HISTOGRAM-BASED BLIND SOURCE SEPARATION OF MORE SOURCES THAN SENSORS USING A DUET-ESPRIT TECHNIQUE

Thomas Melia, Scott Rickard and Conor Fearon

Digital Signal Processing Group, University College Dublin, Ireland

ABSTRACT

The Direction of Arrival estimation algorithm ESPRIT is capable of estimating the angles of arrival of N narrowband source signals using $M > N$ anechoic sensor mixtures from a uniform linear array (ULA). Using a similar parameter estimation step, the DUET Blind Source Separation algorithm can demix $N > 2$ speech signals using $M = 2$ anechoic mixtures of the signals. We introduce here the DUET-ESPRIT (DESPRIT) Blind Source Separation algorithm which demixes $N > M$ speech signals using $M \geq 2$ anechoic mixtures.

1. INTRODUCTION

The "cocktail party phenomenon" illustrates the ability of the human auditory system to separate out a single speech source from the cacophony of a crowded room, using only two sensors and with no prior knowledge of the speakers or the channel presented by the room. Efforts to implement a receiver which emulates this sophistication are referred to as Blind Source Separation techniques [1; 2; 3].

Generally, in the anechoic blind source separation model, N time-varying source signals $s_1(t), s_2(t), \ldots, s_N(t)$ propagate across an isotropic, anechoic (direct path), non-dispersive medium and impinge upon an array of M sensors which are situated in the far-field of all sources. Under such conditions the m^{th} mixture can be expressed as

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t - t_{mn}) + n_m(t)$$

and an expression for a vector of M anechoic mixtures is given as

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_M(t) \end{bmatrix} = \begin{bmatrix} a_{11}\delta(t-t_{11}) & \cdots & a_{1N}\delta(t-t_{1N}) \\ a_{21}\delta(t-t_{21}) & \cdots & a_{2N}\delta(t-t_{2N}) \\ \vdots & & \vdots \\ a_{M1}\delta(t-t_{M1}) & \cdots & a_{MN}\delta(t-t_{MN}) \end{bmatrix} \star \begin{bmatrix} s_1(t) \\ \vdots \\ s_N(t) \end{bmatrix} + \begin{bmatrix} n_1(t) \\ n_2(t) \\ \vdots \\ n_M(t) \end{bmatrix}$$

where $n_1(t), n_2(t), \ldots, n_M(t)$ are i.i.d. (independently and identically distributed) and $\star$ denotes convolution.

Generally blind source separation algorithms attempt to retrieve or estimate the source signals $\mathbf{s}(t)$ from the received mixtures $\mathbf{x}(t)$ with little, if any prior information about the mixing matrix or the source signals themselves. Typically blind source separation and direction of arrival techniques such as ESPRIT require the number of sensors to be greater than or equal to the number of sources ($M \geq N$). The DUET blind source separation algorithm [4; 5] can demix $N > 2$ signals using only 2 anechoic mixtures of the signals, providing these signals are W-disjoint orthogonal (WDO). The DUET-ESPRIT (DESPRIT) algorithm presented in this paper stems from an implementation of ESPRIT under a WDO assumption. DUET requires $M = 2$, whereas DESPRIT can be seen as one possible extension of DUET when $M > 2$ mixtures are available. DESPRIT makes similar assumptions to ESPRIT as regards the layout of the sensors, namely that the sensors can be divided into two paired subarrays with each paired couplet of sensors sharing a common displacement vector.

The paper is structured as follows, Section 2 outlines the classic ESPRIT parameter estimation algorithm, Section 3 outlines the DUET blind source separation (BSS) algorithm and how the DUET-like ESPRIT BSS technique we dub DESPRIT emerges, Section 4 gives a summary of the DESPRIT algorithm implemented as a multichannel DUET extension and Section 5 presents simulation results for this new BSS algorithm.

2. DIRECTION OF ARRIVAL ESTIMATION AND SUBSPACE METHODOLOGY

2.1 Narrowband Array Processing

Classic Direction of Arrival estimation techniques such as MUSIC [6] and ESPRIT [7] aim to find the N angles of arrival for N narrowband signals $s_1(t), s_2(t), \ldots, s_N(t)$ impinging upon an array of M sensors. With accurate estimation beamforming can be performed to separate the N signals.

