Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 86484, 19 pages
doi:10.1155/2007/86484

Research Article

Underdetermined Blind Source Separation in Echoic Environments Using DESPRIT

Thomas Melia and Scott Rickard

Sparse Signal Processing Group, University College Dublin, Belfield, Dublin 4, Ireland

Received 1 October 2005; Revised 4 April 2006; Accepted 27 May 2006

Recommended by Andrzej Cichocki

The DUET blind source separation algorithm can demix an arbitrary number of speech signals using $M = 2$ anechoic mixtures of the signals. DUET however is limited in that it relies upon source signals which are mixed in an anechoic environment and which are sufficiently sparse such that it is assumed that only one source is active at a given time frequency point. The DUET-ESPRIT (DESPRIT) blind source separation algorithm extends DUET to situations where $M \geq 2$ sparsely echoic mixtures of an arbitrary number of sources overlap in time frequency. This paper outlines the development of the DESPRIT method and demonstrates its properties through various experiments conducted on synthetic and real world 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 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 [1–3]. The DUET blind source separation method [4] can demix an arbitrary number of speech source signals given just 2 anechoic mixtures of the sources, providing that the time-frequency representations of the sources do not overlap. The technique is limited in the following respects.

(1) It is not obvious how to best extend the technique to a situation where more mixtures are available.
(2) The assumption that only one source is active at a given time-frequency point is limiting, especially when $M > 2$ mixtures may be available.
(3) The anechoic mixing model clearly restricts the types of environments where DUET can be applied.

A number of extensions to the DUET blind source separation method have recently been proposed [5–7] that address these issues. In this paper we summarise and characterise the performance of these extensions, which we believe embody the natural multichannel, echoic extension of DUET. Other authors have proposed different DUET extensions, for example, [8–11] describe multichannel extensions to DUET when $M \geq 2$ mixtures are available. It is recognised in [9–15] that the assumption that only one source is active at a given time-frequency point is quite a harsh restriction to place upon large numbers of speech sources and weakened forms of this assumption are presented in these papers. An echoic extension to DUET is demonstrated in [9] when the mixing parameters are known a priori. In this work, we extend DUET to use $M > 2$ mixtures and in doing so are able to separate multiple sources at each time-frequency point, even when mixing is echoic.

In general, we seek to demix M mixtures of N source signals taken from a uniform linear array of sensors. In the frequency domain we model the M mixtures $X_1(\omega), \dots, X_M(\omega)$ of N source signals $S_1(\omega), \dots, S_N(\omega)$ as

$$\begin{bmatrix} X_1(\omega) \\ X_2(\omega) \\ \vdots \\ X_M(\omega) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ \phi_1(\omega) & \cdots & \phi_N(\omega) \\ \vdots & & \vdots \\ \phi_1^{M-1}(\omega) & \cdots & \phi_N^{M-1}(\omega) \end{bmatrix} \begin{bmatrix} A_1(\omega)S_1(\omega) \\ \vdots \\ A_N(\omega)S_N(\omega) \end{bmatrix} + \begin{bmatrix} V_1(\omega) \\ V_2(\omega) \\ \vdots \\ V_M(\omega) \end{bmatrix}, \tag{1}$$

where $A_n(\omega) = a_n e^{-j\omega d_n}$, a_n and d_n are the attenuation and delay experienced by the nth signal as it propagates to the 1st sensor, $\phi_n(\omega) = \alpha_n e^{-j\omega\delta_n}$, α_n and δ_n are the attenuation and delay experienced by the nth signal as it travels between two adjacent sensors, and $V_1(\omega), V_2(\omega), \ldots, V_M(\omega)$ are independently and identically distributed noise terms. Equivalently in the time domain the mth anechoic mixture $x_m(t)$ of the N source signals, $s_1(t), s_2(t), \ldots, s_N(t)$, can be expressed as

$$x_m(t) = \sum_{n=1}^{N} a_n \alpha_n^{m-1} s_n\left(t - d_n - (m-1)\delta_n\right) + v_m(t), \qquad (2)$$

where the inverse Fourier transform is defined as $f(t) = (1/2\pi)\int_{-\infty}^{\infty} F(\omega)e^{j\omega t}d\omega$. The anechoic mixing model (1) may be altered to become an echoic mixing model by adding columns to the mixing matrix corresponding to echoic paths:

$$\begin{bmatrix} X_1(\omega)\\ X_2(\omega)\\ \vdots\\ X_M(\omega)\end{bmatrix} = \mathcal{A}(\omega)\begin{bmatrix} A_{1,1}(\omega)S_1(\omega)\\ \vdots\\ A_{1,P_1}(\omega)S_1(\omega)\\ \vdots\\ A_{N,1}(\omega)S_N(\omega)\\ \vdots\\ A_{N,P_N}(\omega)S_N(\omega)\end{bmatrix} + \begin{bmatrix} V_1(\omega)\\ V_2(\omega)\\ \vdots\\ V_M(\omega)\end{bmatrix}, \qquad (3)$$

$$\mathcal{A}(\omega) = \begin{bmatrix} 1 & \cdots & 1 & & 1 & \cdots & 1\\ \phi_{1,1}(\omega) & \cdots & \phi_{1,P_1}(\omega) & & \phi_{N,1}(\omega) & \cdots & \phi_{N,P_N}(\omega)\\ \vdots & & \vdots & \cdots & \vdots & & \vdots\\ \phi_{1,1}^{M-1}(\omega) & \cdots & \phi_{1,P_1}^{M-1}(\omega) & & \phi_{N,1}^{M-1}(\omega) & \cdots & \phi_{N,P_N}^{M-1}(\omega) \end{bmatrix}, \qquad (4)$$

where $A_{n,p}(\omega) = a_{n,p}e^{-j\omega d_{n,p}}$, $a_{n,p}$ and $d_{n,p}$ are the attenuation and delay experienced by the nth signal as it propagates along its pth path, to the 1st sensor, $\phi_{n,p}(\omega) = \alpha_{n,p}e^{-j\omega\delta_{n,p}}$, $\alpha_{n,p}$ and $\delta_{n,p}$ are the attenuation and delay experienced by the nth signal as it propagates between two adjacent sensors along its pth path and P_n is the number of paths the nth source signal travels upon to reach the sensor array. Equivalently in the time domain the mth echoic mixture can be expressed as

$$x_m(t) = \sum_{n=1}^{N}\sum_{p=1}^{P_n} a_{n,p}\alpha_{n,p}^{m-1} s_n\left(t - d_{n,p} - (m-1)\delta_{n,p}\right) + v_m(t). \qquad (5)$$

This model has the same form as (1) but now there are $N' \geq N$ signals being received by the sensor array, some of these signals will be originated from the same source. Figure 1 illustrates a simple anechoic mixing procedure and a related echoic mixing procedure. Our treatment assumes a uniform linear array with spacing $\leq c/2f_{\max}$ throughout, where $f_{\max}$ is the maximum frequency of interest and c is the speed at which the signals propagate. Furthermore it is assumed that the sensor array is located sufficiently far away from the source locations that planar wave propagation occurs, although not previously stated, this assumption is implicit in the mixing models (1) and (3).

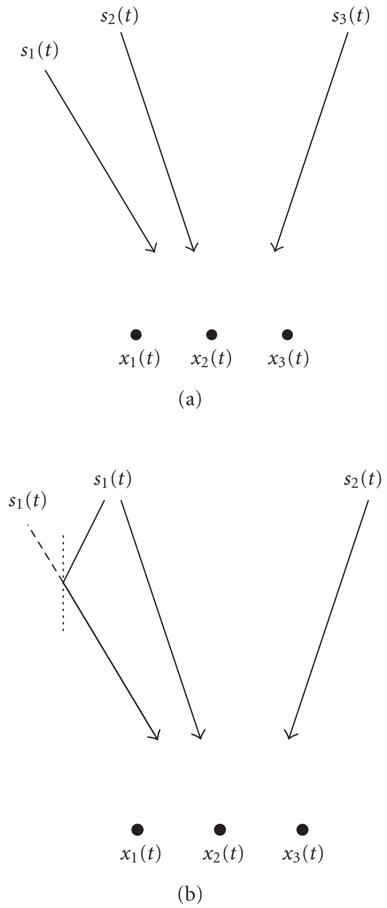


FIGURE 1: 3 sensors pick up 3 anechoic mixtures of 3 signals (a) and 3 echoic mixtures of 2 signals (b).

The goal of a blind source separation method is to estimate the source signals $s_1(t), s_2(t), \ldots, s_N(t)$ from the mixture signals $x_1(t), x_2(t), \ldots, x_M(t)$. This paper describes a time-frequency domain approach to this problem. Such transform domain approaches are a popular way of extending independent component analysis type algorithms to the convolved mixture problem [16–18] but they must overcome the well-known permutation ambiguity [19]. DUET (which we extend in this paper to a sparse convolutive model) overcomes the permutation problem by parameterising the mixing model. In the 2-channel case ($M = 2$) with anechoic mixing ($P_n = 1$), the DUET algorithm can perform blind source separation even when $N > 2$ sources are present and it is unaffected by the permutation ambiguity. DUET relies on the sparsity of speech in the time-frequency domain, a key assumption in many papers [8–15, 20, 21]. Sparsity is defined in various ways in the literature. We take sparsity to mean that a small percentage of the time-frequency points contain a large percentage of the signal power. Moreover the significant power containing coefficients for two different speech signals rarely overlap. This leads to the W-disjoint orthogonal (WDO) property [4]

$$S_n(\omega,\tau)S_l(\omega,\tau) = 0 \quad \forall \omega, \tau, n \neq l, \tag{6}$$

where the time-frequency representation of the signal $s_n(t)$ is given by the windowed Fourier transform

$$S_n(\omega,\tau) = \int_{-\infty}^{\infty} W(t-\tau)s_n(t)e^{-j\omega t}dt, \tag{7}$$

where $W(t)$ is a window function. Note that this is a mathematical idealisation and in practice it is sufficient that $|S_n(\omega,\tau)S_l(\omega,\tau)|$ be small with high probability [4, 8]. The DUET algorithm uses this assumption to separate N speech signals from one anechoic mixture of the signals by partitioning the time-frequency plane. In order to determine the demixing partitions, DUET uses two mixtures: $x_1(t)$ and $x_2(t)$. For simplicity consider the case where $W(t) = 1$, in which case the system model (1) becomes

$$\begin{bmatrix} X_1(\omega) \\ X_2(\omega) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ \alpha_1 e^{-j\omega\delta_1} & \cdots & \alpha_N e^{-j\omega\delta_N} \end{bmatrix} \begin{bmatrix} A_1(\omega)S_1(\omega) \\ \vdots \\ A_N(\omega)S_N(\omega) \end{bmatrix} + \begin{bmatrix} V_1(\omega) \\ V_2(\omega) \end{bmatrix}. \tag{8}$$

As the planar wave from the nth source $s_n(t)$ travels across the two-element array, the signal seen by the first sensor is attenuated or amplified by a real scalar, α_n, and delayed by δ_n seconds before it reaches the second sensor. Without loss of generality the N channel coefficients $A_1(\omega),\ldots,A_N(\omega)$ can be absorbed by the N source signals, that is, $A_n(\omega)S_n(\omega) \rightarrow S_n(\omega)$, $n = 1,\ldots,N$. In the no-noise case, with W-disjoint orthogonal sources, the two mixtures of the sources are related to at most one of the source signals at any given point in the frequency domain. That is

$$\begin{bmatrix} X_1(\omega) \\ X_2(\omega) \end{bmatrix} = \begin{bmatrix} 1 \\ \alpha_n e^{-j\omega\delta_n} \end{bmatrix} \begin{bmatrix} S_n(\omega) \end{bmatrix} \tag{9}$$

for a given value of frequency $\omega \in \Omega_n$, where

$$\Omega_n = \{\omega : S_n(\omega) \neq 0\} \tag{10}$$

defines the support of $S_n(\omega)$. For such values of ω, the attenuation and delay parameters for the nth source can be determined by

$$\alpha_n = \left|\frac{X_2(\omega)}{X_1(\omega)}\right|, \qquad \delta_n = -\frac{1}{\omega}\angle\left\{\frac{X_2(\omega)}{X_1(\omega)}\right\}, \tag{11}$$

where $\angle\{\alpha e^{j\beta}\} = \beta$. Scanning across ω in the support of the mixtures, (11) will take on N distinct attenuation and delay value pairings; these N pairings are the mixing parameters. When noise is present, (11) will be approximately satisfied and a two-dimensional histogram in attenuation-delay space constructed using (11) will contain N peaks, one for each source, with peak locations corresponding to the mixing parameters. Labelling each ω with the peak its corresponding amplitude-delay estimate falls closest to, we partition one of the mixtures in the frequency domain into the original source signals.

