(\omega,\tau),\ldots,\tilde{\phi}_{M/2}(\omega,\tau)$. The mixing parameter estimates are used compute $M/2$ corresponding instantaneous source estimates via

$$\begin{bmatrix}\tilde{S}_1(\omega,\tau)\\ \vdots \\ \tilde{S}_{M/2}(\omega,\tau)\end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ \tilde{\phi}_1(\omega,\tau) & \cdots & \tilde{\phi}_{M/2}(\omega,\tau) \\ \vdots & & \vdots \\ \tilde{\phi}_1^{M-1}(\omega,\tau) & \cdots & \tilde{\phi}_{M/2}^{M-1}(\omega,\tau)\end{bmatrix}^{\dagger}\begin{bmatrix}X_1(\omega,\tau)\\ \vdots \\ X_M(\omega,\tau)\end{bmatrix}.$$

Step 3

The mixing parameter estimates obtained at each time-frequency point are used to construct a two-dimensional histogram according (12), each entry of the histogram has an associated instantaneous source estimate, the squared magnitude of which is used as weighting factor for this entry.

Step 4

The $N' \geq N$ histogram peaks indicate N source signals propagating upon N' paths have been received by the uniform linear array. The N' peak centers are used as estimates of the echoic mixing parameters and act as labels for N' demixed source estimates, according to these labels each of the instantaneous source estimates are assigned to one of the N' demixed source estimates. Beginning with the instantaneous mixing parameter estimates associated with the instantaneous source estimates of lowest power, at each time-frequency point the closest peak center is found and the lowest power instantaneous source estimate is assigned to the appropriate demixed source estimate. The assignment is then carried out for the instantaneous mixing parameter estimates associated with the instantaneous source estimates of next lowest power and so on. Assignments carried out in later stages are allowed to overwrite previous assignments in the belief that the instantaneous mixing parameter estimates associated with the instantaneous signal estimates of greater power are the more reliable, since they have been affected by noise the least. The $N' \geq N$ demixed source estimates are then synthesised back into the time-domain.

5 Experiments

The echoic DESPRIT algorithm is used to demix two 2.4 second long speech signals arriving at a uniform linear array of six sensors upon two and three paths respectively. The power weighted histogram shows five peaks corresponding to the five arrivals at the array, each of the five arrivals is recovered as a scaled and delayed version of one of the original sources.

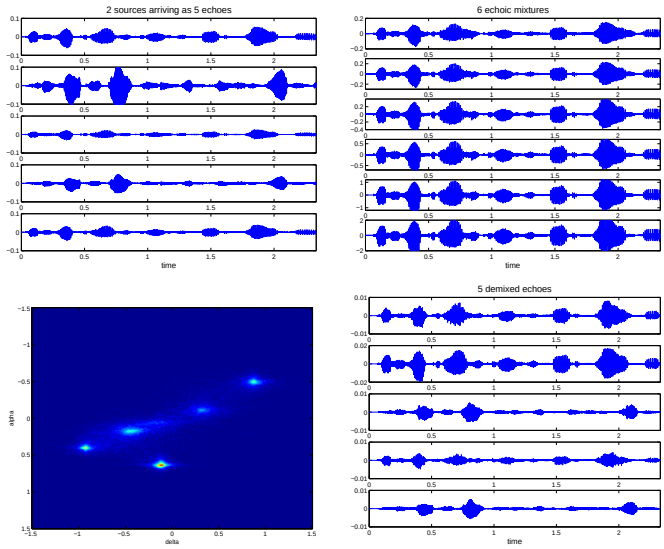

Fig. 2: Echoic DESPRIT is used to demix 6 mixtures of 2 signals traveling upon 2 and 3 paths respectively. Each of the five peaks in the histogram is associated with one signal path. The peak locations correspond to the mixing parameters.

References

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, July 2004.

[2] R. Roy and T. Kailath, "ESPRIT - Estimation of Signal Parameters via Rotational Invariance Techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 7, pp. 984–995, July 1989.

[3] T. Melia, S. Rickard, and C. Fearon, "Histogram-based blind source separation of more sources than sensors using a DUET-ESPRIT technique," in *Proc. European Signal Processing Conference (EUSIPCO'05)*, Antalya, Turkey, September 4–8 2005.

[4] S. Rickard, T. Melia, and C. Fearon, "DESPRIT - Histogram based blind source separation of more sources than sensors using subspace methods," in *Proc. IEEE Workshop on Applications of Signal Processing in Audio and Acoustics*, New Paltz, New York, October 16–19 2005.

[5] M. Haardt and J. A. Nossek, "Unitary ESPRIT: How to obtain increased estimation accuracy with a reduced computational burden," *IEEE Transactions on Signal Processing*, vol. 43, pp. 1232–1242, May 1995.