For narrowband signals of centre frequency ω_0 a time lag can be approximated by a phase rotation, i.e. $s(t-\tau) \approx s(t)e^{-j\omega_0\tau}$, where $s(t)$ is the complex representation of a real signal. As a result the m^{th} mixture can be expressed as

$$x_m(t) = \sum_{n=1}^{N} a_{mn} e^{-j\omega_0 t_{mn}} s_n(t) + n_m(t)$$

and by letting $a_{mn}e^{-j\omega_0 t_{mn}} \rightarrow a_{mn}$ allows M sensor mixtures to be written as

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_M(t) \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ a_{21} & \cdots & a_{2N} \\ \vdots & & \vdots \\ a_{M1} & \cdots & a_{MN} \end{bmatrix} \begin{bmatrix} s_1(t) \\ \vdots \\ s_N(t) \end{bmatrix} + \begin{bmatrix} n_1(t) \\ n_2(t) \\ \vdots \\ n_M(t) \end{bmatrix}$$

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}_{\mathbf{x}}(t) \tag{1}$$

where the mixing matrix $\mathbf{A}$ has complex entries which do not vary significantly with time and each column may be associated with an individual narrowband source signal.

An estimate of the spatial covariance matrix

$$R_{\mathbf{xx}} = E\left\{ [\mathbf{x}(t)] [\mathbf{x}(t)]^H \right\} \tag{2}$$

can be calculated, where $[\,]^H$ denotes a complex conjugate transpose operation. The Singular Value Decomposition (SVD) of $R_{\mathbf{xx}}$ is of the form

$$R_{\mathbf{xx}} = [\ \mathbf{E_s} \mid \mathbf{E_n}\ ] \begin{bmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma} \end{bmatrix} [\ \mathbf{E_s} \mid \mathbf{E_n}\ ]^H, \tag{3}$$

where $\mathbf{\Lambda}$ is a diagonal matrix with the N dominant entries associated with N signals, the $M-N$ remaining singular values are comparable to the noise variance and are contained in the diagonal matrix $\mathbf{\Sigma}$, the N column vectors of $\mathbf{E_s}$ are associated with the N dominant singular values and the $M-N$ column vectors of $\mathbf{E_n}$ are associated with the $M-N$ remaining singular values. The subspace spanned by $\mathbf{E_s}$ is known as the signal subspace and the orthogonal subspace spanned by $\mathbf{E_n}$ is known as the noise subspace.

2.2 ESPRIT

The ESPRIT algorithm relies on two subarrays of sensors. Each element of the first subarray is displaced in space from the corresponding element of the second subarray by the same displacement vector. As a result, the time lag between each sensor pair for a signal travelling between the two subarrays is constant. Without loss of generality we assume that the original sensor array is a uniformly spaced linear array consisting of M sensors, as a result we may subdivide the array of M sensors into two such subarrays of $M-1$ sensors each. The first subarray contains sensors $1,\ldots,M-1$ and the second subarray contains sensors $2,\ldots,M$.

The $M-1$ mixtures from the second array can be represented as

$$\mathbf{y}(t) = \mathbf{A}\,\mathbf{\Phi}\,\mathbf{s}(t) + \mathbf{n_y}(t)$$

where the diagonal matrix

$$\mathbf{\Phi} = \begin{bmatrix} e^{-j\omega_0\delta_1} & & \\ & \ddots & \\ & & e^{-j\omega_0\delta_N} \end{bmatrix},$$

contains delay terms $\delta_n, n = 1,\ldots,N$, which are unique to each source signal and are related geometrically to the angle of arrival, i.e. $\delta_n = \Delta\cos(\theta_n)/c$ where Δ is the distance between the two subarrays, θ_n is the angle of arrival of the n^{th} signal onto the array and c is the propagation speed.