Using the narrowband assumption in the time-frequency domain, that is, if $s_1(t) = s(t)$ and $s_2(t) = s(t-\delta)$ then for all $\delta < \Delta_{\max}$,

$$S_2(\omega,\tau) \approx e^{-j\omega\delta}S_1(\omega,\tau) \tag{12}$$

for some max delay $\Delta_{\max}$, the expression (11) can be extended to the time-frequency domain. Neglecting the effect of noise and assuming (6) is strictly satisfied, the attenuation and delay parameters of the nth signal are then given by

$$\alpha_n = \left|\frac{X_2(\omega,\tau)}{X_1(\omega,\tau)}\right|, \qquad \delta_n = -\frac{1}{\omega}\angle\left\{\frac{X_2(\omega,\tau)}{X_1(\omega,\tau)}\right\} \tag{13}$$

for $(\omega,\tau) \in \Omega_n$, where

$$\Omega_n = \{(\omega,\tau) : S_n(\omega,\tau) \neq 0\} \tag{14}$$

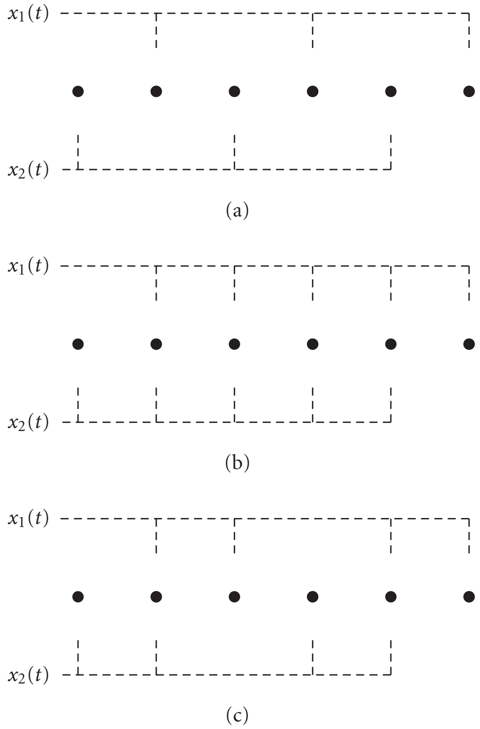
defines the support of $S_n(\omega,\tau)$. Now, similarly scanning across (ω,τ) in the support of the mixtures, (13) will take on N distinct attenuation and delay value pairings, the mixing parameters. When noise is present and (6) is approximately satisfied, (13) will be approximately satisfied and a two-dimensional histogram in attenuation-delay space constructed using (13) will again contain N peaks, one for each source, with peak locations corresponding to the mixing parameters. Labelling each (ω,τ) with the peak its corresponding amplitude-delay estimate falls closest to, one of the mixtures is then partitioned in the time-frequency domain into the original source signals.

The remainder of this paper has the following structure. Section 2 describes the classic ESPRIT direction of arrival estimation algorithm and the development of the hard DESPRIT, soft DESPRIT, and echoic DESPRIT extensions to the DUET blind source separation technique. Section 3 gives an algorithmic description of the echoic DESPRIT technique. Section 4 describes a set of synthetic and real-room experiments designed to demonstrate properties and advantages of the hard DESPRIT, soft DESPRIT, and echoic DESPRIT extensions to the DUET blind source separation technique.

2. THE DESPRIT TECHNIQUE

2.1. The ESPRIT direction of arrival estimation algorithm

Classic direction of arrival estimation techniques such as MUSIC [22] and ESPRIT [23] aim to find the N angles of arrival of N uncorrelated narrowband signals $s_1(t), s_2(t), \ldots, s_N(t)$ as they impinge onto an array of M sensors. With accurate estimation, beamforming can be performed to separate the N signals. We present here a synopsis of the ESPRIT algorithm, for further details consult [23–25].



(a)

(b)

(c)

Figure 2: ESPRIT subarray separation of a uniform linear array in the case of $\mathcal{M} = M/2$, $\mathcal{M} = M-1$, and $M/2 < \mathcal{M} < M-1$.

For narrowband signals of centre frequency ω_0, a time lag can be approximated by a phase rotation, that is, for all $\delta < \Delta_{\max}$,

$$\hat{s}(t-\delta) \approx e^{-j\omega_0\delta}\hat{s}(t) \tag{15}$$

for some max delay $\Delta_{\max}$, where $\hat{s}(t)$ is the complex analytic representation of real signal $s(t)$. In this section only, all functions of time are assumed to be in their complex analytic representation and for notational simplicity we will drop the $\{\hat{\cdot}\}$ from them. ESPRIT separates the M mixtures into two subsets of $\mathcal{M}$ mixtures each, where $M/2 \le \mathcal{M} \le M-1$. The first subarray of $\mathcal{M}$ sensors must be displaced from a second identical subarray of $\mathcal{M}$ sensors by a common displacement vector. In the case of a uniform linear array (see Figure 2), the subarrays can be chosen to maximise overlap, that is, $\mathcal{M} = M-1$ and the output of the first subarray may be expressed as

$$\begin{bmatrix} x_1(t)\\ x_2(t)\\ \vdots\\ x_{M-1}(t)\end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1\\ \phi_1(\omega_0) & \cdots & \phi_N(\omega_0)\\ \vdots & & \vdots\\ \phi_1^{M-2}(\omega_0) & \cdots & \phi_N^{M-2}(\omega_0)\end{bmatrix} \times \begin{bmatrix} A_1(\omega_0)s_1(t)\\ \vdots\\ A_N(\omega_0)s_N(t)\end{bmatrix} + \begin{bmatrix} v_1(t)\\ v_2(t)\\ \vdots\\ v_{M-1}(t)\end{bmatrix} \tag{16}$$

and the output of the second subarray may be expressed as

$$\begin{bmatrix} x_2(t)\\ x_3(t)\\ \vdots\\ x_M(t)\end{bmatrix} = \begin{bmatrix} \phi_1(\omega_0) & \cdots & \phi_N(\omega_0)\\ \phi_1^2(\omega_0) & \cdots & \phi_N^2(\omega_0)\\ \vdots & & \vdots\\ \phi_1^{M-1}(\omega_0) & \cdots & \phi_N^{M-1}(\omega_0)\end{bmatrix} \times \begin{bmatrix} A_1(\omega_0)s_1(t)\\ \vdots\\ A_N(\omega_0)s_N(t)\end{bmatrix} + \begin{bmatrix} v_2(t)\\ v_3(t)\\ \vdots\\ v_M(t)\end{bmatrix}, \tag{17}$$

where $\phi_n(\omega_0) = \alpha_n e^{-j\omega_0\delta_n}$, and α_n and δ_n are the attenuation and delay experienced by the nth signal as it travels from the first subarray to the second. Both data vectors can be stacked to form a $2(M-1)\times 1$ time-varying vector

$$\mathbf{z}(t) = \begin{bmatrix} \mathbf{x}_1(t)\\ \mathbf{x}_2(t)\end{bmatrix} = \begin{bmatrix} \mathbf{A}(\omega_0)\\ \mathbf{A}(\omega_0)\mathbf{\Phi}(\omega_0)\end{bmatrix}[\mathbf{s}(t)] + [\mathbf{v}(t)], \tag{18}$$

where

$$\mathbf{A}(\omega_0) = \begin{bmatrix} A_1(\omega_0) & \cdots & A_N(\omega_0)\\ A_1(\omega_0)\phi_1(\omega_0) & \cdots & A_N(\omega_0)\phi_N(\omega_0)\\ \vdots & & \vdots\\ A_1(\omega_0)\phi_1^{M-2}(\omega_0) & \cdots & A_N(\omega_0)\phi_N^{M-2}(\omega_0)\end{bmatrix},$$

$$\mathbf{\Phi}(\omega_0) = \begin{bmatrix} \phi_1(\omega_0) & & \\ & \ddots & \\ & & \phi_N(\omega_0)\end{bmatrix}, \tag{19}$$

and the entries of $\mathbf{v}(t)$ are noise terms. It follows that the spatial covariance matrix

$$\mathbf{R}_{zz} \doteq E\{[\mathbf{z}(t)][\mathbf{z}(t)]^H\} \tag{20}$$

is of the form

$$\mathbf{R}_{zz} = \begin{bmatrix} \mathbf{A}(\omega_0)\\ \mathbf{A}(\omega_0)\mathbf{\Phi}(\omega_0)\end{bmatrix}\mathbf{R}_{ss}\begin{bmatrix} \mathbf{A}(\omega_0)\\ \mathbf{A}(\omega_0)\mathbf{\Phi}(\omega_0)\end{bmatrix}^H + \mathbf{R}_{\mathbf{vv}}, \tag{21}$$

where

$$\mathbf{R}_{ss} = E\{[\mathbf{s}(t)][\mathbf{s}(t)]^H\}, \qquad \mathbf{R}_{\mathbf{vv}} = E\{[\mathbf{v}(t)][\mathbf{v}(t)]^H\}, \tag{22}$$

and $E\{\cdot\}$ is the expectation operator. ESPRIT assumes $\mathbf{R}_{ss}$ is of full rank and thus for a high signal-to-noise ratio the singular value decomposition (SVD) of $\mathbf{R}_{zz}$ can be computed to give

$$\mathbf{R}_{zz}\begin{bmatrix} \mathbf{E}_s & \mathbf{E}_v\end{bmatrix}\begin{bmatrix} \mathbf{\Lambda} & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma}\end{bmatrix}\begin{bmatrix} \mathbf{E}_s & \mathbf{E}_v\end{bmatrix}^H, \tag{23}$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 + \sigma_1^2 & & \\ & \ddots & \\ & & \lambda_N + \sigma_N^2 \end{bmatrix}, \tag{24}$$
$$\Sigma = \begin{bmatrix} \sigma_{N+1}^2 & & \\ & \ddots & \\ & & \sigma_{2(M-1)}^2 \end{bmatrix},$$

$\lambda_1,\dots,\lambda_N \gg \sigma_1^2,\dots,\sigma_{2(M-1)}^2$, $\lambda_1,\lambda_2,\dots,\lambda_N$ are related to the source signal powers and $\sigma_1^2,\sigma_2^2,\dots,\sigma_{2(M-1)}^2$ are related to the variance of the sensor noise. The N column vectors of $\mathbf{E_s}$ are associated with the singular values of Λ and they are said to span the signal subspace. The $2M-N-2$ column vectors of $\mathbf{E_v}$ associated with the singular values of Σ span the nullspace of $\mathbf{E_s}$, which is often referred to as the noise subspace. (It is understood that $\mathbf{R}_{zz}$ and its singular value decomposition (23) have a dependence upon the centre frequency ω_0, the notation omits reference to this variable.) It follows that for high signal-to-noise ratios there exists a nonsingular matrix $\mathbf{S}$, such that

$$\mathbf{E_s} = \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \end{bmatrix} \approx \begin{bmatrix} \mathbf{A}(\omega_0) \\ \mathbf{A}(\omega_0)\mathbf{\Phi}(\omega_0) \end{bmatrix} \mathbf{S}, \tag{25}$$

where $\mathbf{E}_1$ and $\mathbf{E}_2$ are the signal subspaces corresponding to the first and second subarrays, respectively. Providing that $\mathbf{E}_1$ and $\mathbf{E}_2$ are of rank N, the diagonal matrix $\mathbf{\Phi}(\omega_0)$ is related to $\mathbf{E}_1^\dagger \mathbf{E}_2$ via a similarity transform

$$\mathbf{E}_1^\dagger \mathbf{E}_2 \approx \mathbf{S}^{-1}\mathbf{\Phi}(\omega_0)\mathbf{S}, \tag{26}$$

where $[\cdot]^\dagger$ denotes the Moore-Penrose pseudoinverse, a least-square solution to the nosnquare matrix inverse. The ESPRIT algorithm may be summarised in the following way.

Step 1. M narrowband mixtures $x_1(t),\dots,x_M(t)$ of centre frequency ω_0 are sampled at the K adjacent time points $t_1,\dots,t_K$, these sampled mixtures are used to construct the data matrix

$$\mathbf{z} = \begin{bmatrix} x_1(t_1) & \cdots & x_1(t_K) \\ \vdots & & \vdots \\ x_{M-1}(t_1) & \cdots & x_{M-1}(t_K) \\ x_2(t_1) & \cdots & x_2(t_K) \\ \vdots & & \vdots \\ x_M(t_1) & \cdots & x_M(t_K) \end{bmatrix} \tag{27}$$

and an estimate of the spatial covariance matrix is computed

$$\widehat{\mathbf{R}}_{zz} = \mathrm{zz}^H. \tag{28}$$

Step 2. The singular value decomposition (23) is computed:

$$\widehat{\mathbf{R}}_{zz} \Longrightarrow \begin{bmatrix} \mathbf{E_1} & \mathbf{E}_{\mathbf{v}_1} \\ \mathbf{E_2} & \mathbf{E}_{\mathbf{v}_2} \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma} \end{bmatrix} \begin{bmatrix} \mathbf{E_1} & \mathbf{E}_{\mathbf{v}_1} \\ \mathbf{E_2} & \mathbf{E}_{\mathbf{v}_2} \end{bmatrix}^H \tag{29}$$

($\mathbf{E}_{\mathbf{v}_1}$ and $\mathbf{E}_{\mathbf{v}_2}$ are the top and bottom $M-1$ rows of $\mathbf{E_v}$).

Step 3. The N mixing parameters are estimated via an eigenvalue decomposition

$$(\widetilde{\phi}_1(\omega_0),\dots,\widetilde{\phi}_N(\omega_0)) = \text{eigs}\,\{\mathbf{E}_1^\dagger \mathbf{E}_2\}, \tag{30}$$

where eigs$\{\mathbf{H}\}$ denotes the eigenvalues of the matrix $\mathbf{H}$.

2.1.1. Simplification of ESPRIT technique

As an example we consider the no-noise mixing model

$$\begin{bmatrix} x_1(t_1) & \cdots & x_1(t_K) \\ x_2(t_1) & \dots & x_2(t_K) \\ x_3(t_1) & \dots & x_3(t_K) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \phi_1(\omega_0) & \phi_2(\omega_0) \\ \phi_1^2(\omega_0) & \phi_2^2(\omega_0) \end{bmatrix} \begin{bmatrix} s_1(t_1) & \dots & s_1(t_K) \\ s_2(t_1) & \dots & s_2(t_K) \end{bmatrix}, \tag{31}$$

the spatial covariance matrix is constructed according to Step 1:

$$\widehat{\mathbf{R}}_{zz} = \begin{bmatrix} x_1(t_1) & \cdots & x_1(t_K) \\ x_2(t_1) & \cdots & x_2(t_K) \\ x_2(t_1) & \cdots & x_2(t_K) \\ x_3(t_1) & \cdots & x_3(t_K) \end{bmatrix} \times \begin{bmatrix} x_1^*(t_1) & x_2^*(t_1) & x_2^*(t_1) & x_3^*(t_1) \\ \vdots & \vdots & \vdots & \vdots \\ x_1^*(t_K) & x_2^*(t_K) & x_2^*(t_K) & x_3^*(t_K) \end{bmatrix} \tag{32}$$

and the singular value decomposition is computed as in Step 2 yielding the 2×2 signal subspace matrices $\mathbf{E}_1$ and $\mathbf{E}_2$. The mixing parameter estimates $\widetilde{\phi}_1(\omega_0)$ and $\widetilde{\phi}_2(\omega_0)$ are then given by Step 3

$$(\widetilde{\phi}_1(\omega_0),\widetilde{\phi}_2(\omega_0)) = \text{eigs}\,\{\mathbf{E}_1^{-1}\mathbf{E}_2\}. \tag{33}$$

The computation of the singular value decomposition in Step 2 is not strictly necessary in this case, $\mathbf{E}_1$ and $\mathbf{E}_2$ may be simply replaced by

$$\mathbf{E}_1 = \begin{bmatrix} x_1(t_1) & x_1(t_2) \\ x_2(t_1) & x_2(t_2) \end{bmatrix}, \qquad \mathbf{E}_2 = \begin{bmatrix} x_2(t_1) & x_2(t_2) \\ x_3(t_1) & x_3(t_2) \end{bmatrix} \tag{34}$$

since

$$\begin{aligned} &\begin{bmatrix} x_1(t_1) & x_1(t_2) \\ x_2(t_1) & x_2(t_2) \end{bmatrix}^{-1} \begin{bmatrix} x_2(t_1) & x_2(t_2) \\ x_3(t_1) & x_3(t_2) \end{bmatrix} \\ &\quad = \begin{bmatrix} s_1(t_1) & s_1(t_2) \\ s_2(t_1) & s_2(t_2) \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 \\ \phi_1(\omega_0) & \phi_2(\omega_0) \end{bmatrix}^{-1} \\ &\qquad \times \begin{bmatrix} \phi_1(\omega_0) & \phi_2(\omega_0) \\ \phi_1^2(\omega_0) & \phi_2^2(\omega_0) \end{bmatrix} \begin{bmatrix} s_1(t_1) & s_1(t_2) \\ s_2(t_1) & s_2(t_2) \end{bmatrix} \\ &\quad = \begin{bmatrix} s_1(t_1) & s_1(t_2) \\ s_2(t_1) & s_2(t_2) \end{bmatrix}^{-1} \begin{bmatrix} \phi_1(\omega_0) & 0 \\ 0 & \phi_2(\omega_0) \end{bmatrix} \begin{bmatrix} s_1(t_1) & s_1(t_2) \\ s_2(t_1) & s_2(t_2) \end{bmatrix}, \end{aligned} \tag{35}$$

where t_1 and t_2 are two adjacent sample points. As in (26) the mixing parameters are related to $\mathbf{E}_1^{-1}\mathbf{E}_2$ via a similarity transform, that is,

$$\mathbf{E}_1^{-1}\mathbf{E}_2 = \mathbf{S}^{-1}\boldsymbol{\Phi}(\omega_0)\mathbf{S},$$
$$\mathbf{S} = \begin{bmatrix} s_1(t_1) & s_1(t_2) \\ s_2(t_1) & s_2(t_2) \end{bmatrix}, \qquad \boldsymbol{\Phi}(\omega_0) = \begin{bmatrix} \phi_1(\omega_0) & 0 \\ 0 & \phi_2(\omega_0) \end{bmatrix}. \tag{36}$$

It follows that in general for M noiseless mixtures Step 3 may be modified to become

$$(\widetilde{\phi}_1(\omega_0), \ldots, \widetilde{\phi}_{M-1}(\omega_0)) = \text{eigs}\,\{\mathbf{E}_1^{-1}\mathbf{E}_2\}, \tag{37}$$

where

$$\mathbf{E}_1 = \begin{bmatrix} x_1(t_1) & \cdots & x_1(t_{M-1}) \\ \vdots & & \vdots \\ x_{M-1}(t_1) & \cdots & x_{M-1}(t_{M-1}) \end{bmatrix},$$
$$\mathbf{E}_2 = \begin{bmatrix} x_2(t_1) & \cdots & x_2(t_{M-1}) \\ \vdots & & \vdots \\ x_M(t_1) & \cdots & x_M(t_{M-1}) \end{bmatrix}, \tag{38}$$

and $t_1, t_2, \ldots, t_{M-1}$ are adjacent time samples.

It is also possible to switch the order of the matrix multiplication, that is,

$$(\widetilde{\phi}_1(\omega_0), \ldots, \widetilde{\phi}_{M-1}(\omega_0)) = \text{eigs}\,\{\mathbf{E}_2\mathbf{E}_1^{\dagger}\}; \tag{39}$$

this approach removes the restriction that $M-1$ time samples are used to estimate $M-1$ mixing parameters, now $K \geq M-1$ samples may be used to estimate $M-1$ mixing parameters. This can be shown for the $M = 3$ case:

$$\mathbf{E}_1 = \begin{bmatrix} x_1(t_1) & \cdots & x_1(t_K) \\ x_2(t_1) & \cdots & x_2(t_K) \end{bmatrix},$$
$$\mathbf{E}_2 = \begin{bmatrix} x_1(t_1) & \cdots & x_1(t_K) \\ x_2(t_1) & \cdots & x_2(t_K) \end{bmatrix},$$
$$\begin{aligned} \mathbf{E}_2\mathbf{E}_1^{\dagger} &= \begin{bmatrix} x_2(t_1) & \cdots & x_2(t_K) \\ x_3(t_1) & \cdots & x_3(t_K) \end{bmatrix} \begin{bmatrix} x_1(t_1) & \cdots & x_1(t_K) \\ x_2(t_1) & \cdots & x_2(t_K) \end{bmatrix}^{\dagger} \\ &= \begin{bmatrix} \phi_1(\omega_0) & \phi_2(\omega_0) \\ \phi_1^2(\omega_0) & \phi_2^2(\omega_0) \end{bmatrix} \begin{bmatrix} s_1(t_1) & \cdots & s_1(t_K) \\ s_2(t_1) & \cdots & s_2(t_K) \end{bmatrix} \\ &\quad \times \begin{bmatrix} s_1(t_1) & \cdots & s_1(t_K) \\ s_2(t_1) & \cdots & s_2(t_K) \end{bmatrix}^{\dagger} \begin{bmatrix} 1 & 1 \\ \phi_1(\omega_0) & \phi_2(\omega_0) \end{bmatrix}^{-1} \\ &= \begin{bmatrix} 1 & 1 \\ \phi_1(\omega_0) & \phi_2(\omega_0) \end{bmatrix} \begin{bmatrix} \phi_1(\omega_0) & 0 \\ 0 & \phi_2(\omega_0) \end{bmatrix} \\ &\quad \times \begin{bmatrix} 1 & 1 \\ \phi_1(\omega_0) & \phi_2(\omega_0) \end{bmatrix}^{-1} \\ &= \mathbf{A}(\omega_0)\boldsymbol{\Phi}(\omega_0)\mathbf{A}^{-1}(\omega_0), \end{aligned} \tag{40}$$

where

$$\mathbf{A}(\omega_0) = \begin{bmatrix} 1 & 1 \\ \phi_1(\omega_0) & \phi_2(\omega_0) \end{bmatrix},$$
$$\boldsymbol{\Phi}(\omega_0) = \begin{bmatrix} \phi_1(\omega_0) & 0 \\ 0 & \phi_2(\omega_0) \end{bmatrix}. \tag{41}$$

Again it follows that in general for M mixtures Step 3 may be modified to become

$$(\widetilde{\phi}_1(\omega_0), \ldots, \widetilde{\phi}_{M-1}(\omega_0)) = \text{eigs}\,\{\mathbf{E}_2\mathbf{E}_1^{\dagger}\}, \tag{42}$$

where

$$\mathbf{E}_1 = \begin{bmatrix} x_1(t_1) & \cdots & x_1(t_K) \\ \vdots & & \vdots \\ x_{M-1}(t_1) & \cdots & x_{M-1}(t_K) \end{bmatrix},$$
$$\mathbf{E}_2 = \begin{bmatrix} x_2(t_1) & \cdots & x_2(t_K) \\ \vdots & & \vdots \\ x_M(t_1) & \cdots & x_M(t_K) \end{bmatrix}, \tag{43}$$

and $t_1, t_2, \ldots, t_K$ are adjacent time samples with $K \geq M - 1$. The simplified ESPRIT algorithm may be summarised as follows.