Both data vectors can be stacked to form

$$[\mathbf{z}(t)] = \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A} \\ \mathbf{A\Phi} \end{bmatrix} [\ \mathbf{s}(t)\ ] + \begin{bmatrix} \mathbf{n_x}(t) \\ \mathbf{n_y}(t) \end{bmatrix} \quad (4)$$

which is a $2(M-1)\times 1$ vector of mixtures. It follows that the SVD of the spatial covariance matrix $\mathbf{R_{zz}}$ can be computed

$$R_{\mathbf{zz}} = \left[\begin{array}{c|c} \mathbf{E_x} & \mathbf{E_{n_x}} \\ \mathbf{E_y} & \mathbf{E_{n_y}} \end{array}\right] \begin{bmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma} \end{bmatrix} \left[\begin{array}{c|c} \mathbf{E_x} & \mathbf{E_{n_x}} \\ \mathbf{E_y} & \mathbf{E_{n_y}} \end{array}\right]^H.$$

For the no-noise case, the mixing matrix spans the same space as the signal subspace, i.e. there exists a non-singular matrix $\mathbf{T}$ such that

$$\begin{bmatrix} \mathbf{E_x} \\ \mathbf{E_y} \end{bmatrix} = \begin{bmatrix} \mathbf{A} \\ \mathbf{A\Phi} \end{bmatrix} \mathbf{T} \quad (5)$$

furthermore the diagonal matrix Φ is related to $[\mathbf{E_x}^{\dagger}\mathbf{E_y}]$ via a similarity transform

$$\mathbf{\Phi} = \mathbf{T}\left[\mathbf{E_x}^{\dagger}\mathbf{E_y}\right]\mathbf{T}^{-1}. \quad (6)$$

where $(\,)^{\dagger}$ denotes the Moore-Penrose pseudo-inverse. As a result the N angles of arrival (θ_n, $n = 1,\ldots,N$) can be recovered from the N complex eigenvalues of $[\mathbf{E_x}^{\dagger}\mathbf{E_y}]$, which are of the form

$$e^{-j\omega_0(\Delta\cos(\theta_n)/c)}\ n = 1,\ldots,N.$$

The original ESPRIT algorithm is a time-domain based technique, where $\mathbf{R_{zz}}$ is approximated by a time average

$$\mathbf{R_{zz}} \approx \frac{1}{T}\int_{-T/2}^{T/2} [\mathbf{z}(t)]\,[\mathbf{z}(t)]^H\,dt\,.$$

A frequency domain based approach is also possible with the ESPRIT algorithm being performed at each point in the frequency domain using the covariance matrix

$$\mathbf{R_{ZZ}}(\omega) = [\mathbf{Z}(\omega)]\,[\mathbf{Z}(\omega)]^H\,,$$

where $\mathbf{Z}(\omega)$ is Fourier Transform of $\mathbf{z}(t)$. Such a frequency domain approach has the advantage that the narrowband assumption placed upon the source signals is no longer necessary. However, at each frequency the N signal subspace vectors are permutated and so, without knowledge of this random permutation, combining results across frequencies becomes difficult [8].

3. BLIND SOURCE SEPARATION OF W-DISJOINT ORTHOGONAL SIGNALS

3.1 DUET

DUET handles this permutation problem by mapping each delay estimate to a source using a weighted histogram. DUET makes a further simplifying assumption which ESPRIT does not require. The DUET method relies on the concept of *approximate W-disjoint orthogonality* (WDO), a measure of sparsity which quantifies the non-overlapping nature of the time-frequency representations of the sources. This property is exploited to facilitate the separation of any number of sources blindly from just two mixtures using the spatial signatures of each source. These spatial signatures arise out of the separation of the measuring sensors which produces a relative arrival delay, δ_i, and a relative attenuation factor, α_i, for the i^{th} source.

Using a windowed Fourier transform

$$S_i^W(\omega,\tau) = \int_{-\infty}^{\infty} W(t-\tau)s_i(t)e^{-j\omega t}dt$$

the WDO assumption can be written as

$$S_i^W(\omega,\tau)S_k^W(\omega,\tau) = 0, \forall\,\omega,\tau, i \neq k\,. \quad (7)$$