Step 1. $K \geq M - 1$ time samples of M narrowband mixtures $x_1(t), x_2(t), \ldots, x_M(t)$ are used to construct the matrices

$$\mathbf{E}_1 = \begin{bmatrix} x_1(t_1) & \ldots & x_1(t_K) \\ \vdots & & \vdots \\ x_{M-1}(t_1) & \cdots & x_{M-1}(t_K) \end{bmatrix},$$
$$\mathbf{E}_2 = \begin{bmatrix} x_2(t_1) & \cdots & x_2(t_K) \\ \vdots & & \vdots \\ x_M(t_1) & \cdots & x_M(t_K) \end{bmatrix}. \tag{44}$$

Step 2. The $M - 1$ mixing parameters are estimated via an eigenvalue decomposition

$$(\widetilde{\phi}_1(\omega_0), \ldots, \widetilde{\phi}_{M-1}(\omega_0)) = \text{eigs}\,\{\mathbf{E}_2\mathbf{E}_1^{\dagger}\}. \tag{45}$$

2.1.2. Combining DUET and ESPRIT

The $M - 1$ eigenvalues obtained in (37) or in (42) serve as $M - 1$ mixing parameter estimates $\widetilde{\phi}_1(\omega_0), \ldots, \widetilde{\phi}_{M-1}(\omega_0)$ and the $M - 1$ attenuation and delay estimates are then given as

$$\widetilde{\alpha}_m = |\widetilde{\phi}_m(\omega_0)|,$$
$$\widetilde{\delta}_m = -\frac{1}{\omega_0}\angle\widetilde{\phi}_m(\omega_0), \quad m = 1, \ldots, M - 1 \tag{46}$$

(it may be noted that the classic ESPRIT algorithm makes the assumption that the attenuation parameters are unity, i.e., $\alpha_1 = \alpha_2 = \cdots = \alpha_{M-1} = 1$). The $M - 1$ delay estimates $\widetilde{\delta}_1, \ldots, \widetilde{\delta}_{M-1}$ are related to $M - 1$ angle of arrival estimates $\widetilde{\theta}_1, \ldots, \widetilde{\theta}_{M-1}$ onto the line of the sensor array via

$$\widetilde{\delta}_m = \frac{D}{c}\cos\{\widetilde{\theta}_m\}, \quad m = 1, 2, \ldots, M - 1, \tag{47}$$

Table 1: Summary of the properties of the three extensions to DUET, where the number of echoic paths is the number of extra (nondirect) paths.

	Sensors utilised	Sources demixed at (ω, τ)	Echoic paths demixed at (ω, τ)
Classic DUET	$M = 2$	$R = 1$	$P = 0$
Hard DESPRIT	$M \geq 2$	$R = 1$	$P = 0$
Soft DESPRIT	$M \geq 2$	$R = M - 1$	$P = 0$
Echoic DESPRIT	$M \geq 2$	$R = \lfloor M/2 \rfloor - P$	$P = \lfloor M/2 \rfloor - R$

where c is the propagation speed and D is the array spacing. Since the attenuation and the delay estimates $(\tilde{\alpha}_1, \tilde{\delta}_1), \dots, (\tilde{\alpha}_{M-1}, \tilde{\delta}_{M-1})$ used in the DUET algorithm to construct the power weighted histogram are also estimated by the ESPRIT algorithm, it is possible to combine both techniques to form a hybrid DUET-ESPRIT technique, which is discussed in the next section. Also in adapting ESPRIT for using with DUET, the narrowband assumption on complex analytic representations (15) is replaced with the narrowband assumption on time-frequency representations (12).

2.2. DESPRIT algorithm outline

The combined DUET-ESPRIT technique (DESPRIT) may be used to extend the DUET blind source separation algorithm to

(1) the multichannel case ($M \geq 2$) using hard DESPRIT, discussed in Section 2.2.1,

(2) the weakened WDO case (where sources may overlap in the time-frequency domain) using soft DESPRIT, discussed in Section 2.2.2,

(3) and the echoic mixing case using echoic DESPRIT, discussed in Section 2.3.

The properties of these extensions are summarised in Table 1. All three of these extensions have the same general outline.

Step 1. An M-element uniform linear array receives M mixtures $x_1(t), x_2(t), \dots, x_M(t)$ of N signals $s_1(t), s_2(t), \dots, s_N(t)$. These M mixtures are transformed into the time-frequency domain using the windowed Fourier transform.

Step 2. Centred at each sample point in the time-frequency domain, the ESPRIT algorithm is performed and the mixing parameters of the source signals active at that point are estimated.

Step 3. The mixing parameter estimates are used to create a weighted histogram, a technique borrowed from the DUET algorithm. The peaks of the histogram indicate sources and the centres of these peaks are used as estimates of the associated mixing parameters.

Step 4. Demixing is performed by inverting a local mixing matrix dependent on the sources active at each time-frequency point. The resulting demixed components are partitioned and combined in a maximum-likelihood align and sum estimator using the labels from the histogram to produce the demixture time-frequency representations.

2.2.1. Hard DESPRIT: a multichannel DUET extension

The hard DESPRIT technique extends DUET to handle $M > 2$ mixtures but still assumes at most one source active at any time-frequency point and an anechoic mixing model. Similar to (20) the time-frequency spatial covariance matrix may be defined as

$$\mathbf{R}_{\mathbf{ZZ}} \doteq \mathrm{E}\{\mathbf{Z}(\omega, \tau)\mathbf{Z}^H(\omega, \tau)\}, \tag{48}$$

where

$$\mathbf{Z}(\omega, \tau) = \begin{bmatrix} X_1(\omega, \tau) \\ \vdots \\ X_{M-1}(\omega, \tau) \\ X_2(\omega, \tau) \\ \vdots \\ X_M(\omega, \tau) \end{bmatrix} \tag{49}$$

and $X_m(\omega, \tau) = \int_{-\infty}^{\infty} W(t-\tau)x_m(t)e^{-j\omega t}dt$. (Again it is understood that $\mathbf{R}_{\mathrm{ZZ}}$ and its singular value decomposition have a dependence upon the time-frequency point (ω, τ), the notation omits reference to these variables.) Under a strong WDO assumption (6) only one source signal is active at each time-frequency point, as a result $\mathbf{R}_{\mathrm{ZZ}}$ is at most rank one and has a singular value decomposition of the form

$$\mathbf{R}_{\mathrm{ZZ}} = \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \end{bmatrix}_{2(M-1)\times 1} \begin{bmatrix} \mathbf{E}_1^H & \mathbf{E}_2^H \end{bmatrix}_{1\times 2(M-1)}. \tag{50}$$

It follows that

$$\tilde{\phi}_n(\omega, \tau) = \mathbf{E}_1^{\dagger}\mathbf{E}_2 \quad \forall(\omega, \tau) \in \Omega_n \tag{51}$$

is a complex scalar corresponding to the estimated mixing parameter of the nth source signal. Furthermore when the expectation operator $\mathrm{E}\{\cdot\}$ is approximated using an instantaneous estimate, $\tilde{\phi}_n(\omega, \tau)$ is given by

$$\tilde{\phi}_n(\omega, \tau) = \left[\mathbf{X}_1(\omega, \tau)\right]^{\dagger}\left[\mathbf{X}_2(\omega, \tau)\right] \quad \forall(\omega, \tau) \in \Omega_n, \tag{52}$$

where $\mathbf{X}_1(\omega, \tau) = [X_1(\omega, \tau), \dots, X_{M-1}(\omega, \tau)]^T$ and $\mathbf{X}_2(\omega, \tau) = [X_2(\omega, \tau), \dots, X_M(\omega, \tau)]^T$, this expression may be restated as

$$\tilde{\phi}_n(\omega, \tau) = \frac{\sum_m^{M-1} X_m^*(\omega, \tau)X_{m+1}(\omega, \tau)}{\sum_m^{M-1} X_m^*(\omega, \tau)X_m(\omega, \tau)} \quad \forall(\omega, \tau) \in \Omega_n. \tag{53}$$

In the $M = 2$ case, this expression corresponds to the DUET parameter estimation step (13) and in general for the $M \geq 2$ case, it corresponds to the parameter estimation step of a multichannel DUET extension [5].

2.2.2. Soft DESPRIT: the weakened WDO assumption

The soft DESPRIT technique extends DUET to handle $M > 2$ mixtures and also allows for more than one source to be active at a given time-frequency point. It assumes, as DUET and hard DESPRIT do, anechoic mixing. Soft DESPRIT is an implementation of DESPRIT under a weakened WDO assumption [6]:

$$S_{n_1}(\omega,\tau) \times \cdots \times S_{n_M}(\omega,\tau) = 0 \quad \forall \omega, \tau, n_l \neq n_k,\ l \neq k. \tag{54}$$

This weakened WDO assumption allows source signals to overlap in the time-frequency domain, with up to $M-1$ source signals coexisting at any given time-frequency point. Since the strong WDO assumption (6) used by DUET is only ever approximately true, the weakened WDO assumption may be adopted as a more realistic source model. The spatial covariance matrix (48) may be approximated as

$$\mathbf{R}_{ZZ} \approx \frac{1}{2\kappa+1}\sum_{k=-\kappa}^{k=\kappa}\left[\mathbf{Z}(\omega,\tau+k\Delta T)\right]\left[\mathbf{Z}(\omega,\tau+k\Delta T)\right]^H, \tag{55}$$

where ΔT is the separation between adjacent time samples in the time-frequency domain and $\kappa \geq M/2-1$. The expectation operator $\mathrm{E}\{\cdot\}$ is approximated by averaging over the 2κ samples adjacent to the time-frequency point of interest.

In accordance with our simplified ESPRIT algorithm, the $M-1$ mixing parameter estimates $\widetilde{\phi}_1(\omega,\tau), \widetilde{\phi}_2(\omega,\tau), \ldots, \widetilde{\phi}_{M-1}(\omega,\tau)$ are given by (42)

$$\left(\widetilde{\phi}_1(\omega,\tau),\ldots,\widetilde{\phi}_{M-1}(\omega,\tau)\right) = \operatorname{eigs}\{\mathbf{E}_2\mathbf{E}_1^{\dagger}\}, \tag{56}$$

where

$$\begin{gathered}\mathbf{E}_1 = \begin{bmatrix} x_1(\omega,\tau_1) & \ldots & x_1(\omega,\tau_K) \\ \vdots & & \vdots \\ x_{M-1}(\omega,\tau_1) & \cdots & x_{M-1}(\omega,\tau_K)\end{bmatrix}, \\ \mathbf{E}_2 = \begin{bmatrix} x_2(\omega,\tau_1) & \cdots & x_2(\omega,\tau_K) \\ \vdots & & \vdots \\ x_M(\omega,\tau_1) & \cdots & x_M(\omega,\tau_K)\end{bmatrix},\end{gathered} \tag{57}$$

and $\tau_1, \tau_2, \ldots, \tau_K$ are adjacent time points with $K \geq M-1$.

2.3. Echoic DESPRIT: extending to reverberant environments

The echoic DESPRIT extension to DUET leverages $M > 2$ mixtures to demix up to $\lfloor M/2 \rfloor$ sources from each time-frequency point, as in the soft DESPRIT extension. However in echoic DESPRIT the $\lfloor M/2 \rfloor$ sources can consist of the same source arriving on different paths ($\lfloor\cdot\rfloor$ denotes rounding down to the nearest integer).

2.3.1. Mixing parameter estimation of coherent source signals

The echoic mixing model (3) makes the assumption that a source signal $s_n(t)$ propagates upon P_n distinct echoic paths to the sensor array. In order to successfully demix echoic mixtures, it follows that a parameter estimation step must allow for source signals to be coherent (i.e., fully correlated). Both the DUET and the classic ESPRIT algorithms face problems when source signals are coherent.

2.3.2. DUET fails for coherent source signals

For DUET in the no-noise case and $W(t) = 1$, $M = 2$ mixtures of $N = 2$ source signals are of the form

$$\begin{bmatrix} X_1(\omega) \\ X_2(\omega)\end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \phi_1(\omega) & \phi_2(\omega)\end{bmatrix}\begin{bmatrix} A_1(\omega)S_1(\omega) \\ A_2(\omega)S_2(\omega)\end{bmatrix}, \tag{58}$$

if the 2 sources are coherent, $S_1(\omega) = S_2(\omega) = S(\omega)$, then

$$\begin{aligned} X_1(\omega) &= \left(A_1(\omega)+A_2(\omega)\right)S(\omega), \\ X_2(\omega) &= \left(A_1(\omega)\alpha_1 e^{-j\omega\delta_1} + A_2(\omega)\alpha_2 e^{-j\omega\delta_2}\right)S(\omega).\end{aligned} \tag{59}$$

The DUET parameter estimation step yields

$$\begin{aligned} \widetilde{\alpha}(\omega) &= \left|\frac{X_2(\omega)}{X_1(\omega)}\right| = \left|\frac{A_1(\omega)\alpha_1 e^{-j\omega\delta_1} + A_2(\omega)\alpha_2 e^{-j\omega\delta_2}}{A_1(\omega)+A_2(\omega)}\right|, \\ \widetilde{\delta}(\omega) &= -\frac{1}{\omega}\angle\frac{X_2(\omega)}{X_1(\omega)} \\ &= -\frac{1}{\omega}\angle\frac{A_1(\omega)\alpha_1 e^{-j\omega\delta_1} + A_2(\omega)\alpha_2 e^{-j\omega\delta_2}}{A_1(\omega)+A_2(\omega)}\end{aligned} \tag{60}$$

at each frequency point, which will not result in a peak in the weighted histogram corresponding to the mixing parameter pair of either arrivals, as $\widetilde{\alpha}(\omega)$ and $\widetilde{\delta}(\omega)$ depend on ω. DUET fails in this case to correctly estimate the 2 mixing parameter pairs and this failing is true in general for N coherent sources $S_1(\omega) = \cdots = S_N(\omega) = S(\omega)$.

2.3.3. ESPRIT fails for N coherent source signals

For ESPRIT in the no noise case, M mixtures of N narrowband coherent source signals of centre frequency ω_0, are of the form

$$\mathbf{z}(t) = \begin{bmatrix} \mathbf{A}(\omega_0) \\ \mathbf{A}(\omega_0)\mathbf{\Phi}(\omega_0)\end{bmatrix}\begin{bmatrix} s(t) \\ \vdots \\ s(t)\end{bmatrix}. \tag{61}$$

The spatial covariance matrix may be written as

$$\mathbf{R}_{zz} = \mathrm{E}\left\{\begin{bmatrix}\mathbf{A}(\omega_0)\\ \mathbf{A}(\omega_0)\mathbf{\Phi}(\omega_0)\end{bmatrix}\begin{bmatrix}s(t)\\ \vdots\\ s(t)\end{bmatrix} \times \begin{bmatrix}s^*(t) & \cdots & s^*(t)\end{bmatrix}\begin{bmatrix}\mathbf{A}(\omega_0)\\ \mathbf{A}(\omega_0)\mathbf{\Phi}(\omega_0)\end{bmatrix}^H\right\}, \tag{62}$$

$$\mathbf{R}_{zz} = \mathrm{E}\{s(t)s^*(t)\}\begin{bmatrix}\mathbf{A}(\omega_0)\\ \mathbf{A}(\omega_0)\mathbf{\Phi}(\omega_0)\end{bmatrix}\begin{bmatrix}1 & \dots & 1\\ \vdots & & \vdots\\ 1 & \dots & 1\end{bmatrix}_{N\times N} \times \begin{bmatrix}\mathbf{A}(\omega_0)\\ \mathbf{A}(\omega_0)\mathbf{\Phi}(\omega_0)\end{bmatrix}^H.$$

Since an $N \times N$ matrix of all ones is of rank one, the rank of $\mathbf{R}_{zz}$ will be at most one, and for the rank one case the singular value decomposition will be of the form

$$\mathbf{R}_{zz} = \begin{bmatrix}\mathbf{E}_1\\ \mathbf{E}_2\end{bmatrix}_{2(M-1)\times 1}\begin{bmatrix}\mathbf{E}_1^H & \mathbf{E}_2^H\end{bmatrix}_{1\times 2(M-1)}, \tag{63}$$

it follows that

$$[\mathbf{E}_1]^{\dagger}_{M-1\times 1}[\mathbf{E}_2]_{1\times M-1} \tag{64}$$

will also be of rank one and so only a single mixing parameter estimate

$$\tilde{\phi}(\omega_0) = \frac{\mathbf{A}(\omega_0)\mathbf{\Phi}(\omega_0)\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}_{N\times 1}}{\mathbf{A}(\omega_0)\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}_{N\times 1}} \tag{65}$$

may be obtained, thus ESPRIT fails in echoic environments.

2.3.4. Unitary ESPRIT for 2 coherent source signals

It is demonstrated in [26] that the unitary ESPRIT algorithm has the ability to estimate the angles of arrival of 2 completely coherent narrowband source signals. This property relies upon a modified data matrix construction technique which may be stated as

$$\mathbf{z}(t) = \begin{bmatrix}x_1(t) & x_{M-1}^*(t)\\ \vdots & \vdots\\ x_{M-1}(t) & x_2^*(t)\\ x_2(t) & x_M^*(t)\\ \vdots & \vdots\\ x_M(t) & x_1^*(t)\end{bmatrix}. \tag{66}$$

In the no noise case, M mixtures of 2 narrowband source signals of centre frequency ω_0 have a corresponding data matrix of the form

$$\mathbf{z}(t) = \begin{bmatrix}\mathbf{A}(\omega_0)\\ \mathbf{A}(\omega_0)\mathbf{\Phi}(\omega_0)\end{bmatrix}\mathbf{\Psi}(\omega_0)s(t), \tag{67}$$

where

$$\mathbf{A}(\omega_0) = \begin{bmatrix}A_1 & A_2\\ A_1e^{-j\omega_0\delta_1} & A_2e^{-j\omega_0\delta_2}\\ \vdots & \vdots\\ A_1e^{-j\omega_0(M-2)\delta_1} & A_2e^{-j\omega_0(M-2)\delta_2}\end{bmatrix},$$
$$\mathbf{\Phi}(\omega_0) = \begin{bmatrix}e^{-j\omega_0\delta_1} & 0\\ 0 & e^{-j\omega_0\delta_2}\end{bmatrix},$$
$$\mathbf{\Psi}(\omega_0) = \begin{bmatrix}1 & e^{j\omega_0(M-1)\delta_1}\\ 1 & e^{j\omega_0(M-1)\delta_2}\end{bmatrix}, \tag{68}$$

and the attenuation parameters are assumed to be unity, that is, $\alpha_1 = \cdots = \alpha_N = 1$. The spatial covariance matrix (20) is of the form

$$\mathbf{R}_{zz} = \mathrm{E}\{s(t)s^*(t)\}\begin{bmatrix}\mathbf{A}(\omega_0)\\ \mathbf{A}(\omega_0)\mathbf{\Phi}(\omega_0)\end{bmatrix} \times \mathbf{\Psi}(\omega_0)\mathbf{\Psi}^H(\omega_0)\begin{bmatrix}\mathbf{A}(\omega_0)\\ \mathbf{A}(\omega_0)\mathbf{\Phi}(\omega_0)\end{bmatrix}^H \tag{69}$$

and its singular value decomposition is of the form

$$\mathbf{R}_{zz} = \begin{bmatrix}\mathbf{E}_1\\ \mathbf{E}_2\end{bmatrix}\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{bmatrix}\mathbf{E}_1^H & \mathbf{E}_2^H\end{bmatrix} \tag{70}$$

since $\mathbf{\Psi}(\omega_0)$ is at most rank 2, and it follows that

$$[\mathbf{E}_1]^{\dagger}[\mathbf{E}_2] \tag{71}$$

is at most rank 2 and so can yield at most 2 mixing parameter estimates $\tilde{\phi}_1$ and $\tilde{\phi}_2$.

When $N > 2$ coherent sources are present, $\mathbf{\Psi}(\omega_0)$ is of the form

$$\mathbf{\Psi}(\omega_0) = \begin{bmatrix}1 & e^{j\omega_0(M-1)\delta_1}\\ \vdots & \vdots\\ 1 & e^{j\omega_0(M-1)\delta_N}\end{bmatrix} \tag{72}$$

and since it is only ever rank 2, it follows that only 2 parameter estimates are available.