Assuming all sources are W-disjoint orthogonal, at a given time-frequency point only one of the N sources will have a non-zero value. This allows DUET to perform separation using only two mixtures. Thus the equations for the mixtures can be written as follows:

$$\begin{bmatrix} X^W(\omega,\tau) \\ Y^W(\omega,\tau) \end{bmatrix} = \begin{bmatrix} 1 \\ \alpha_i e^{-j\omega\delta_i} \end{bmatrix} S_i^W(\tau,\omega) \quad (8)$$

where we have defined $s_i(t)$ to be the i^{th} source measured at $x(t)$. From this we can determine expressions for the mixing parameters at each point in the time-frequency domain of each of the mixtures $X^W(\omega,\tau)$ and $Y^W(\omega,\tau)$. Approximate W-disjoint orthogonality suggests that the parameters at each point are equal to or, at least, tend towards those for one source only.

$$(\hat{\alpha}_j, \hat{\delta}_j) = \left(\log\left|\frac{Y^W(\omega,\tau)}{X^W(\omega,\tau)}\right|, \mathrm{Im}\left(\log\left(\frac{Y^W(\omega,\tau)}{X^W(\omega,\tau)}\right)\right)/\omega\right) \quad (9)$$

Note that, due to approximate nature of W-disjoint orthogonality along with the presence of noise, the mixing parameters in (9) are only estimates of the true values. If we calculated these parameter estimates at every point in time-frequency space, we would expect the results to cluster around the true values of the actual mixing parameters. N sources produces N pairs of mixing parameters which creates N peaks in the parameter space histogram. We can then use these mixing parameter estimates to partition the time-frequency representation of one mixture to recover the source estimates. It may be noted that the phase is defined modulo-π in (9), with closely space sensors of maximum separation $\Delta_{\max} = \pi f_{\max}$ (where $f_{\max}$ is the highest frequency with non-negligible energy content) phase wrapping is not a problem.

3.2 DUET-ESPRIT (DESPRIT)

The ESPRIT algorithm can be performed at each point in the time-frequency domain using the localised spatial covariance matrix

$$\mathbf{R_{ZZ}}(\omega,\tau) = \mathrm{E}\left\{ \begin{bmatrix} \mathbf{X}^W(\omega,\tau) \\ \mathbf{Y}^W(\omega,\tau) \end{bmatrix} \begin{bmatrix} \mathbf{X}^W(\omega,\tau)^H & \mathbf{Y}^W(\omega,\tau)^H \end{bmatrix} \right\} \tag{10}$$

the singular value decomposition of $\mathbf{R_{ZZ}}(\omega,\tau)$ at each time-frequency point is of the form

$$\mathbf{R_{ZZ}}(\omega,\tau) = \left[\begin{array}{c|c} \mathbf{E_X} & \mathbf{E_{n_X}} \\ \mathbf{E_Y} & \mathbf{E_{n_Y}} \end{array}\right] \begin{bmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma} \end{bmatrix} \left[\begin{array}{c|c} \mathbf{E_X} & \mathbf{E_{n_X}} \\ \mathbf{E_Y} & \mathbf{E_{n_Y}} \end{array}\right]^H,$$

From equation (6) Φ may be recovered via an eigenvalue decomposition

$$\Phi(\omega,\tau) = \mathbf{T}\left[\mathbf{E_X}^{\dagger}(\omega,\tau)\mathbf{E_Y}(\omega,\tau)\right]\mathbf{T}^{-1}$$

at a given time-frequency point, up to N signals may be present and the resulting $N \times N$ diagonal matrix $\Phi(\omega,\tau)$ has up to N non-zero entries which are of the form

$$\phi_i = \alpha_i e^{-j\omega\delta_i},\ i = 1,\ldots,N$$

where α_i and δ_i are the attenuation and delay parameters for the i^{th} source. It is discussed in section 3.1 how the DUET BSS algorithm constructs a two dimensional histogram of these parameters to identify any number of sources and ultimately separate them if they can be assumed to strongly W-disjoint orthogonal. By borrowing from both techniques a hybrid DUET-ESPRIT (DESPRIT) blind source separation algorithm is possible.

This DESPRIT algorithm estimates the delay (equivalently the angle of arrival) and the attenuation of N WDO source signals as they pass across an ESPRIT-like array of sensor pairs using two or more anechoic mixtures. Providing each source has a unique attenuation and delay estimate, a two dimensional histogram will have N peaks corresponding to N source signals. The centre of each peak provides an accurate estimate of the actual attenuation and delay of each source. Since the attenuation and delay parameter estimation is performed at each time-frequency point, the estimates for the mixing parameters of the N sources can be used to partition the time-frequency plane into N regions where the WDO sources are active. As a result N time-frequency masks with non-zero values at active time-frequency points and zeros elsewhere can be applied to any of the mixtures to demix these N source signals.