2.3.5. A new ESPRIT technique for N coherent source signals

It is possible to augment the data matrix construction technique (66) by increasing the number of columns in $\boldsymbol{\Psi}(\omega_0)$ to N, this will make it possible for $\boldsymbol{\Psi}(\omega_0)$ to be of rank N and so it is possible to estimate the mixing parameters of N coherent source signals. Hence adding structure across the columns of $\mathbf{z}(t)$ allows parameter estimation of correlated and even completely coherent sources. M mixtures of N possibly coherent narrowband source signals of centre frequency ω_0 are stacked in a matrix of the form

$$\mathbf{z}(t) = \begin{bmatrix} x_1(t) & x_2(t) & \cdots & x_{\lfloor M/2 \rfloor}(t) \\ \vdots & \vdots & & \vdots \\ x_{\lceil M/2 \rceil} & x_{\lceil M/2 \rceil + 1}(t) & \cdots & x_{M-1}(t) \\ x_2(t) & x_3(t) & \cdots & x_{\lfloor M/2 \rfloor + 1}(t) \\ \vdots & \vdots & & \vdots \\ x_{\lceil M/2 \rceil + 1}(t) & x_{\lceil M/2 \rceil + 2}(t) & \cdots & x_M(t) \end{bmatrix}, \qquad (73)$$

where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ denote rounding up and down to the nearest integer. In the no-noise case this may be rewritten as

$$\mathbf{z}(t) = \begin{bmatrix} \mathbf{A}(\omega_0) \\ \mathbf{A}(\omega_0)\boldsymbol{\Phi}(\omega_0) \end{bmatrix} \boldsymbol{\Psi}(\omega_0) s(t), \qquad (74)$$

where

$$\boldsymbol{\Psi}(\omega_0) = \begin{bmatrix} 1 & \phi_1(\omega_0) & \cdots & \phi_1^{\lfloor M/2 \rfloor - 1}(\omega_0) \\ \vdots & \vdots & & \vdots \\ 1 & \phi_N(\omega_0) & \cdots & \phi_N^{\lfloor M/2 \rfloor - 1}(\omega_0) \end{bmatrix}. \qquad (75)$$

The spatial covariance matrix

$$\mathbf{R}_{zz} = \mathrm{E}\{\mathbf{z}(t)\mathbf{z}^H(t)\} \qquad (76)$$

is of the form

$$\begin{aligned} &= \mathrm{E}\{s(t)s^*(t)\} \begin{bmatrix} \mathbf{A}(\omega_0) \\ \mathbf{A}(\omega_0)\boldsymbol{\Phi}(\omega_0) \end{bmatrix} \\ &\quad \times \boldsymbol{\Psi}(\omega_0)\boldsymbol{\Psi}^H(\omega_0) \begin{bmatrix} \mathbf{A}(\omega_0) \\ \mathbf{A}(\omega_0)\boldsymbol{\Phi}(\omega_0) \end{bmatrix}^H, \end{aligned} \qquad (77)$$

and by choosing $M \geq 2N$, $\mathbf{R}_{zz}$ will have a maximum possible rank of N. For $\mathbf{R}_{zz}$ of rank N there exists a singular value decomposition

$$\mathbf{R}_{zz} = \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix} \begin{bmatrix} \mathbf{E}_1 \\ \mathbf{E}_2 \end{bmatrix}^H, \qquad (78)$$

and it follows that the N eigenvalues of $[\mathbf{E}_1]^{-1}[\mathbf{E}_2]$ are the mixing parameters $\phi_1, \dots, \phi_N$.

```
for $\omega = (-L/2 : 1 : L/2 - 1)2\pi/LT$ do
    for $\tau = (0 : \Delta : K - 1)T$ do
        $X_1(\omega, \tau) = \sum_{k=0}^{K-1} W(kT - \tau)x_1(kT)e^{-j\omega kT}$
        $\vdots$
        $X_M(\omega, \tau) = \sum_{k=0}^{K-1} W(kT - \tau)x_M(kT)e^{-j\omega kT}$
    end
end
```

Algorithm 1

Our simplified ESPRIT algorithm (Section 2.1.1) may be adapted to this new technique.

Step 1. M narrowband mixtures $x_1(t), \dots, x_M(t)$ are used to construct the matrices

$$\begin{aligned} \mathbf{E}_1 &= \begin{bmatrix} x_1(t) & \cdots & x_{\lfloor M/2 \rfloor}(t) \\ \vdots & & \vdots \\ x_{\lceil M/2 \rceil}(t) & \cdots & x_{M-1}(t) \end{bmatrix}, \\ \mathbf{E}_2 &= \begin{bmatrix} x_2(t) & \cdots & x_{\lfloor M/2 \rfloor + 1}(t) \\ \vdots & & \vdots \\ x_{\lceil M/2 \rceil + 1}(t) & \cdots & x_M(t) \end{bmatrix}. \end{aligned} \qquad (79)$$

Step 2. The $\lfloor M/2 \rfloor$ mixing parameters estimates are obtained via an eigenvalue decomposition

$$(\widetilde{\phi}_1(\omega_0), \dots, \widetilde{\phi}_{\lfloor M/2 \rfloor}(\omega_0)) = \mathrm{eigs}\{\mathbf{E}_2\mathbf{E}_1^{\dagger}\}. \qquad (80)$$

Using this new technique a uniform linear array of M sensors may be used to estimate the mixing parameters of one signal travelling on P echoic paths, providing $M \geq 2P$. It follows that this technique will allow the DESPRIT algorithm to demix M echoic mixtures of an arbitrary number of speech source signals providing the maximum number of echoic paths is at most half the number of sensors in the uniform linear array.

3. ALGORITHMIC DESCRIPTION

Step 1. A uniform linear array of M sensors receives M possibly echoic mixtures

$$x_1(t), x_2(t), \dots, x_M(t) \qquad (81)$$

of N speech signals. These M mixture signals are sampled every T seconds, and a window $W(t)$ of length $L \ll KT$ seconds is shifted by multiples of ΔT seconds to perform K/Δ L-point discrete windowed Fourier transforms upon K samples of each mixture (see Algorithm 1).

```
for $\omega = (-L/2:1:L/2-1)2\pi/LT$ do
    for $\tau = (0:\Delta:K-1)T$ do
        $\mathbf{E}_1 = \begin{bmatrix} X_1(\omega,\tau) & \dots & X_{\lfloor M/2\rfloor}(\omega,\tau) \\ \vdots & & \vdots \\ X_{\lceil M/2\rceil}(\omega,\tau) & \dots & X_{M-1}(\omega,\tau) \end{bmatrix}$
        $\mathbf{E}_2 = \begin{bmatrix} X_2(\omega,\tau) & \dots & X_{\lfloor M/2\rfloor+1}(\omega,\tau) \\ \vdots & & \vdots \\ X_{\lceil M/2\rceil+1}(\omega,\tau) & \dots & X_M(\omega,\tau) \end{bmatrix}$
        $(\tilde{\phi}_1,\dots,\tilde{\phi}_{\lfloor M/2\rfloor}) = \text{eigs}\left\{\left[\mathbf{E}_2\right]\left[\mathbf{E}_1\right]^{\dagger}\right\}$
    end
end
```

Algorithm 2

```
$H_{\alpha,\delta} = \mathbf{0}_{\mathcal{A}\times\mathcal{D}}$
for $i = 1:1:\lfloor M/2\rfloor$ do
    for $a = \min_\alpha : (\max_\alpha - \min_\alpha)/\mathcal{A} : \max_\alpha$ do
        for $d = \min_\delta : (\max_\delta - \min_\delta)/\mathcal{D} : \max_\delta$ do
            if $|\tilde{\alpha}_i(\omega,\tau) - a| < (\max_\alpha - \min_\alpha)/2\mathcal{A}$ do
            if $|\tilde{\delta}_i(\omega,\tau) - d| < (\max_\delta - \min_\delta)/2\mathcal{D}$ do
            $H_{\alpha,\delta}(a,d) = H_{\alpha,\delta}(a,d) + |\tilde{S}_i(\omega,\tau)|^2$
        end
    end
end
```

Algorithm 3

$W(t)$ is chosen such that the class of source signals of interest satisfy the W-disjoint orthogonal assumption as much as possible, for speech $W(t)$ is chosen to be an $L = 30$-millisecond long Hamming window [4] and $\Delta = L/2T$.

Step 2. At each time-frequency point a simplified ESPRIT parameter estimation step (Section 2.1.1) is performed, the $\lfloor M/2\rfloor$ estimated mixing parameters are used to perform a demixing step at each time-frequency point via an inversion of the estimated mixing matrix and the Moore-Penrose pseudoinverse $[\cdot]^\dagger$ is used to invert nonsquare matrices (see Algorithm 2).

Step 3. At each time-frequency point and for $i = 1,2,\dots,\lfloor M/2\rfloor$ the relative attenuation and delay mixing parameter estimates are calculated:

$$\tilde{\alpha}_i(\omega,\tau) = \left|\tilde{\phi}_i(\omega,\tau)\right|, \tilde{\delta}_i(\omega,\tau) = -\frac{\operatorname{Im}\{\log_e\{\tilde{\phi}_i(\omega,\tau)\}\}}{\omega}, \qquad (82)$$

an $\mathcal{A}\times\mathcal{D}$ two-dimensional power weighted histogram $H_{\alpha,\delta}$ of the relative attenuation and delay parameters is also constructed (see Algorithm 3):

Step 4. The power weighted histogram $H_{\alpha,\delta}$ will have a number of peaks $N' \geq N$, each represents a signal received by the sensor array, in an echoic environment some of these signals may have originated from the same source. The centres of each of the peaks provide estimates of the mixing parameters $\hat{\alpha}_1,\hat{\delta}_1,\dots,\hat{\alpha}_{N'},\hat{\delta}_{N'}$. Peak detection may be performed using a suitable clustering technique.

Step 5. The permutation ambiguity associated with wideband implementations of narrowband techniques is overcome when each of the $\lfloor M/2\rfloor$ instantaneous source estimates $\tilde{S}_1(\omega,\tau),\dots,\tilde{S}_{\lfloor M/2\rfloor}(\omega,\tau)$ is correctly assigned to one of the $N' \geq N$ demixed estimates at each time-frequency point. Assignment is performed by determining which of the $\lfloor M/2\rfloor$ instantaneous parameter estimates $(\tilde{\alpha}_1(\omega,\tau),\tilde{\delta}_1(\omega,\tau)),\dots,(\tilde{\alpha}_{\lfloor M/2\rfloor}(\omega,\tau),\tilde{\delta}_{\lfloor M/2\rfloor}(\omega,\tau))$ is *closest* to each of the $N' \geq N$ peak centres $(\hat{\alpha}_1,\hat{\delta}_1),\dots,(\hat{\alpha}_{N'},\hat{\delta}_{N'})$. The measure of *closeness* of the ith estimate at (ω,τ) to the nth peak centre is given as

$$\left\{\left|\frac{\tilde{\alpha}_i(\omega,\tau)-\hat{\alpha}_n}{\mathcal{N}_\alpha}\right|^2 + \left|\frac{\tilde{\delta}_i(\omega,\tau)-\hat{\delta}_n}{\mathcal{N}_\delta}\right|^2\right\}, \qquad (83)$$

where $\mathcal{N}_\alpha$ and $\mathcal{N}_\delta$ are normalising factors. Beginning with the instantaneous mixing parameter estimates associated with the instantaneous source estimates of lowest power, at each time-frequency point the closest peak centre 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.

4. EXPERIMENTAL SIMULATIONS

In this section we present the results of experiments conducted on various synthetically generated mixtures and on real-room mixtures. These experiments were designed to demonstrate properties and advantages of the hard DESPRIT, soft DESPRIT, and echoic DESPRIT extensions to the DUET blind source separation algorithm.

4.1. Synthetic mixing experiments

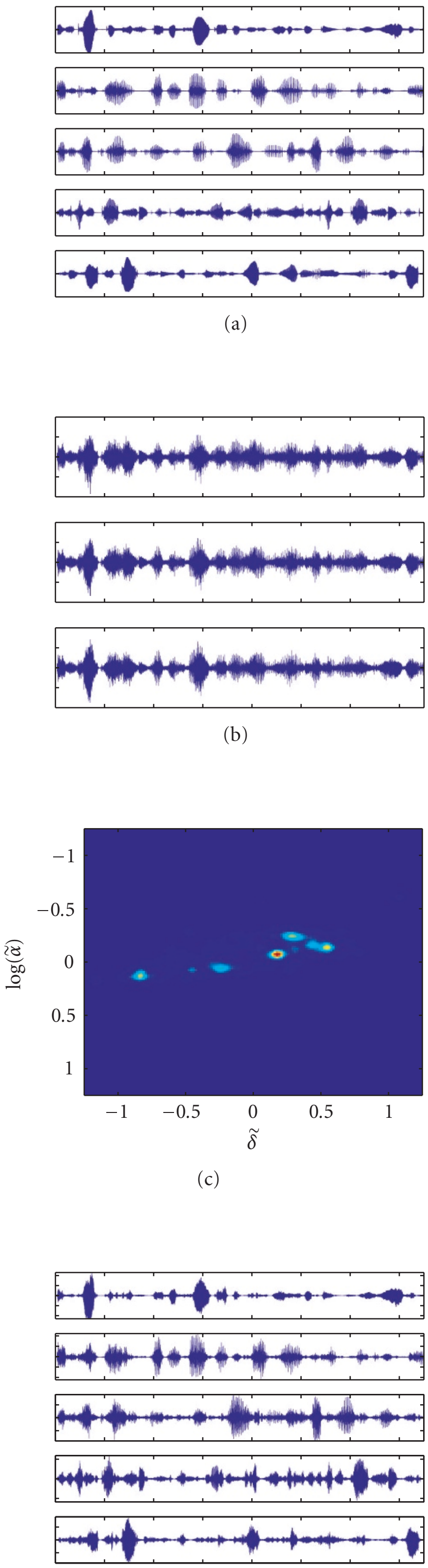
4.1.1. The hard DESPRIT extension

Five 3.75-second long speech signals (sampling frequency 16 kHz) taken from the TIMIT database were synthetically mixed in Matlab to create the three anechoic mixtures corresponding to the signals received by a three element uniform linear array with microphone spacing $D = 2\,\text{cm}$, the mixing parameters used were $\alpha_1, \ldots, \alpha_5 = (1.06, 0.78, 0.87, 1.15, 0.93)$ and $\delta_1, \ldots, \delta_5 = (0.24, 0.29, 0.5, -0.85, 0.17)$ samples.

The results of applying the hard DESPRIT algorithm to the mixtures are presented in Figure 3, as expected five peaks appear in the power weighted histogram at the mixing parameter locations. The mixtures were partitioned, aligned, and combined to produce the five demixtures using a maximum-likelihood approach

$$\widehat{S}_n(\omega, \tau) = \frac{\sum_{m=1}^{M} M_n(\omega, \tau) X_m(\omega, \tau) \left(\widetilde{\phi}^*(\omega, \tau)\right)^{m-1}}{\sum_{m=1}^{M} \left|\widetilde{\phi}(\omega, \tau)\right|^{2(m-1)}}, \quad (84)$$

where $\widehat{S}_n(\omega, \tau)$ is an estimate of the nth source, $M_n(\omega, \tau)$ is the nth binary time-frequency mask (i.e., $M_n(\omega, \tau)$ has value one for the time-frequency points whose associated mixing parameter estimates lie closest to the nth peak and zeros elsewhere), $X_m(\omega, \tau)$ is the mth mixture, and $\widetilde{\phi}(\omega, \tau)$ is the mixing parameter estimate obtained at the time-frequency point (ω, τ). This approach is the multichannel equivalent of [4, equation (53)]. The ability to blindly separate an arbitrary number of N sources from $M \geq 2$ anechoic mixtures is an ability of hard DESPRIT, soft DESPRIT, and echoic DESPRIT inherited from the original DUET algorithm.


(a)

(b)

(c)

(d)

Figure 3: Undetermined blind source separation via hard DESPRIT: (a) 5 speech sources, (b) 3 anechoic mixtures from a uniform linear array, (c) the 2-dimensional power weighted histogram shows 5 peaks from which (d) 5 demixtures are recovered.

4.1.2. The soft DESPRIT extension

Five 1.7-second long speech signals (sampling frequency 16 kHz) taken from the TIMIT database were synthetically mixed in Matlab to create anechoic mixtures corresponding to the signals received by a 2-, 3-, and 4-element uniform linear array with microphone spacing $D = 2\,\text{cm}$, the mixing parameters used were $\alpha_1, \ldots, \alpha_5 = (-0.45, 0.87, 0.32, -0.92, -0.11)$ and $\delta_1, \ldots, \delta_5 = (0.24, 0.29, 0.5, -0.85, 0.17)$ samples.

The soft DESPRIT algorithm was used blindly to demix five source signals from the 2, 3, and 4 anechoic mixtures of these signals. As with hard DESPRIT a two-dimensional mixing parameter estimation histogram was computed, unlike hard DESPRIT, where only a single parameter estimate available at each time-frequency point soft DESPRIT computes $M - 1$ eigenvalue estimates at each time-frequency point and uses these estimates to demix $M - 1$ signal estimates at each time-frequency point. Each of the $M - 1$ parameter estimates was weighted using the associated $M - 1$ signal power estimates to create a single histogram.

In Figure 4 we plot the parameter histograms for 2, 3, and 4 anechoic mixtures on the bottom row, in addition for illustrative purposes we plot the separate histograms associated with each of the eigenvalue estimates. The eigenvalues have been sorted from low to high powers where the powers are given by the associated instantaneous signal power estimates. The average percentage power associated with each histogram is given as a label to the histogram. If a strong WDO assumption was adopted, only the high-power histogram would be considered with any information available from lower-power histograms being disregarded,

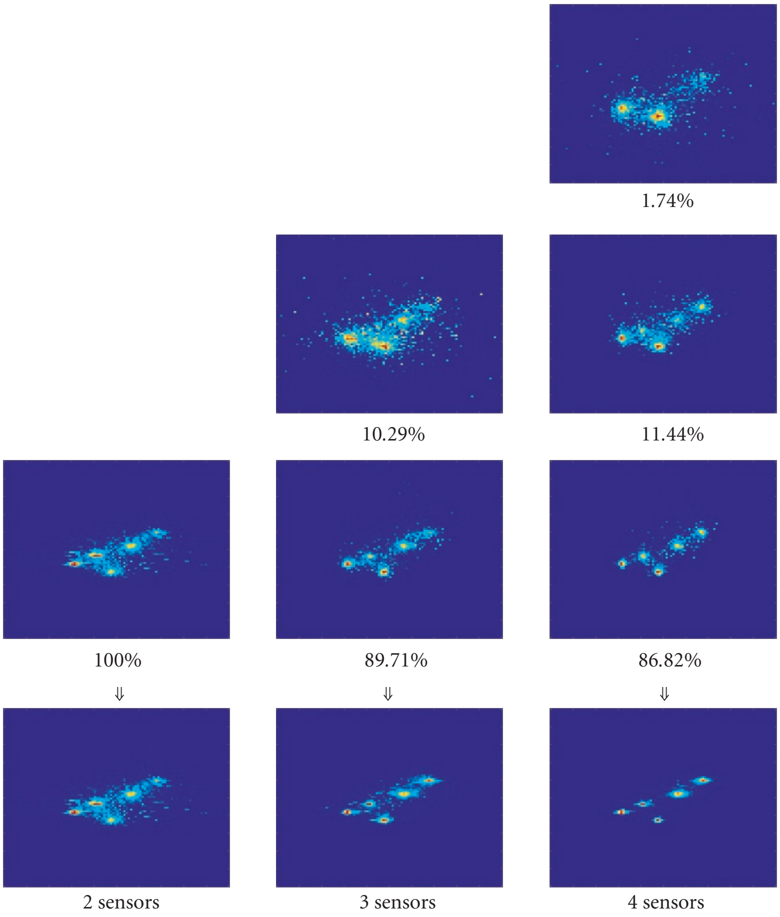


Figure 4: Soft DESPRIT histograms associated with the low to high power source estimates for 5 sources and 2, 3, and 4 anechoic mixtures. The average percentage power associated with each histogram is also given as a label to each component histogram. Each plot has an x-axis with units $-2.5 \leq \tilde{\delta} \leq 2.5$ samples and a y-axis with units $-2.5 \leq \log(\tilde{\alpha}) \leq 2.5$.

however upon examination it is evident that although the lower-power histograms are less clear they do possess information about peak locations, this observation motivates the soft DESPRIT algorithm.

It seems sensible to suggest that in general as the number of sensors increases the histograms become clearer, leading to more accurate source mixing parameter estimates. These plots certainly do show clearer histograms for more sensors but it can also be observed that at least in the case of 2–4 speech sources the first two eigenvalue estimates contain most of the power, this may suggest that increasing the number of sensors beyond $M = 3$ will not be as beneficial as increasing the number of sensors from DUET's original $M = 2$ to $M = 3$. The next section provides a quantitative description of these phenomena.

4.1.3. Hard DESPRIT versus soft DESPRIT

In an effort to quantify what we mean when we refer to a particular histogram being "clearer" and "more accurate" than another we define the following histogram peak measure:

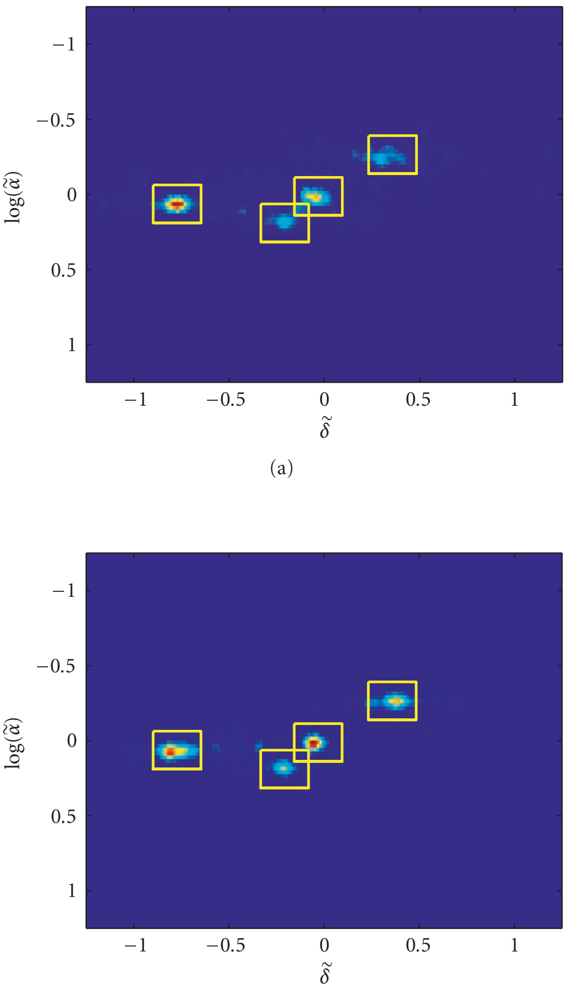
$$\mathcal{P}_{\alpha,\delta} \doteq \frac{\sum_{a,d} Q_{\alpha,\delta}(a,d) H_{\alpha,\delta}(a,d)}{\sum_{a,d} H_{\alpha,\delta}(a,d)}, \tag{85}$$

where

$$Q_{\alpha,\delta}(a,d) \doteq \begin{cases} 1, & |a-\alpha_n| \leq \epsilon_\alpha,\ |d-\delta_n| \leq \epsilon_\delta\ \forall\, n = 1,\ldots,N, \\ 0 & \text{otherwise,} \end{cases} \tag{86}$$

ϵ_α and ϵ_δ are used to define the boundaries of N square regions of the histogram centred upon the mixing parameter pairs $(\alpha_1,\delta_1),\ldots,(\alpha_N,\delta_N)$.