Under a weakened WDO assumption with possibly $M-1$ sources overlapping in the time-frequency domain the parameter estimation step of DUET fails, however the DESPRIT algorithm continues to work well providing that the number of sensors in the ESPRIT-like uniform linear array outnumber the number of sources that may coexist at a particular region in the time-frequency domain. A treatment of the DESPRIT BSS technique operating under a weakened WDO assumption is contained within a future publication. The current paper examines DESPRIT under the DUET strong WDO assumption (at most one source is active for every time-frequency point).

3.3 DESPRIT as a multichannel DUET extension

Under a strong WDO assumption Λ is a 1×1 scalar λ, Σ has all near zero entries and $\begin{bmatrix} \mathbf{E_X}(\omega,\tau) \\ \mathbf{E_Y}(\omega,\tau) \end{bmatrix}$ is a $2m\times 1$ vector so as a result the scalar ϕ is given by

$$\phi = \mathbf{E_X}(\omega,\tau)^{\dagger}\mathbf{E_Y}(\omega,\tau). \tag{11}$$

Furthermore when the expectation operator of equation (10) is approximated by an instantaneous estimate, i.e.

$$\mathbf{R_{ZZ}}(\omega,\tau) = \begin{bmatrix} \mathbf{X}^W(\omega,\tau) \\ \mathbf{Y}^W(\omega,\tau) \end{bmatrix} \begin{bmatrix} \mathbf{X}^W(\omega,\tau)^H & \mathbf{Y}^W(\omega,\tau)^H \end{bmatrix}$$

the expression (11) is equivalent to

$$\phi = \mathbf{X}^W(\omega,\tau)^{\dagger}\mathbf{Y}^W(\omega,\tau) \tag{12}$$

and so in this case the subspace decomposition of the spatial covariance matrix is unnecessary. In the $M=2$ case this implementation of DESPRIT reduces to DUET and for $M>2$ it may be viewed simply as a multichannel DUET extension.

4. THE DESPRIT MULTICHANNEL DUET EXTENSION

Step 1

A uniformly spaced linear array of M sensors receives M anechoic mixtures $x_1(t), x_2(t), \ldots, x_M(t)$, of N WDO source signals. These M signals are represented in the $2(M-1)\times 1$ time-varying vector

$$\mathbf{z}(t) = \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \end{bmatrix}_{2(M-1)\times 1}$$

where $\mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_{M-1}(t))^T$ and $\mathbf{y}(t) = (x_2(t), x_2(t), \ldots, x_M(t))^T$ represent signals taken from the first and second subarrays respectively. K samples are taken at $t = kT$, $k = 0, 1, \ldots, K-1$, where T is the sampling period.

Step 2

A window $W(t)$, of length L is formed and by shifting the position of the window by multiples of Δ seconds, localisation in time is possible.

for $\tau = 0 : \Delta : (K-1)T$

$$\mathbf{z}(t,\tau) = W(t-\tau)\mathbf{z}(t)$$

$$\mathbf{Z}(\omega,\tau) = \mathrm{DFT}(\mathbf{z}(t,\tau))$$

for $\omega = (0:1:L-1)\times 2\pi/LT$

$$\phi(\omega,\tau) = \mathbf{X}(\omega,\tau)^{\dagger}\mathbf{Y}(\omega,\tau)$$

$$\delta(\omega,\tau) = -\mathrm{Im}\left\{\log_e\left\{\phi(\omega,\tau)\right\}\right\}/\omega$$

$$\alpha(\omega,\tau) = |\phi(\omega,\tau)|$$

end
end

Step 3

A two dimensional histogram of the attenuation and delay parameters (α and δ) is constructed, weighting of histogram values is possible using $\mathbf{X}(\omega,\tau)^H\mathbf{X}(\omega,\tau)$ which is proportional to the power of the source present at each time-frequency point. N histogram peaks indicate N source signals, the (α,δ) values corresponding to the centre of each peak are mapped back into the time-frequency domain to indicate in which regions each of the N source signals are active. Peak Detection is performed using a weighted K-means based technique.

Step 4

Under the assumption that the N source signals are strongly W-disjoint orthogonal, a binary time-frequency mask corresponding to the regions of the time-frequency plane where a source is active is created. Applying the n^{th} mask to any of the received mixtures recovers the n^{th} source signal. N such masks are used to separate the N sources.

5. SIMULATION RESULTS

DESPRIT was used to blindly demix four 2.4 seconds long speech signals, using three anechoic mixtures of these signals each having been sampled at 16kHz. Plots of the original source signals, the received mixtures, the two-dimensional histogram and the demixed signals are given in Figure 1, a high SNR of 100dB is assumed.

A straight-forward multiple sensor extension of DUET is Stacked-DUET where the DUET parameter estimation step is performed for each sensor pair in an ESPRIT-type sensor array, a weighted histogram made up of all the parameter estimates from each DUET implementation is used to estimate and demix the sources in a DUET-like fashion.

The DESPRIT multichannel extension of DUET can be compared with Stacked-DUET for the same data as before under noisier conditions. DESPRIT has clear advantages at lower values of Signal to Noise Ratios (SNRs) since an increase in the number of sensors improves parameter estimation when using DESPRIT, but Stacked-DUET fails to improve its parameter estimates when the number of sensors is increased. Figure 2 shows the parameter estimation histograms for DESPRIT and DUET (Stacked-DUET) for 3 sensors at SNR level 70 dB, 3 sensors at 40 dB and 7 sensors at 40 dB.

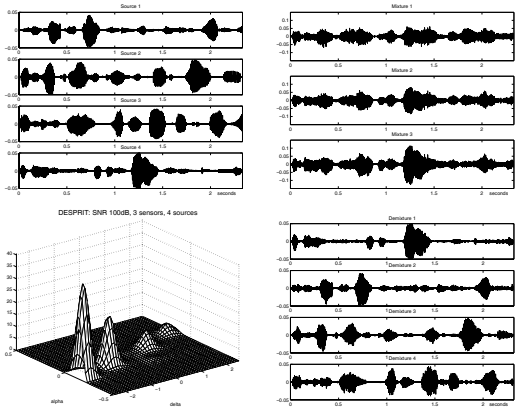

Figure 1: Blind Source Separation using DESPRIT of 4 speech signals from 3 anechoic mixtures.

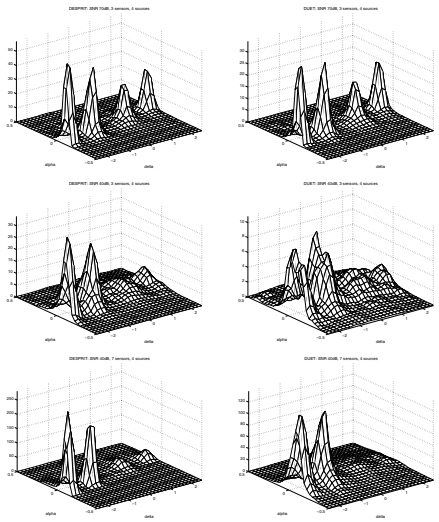

Figure 2: Parameter estimation with DESPRIT(left column) and DUET(right column) using 3 sensors with SNR 70dB, 3 sensors at 40dB and 7 sensors at 40dB.

References

[1] A. J. Bell and T. J. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. vol. 6, pp. 1129–1159, 1995.

[2] P. Comon, "Independent component analysis: A new concept?" *Signal Processing*, vol. vol. 36, no. 8, pp. 287–314, 1994.

[3] A. Hyvarinen, J. Karhunen, and E. Oja, "Independent component anaylsis," *Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications and Control*, 2001.

[4] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. vol. 5, pp. 2985–2988, June 2000.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. vol. 52, no. 7, pp. 1830–1847, July 2004.

[6] R. O. Schmidt, "Multiple emitter location and signal parameter estimation (MUSIC)," *IEEE Trans. on Antennas and Propagation*, vol. AP-34, no. 53, pp. 276–280, March 1986.

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

[8] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, september 2004.