This is illustrated in Figure 5 where the $\mathcal{A} \times \mathcal{D} = 100 \times 200$ histograms for hard DESPRIT and soft DESPRIT are marked with 4 square regions with boundaries defined by $\epsilon_\alpha = 5$ bins and $\epsilon_\delta = 10$ bins and centred upon the mixing parameters $\alpha_1,\ldots,\alpha_4 = (1.08, 1.02, 0.77, 1.21)$



(a)

(b)

Figure 5: $\mathcal{A} \times \mathcal{D} = 100 \times 200$ histograms for hard DESPRIT (a) and soft DESPRIT (b) have 4 square regions with boundaries defined by $\epsilon_\alpha = 5$ and $\epsilon_\delta = 10$ and centred upon the original mixing parameters.

and $\delta_1, \ldots, \delta_4 = (-0.77, -0.04, 0.36, -0.20)$ samples. The peak measure $\mathcal{P}_{\alpha,\delta}$ gives the fraction of signal power contained within square regions defined by $Q_{\alpha,\delta}(a, d)$ compared with the total signal power contained within the histogram $H_{\alpha,\delta}(a, d)$. The peak measure gives an indication of how clear the histogram peaks are and whether or not they are obscured by unwanted noise. The clearer the histogram, the larger the peak measure and the more accurate the final mixing parameter estimates. Ideally a histogram will have $\mathcal{P}_{\alpha,\delta} = 1$ but in practice $\mathcal{P}_{\alpha,\delta} \leq 1$.

The measure was used to compare the histograms generated by the hard DESPRIT and soft DESPRIT algorithms, in Figure 6 we plot the values of $\mathcal{P}_{\alpha,\delta}$ for hard DESPRIT (solid curve) and for soft DESPRIT (dotted curve) when 2 $(\cdot)$, 3$(*)$, and 4 (o) sources were present. The red curve is for the no-noise case (SNR $= \infty$ dB) and the blue curve is in a noisy case (SNR $= 0$ dB). The curves were averaged over 100 trials, in each individual trial the mixing parameters were generated randomly and the sources were chosen randomly the TIMIT speech database.

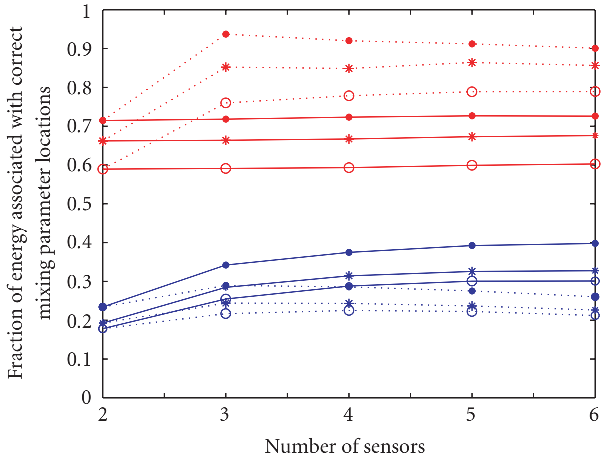


Figure 6: Comparison of the fraction of total histogram energy associated with correct mixing parameter locations for hard DESPRIT (solid) and soft DESPRIT (dotted) when 2$(\cdot)$, 3$(*)$, and 4 (o) sources are present and for signal-to-noise levels ∞ dB (red) and 0 dB (blue).

For high signal-to-noise values, for example, SNR $=$ ∞ dB (red curve), soft DESPRIT produces “clearer” and “more accurate” histograms in the sense of our peak measure (85) compared to hard DESPRIT. The benefit of increasing the number of sensors is evident in the case of soft DESPRIT with most benefit being gained from 3 sensors. This is consistent with Figure 4 where it may be observed that for $M = 4$ mixtures only 1.74% of the total signal power is associated with third eigenvalue estimate and so the first and second eigenvalue estimates make the most significant contribution to the power weighted histogram. Our peak measure suggests that at least in the case of 2–4 speech sources for $M \geq 3$ the first and second eigenvalue estimates are the most useful. There is little benefit in increasing the number of sensors for hard DESPRIT in this case since the peak measure stays relatively constant as the number of sensors increases.

For low signal-to-noise values, for example, SNR $= 0$ dB (blue curve), hard DESPRIT outperforms soft DESPRIT producing histograms with a higher peak measure. As the number of sensors increases, the effect of noise is alleviated and histograms with a higher peak measure are produced. The performance of soft DESPRIT disimproves slightly in this case as the number of sensors increases beyond $M = 3$, this is due to the eigenvalue estimates sensitivity to noise. In this case where 2–4 speech sources are present only the first and second eigenvalues provide useful mixing parameter estimates, the third, fourth, and fifth eigenvalues provide inaccurate estimates and result in a lower peak measure when they are used to compute the power weighted histogram.

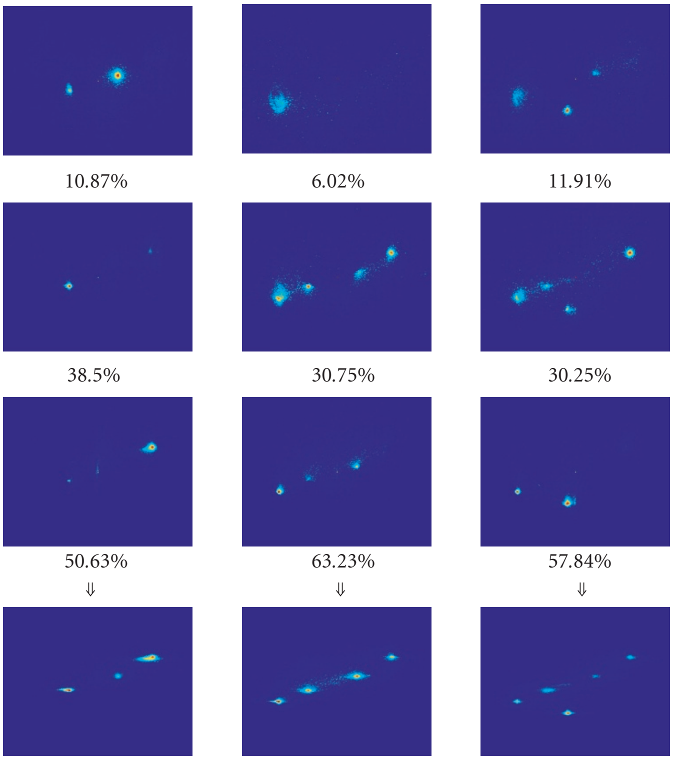


FIGURE 7: Histograms for 6 echoic mixtures of 1 source arriving on 3 paths (a), 2 sources arriving on 2 paths each (b), and 2 sources arriving on 2 paths and 3 paths, respectively, (c). The average percentage power associated with each histogram is also given as a label to each component histogram. Each plot has an x-axis with units $-2.5 \le \widetilde{\delta} \le 2.5$ samples and a y-axis with units $-2.5 \le \log(\widetilde{\alpha}) \le 2.5$.

4.1.4. *The echoic DESPRIT extension*

Three synthetic mixing experiments were performed demonstrating the properties of the echoic DESPRIT algorithm. Two 2.5-second long speech signals (sampling frequency 16 kHz) taken from the TIMIT database were synthetically mixed in Matlab to create six echoic mixtures corresponding to the signals received by a six element uniform linear array with microphone spacing $D = 2$ cm.

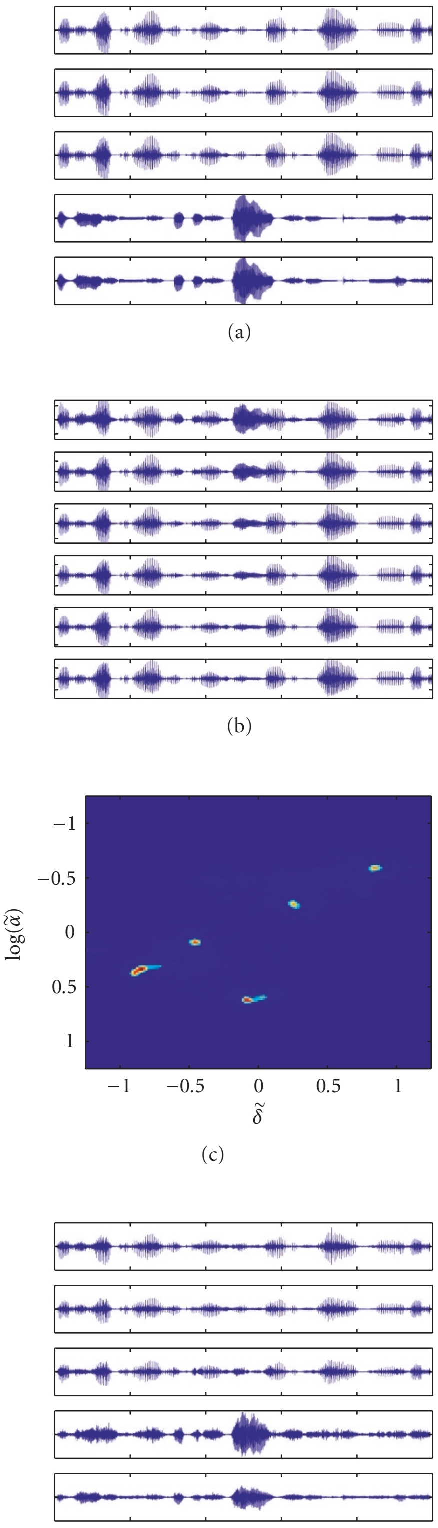
Experiment 1. The first signal was sent upon three paths with corresponding mixing parameters $\alpha_{1,1}, \alpha_{1,2}, \alpha_{1,3} = (1.09, 0.81, 0.55)$, $\delta_{1,1}, \delta_{1,2}, \delta_{1,3} = (-0.45, 0.33, 0.86)$ samples.

Experiment 2. The first signal was sent upon two paths with corresponding mixing parameters $\alpha_{1,1}, \alpha_{1,2} = (1.48, 1.09)$, $\delta_{1,1}, \delta_{1,2} = (-0.9, -0.45)$ samples and the second signal was sent upon two paths with corresponding mixing parameters $\alpha_{2,1}, \alpha_{2,2} = (0.81, 0.55)$, $\delta_{2,1}, \delta_{2,2} = (0.33, 0.86)$ samples.

Experiment 3. The first signal was sent upon three paths with corresponding mixing parameters $\alpha_{1,1}, \alpha_{1,2}, \alpha_{1,3} = (1.84, 1.48, 0.81)$, $\delta_{1,1}, \delta_{1,2}, \delta_{1,3} = (-0.09, -0.9, 0.33)$ samples and the second signal was sent upon two paths with corresponding mixing parameters $\alpha_{2,1}, \alpha_{2,2} = (1.09, 0.55)$, $\delta_{2,1}, \delta_{2,2} = (-0.45, 0.86)$ samples.

Echoic DESPRIT was applied to each of $M = 6$ echoic mixtures generated in Experiments 1, 2, and 3 and the corresponding parameter histograms are plotted on the bottom row of Figure 7. Again for illustrative purposes we have plotted the separate $\lfloor M/2 \rfloor = 3$ histograms associated with each of the eigenvalue estimates. The eigenvalues have been sorted from low to high powers, where the powers are given by the associated instantaneous signal power estimates. The average percentage power associated with each histogram is given as a label to the histogram, comparing with soft DESPRIT, the average percentage of instantaneous power weighting is more evenly spread amongst the component histograms. In Experiment 1 only one source is present, the average percentage power labels are in the same ratio as the square of the mixing parameters $\alpha_{1,1}, \alpha_{1,2}, \alpha_{1,3} = (1.09, 0.81, 0.55)$, since the single source is scaled by these attenuation factors. This may be considered consistent with the observation that each one of the three peaks dominates one of the individual histograms. The average power labels for Experiments 2 and 3 do not have the same interpretation available since two sources are present in these cases.

The results of applying the echoic DESPRIT algorithm in Experiment 3 are presented in Figure 8, as expected five peaks (corresponding to five signals arriving at the array) appear in the power weighted histogram at the mixing



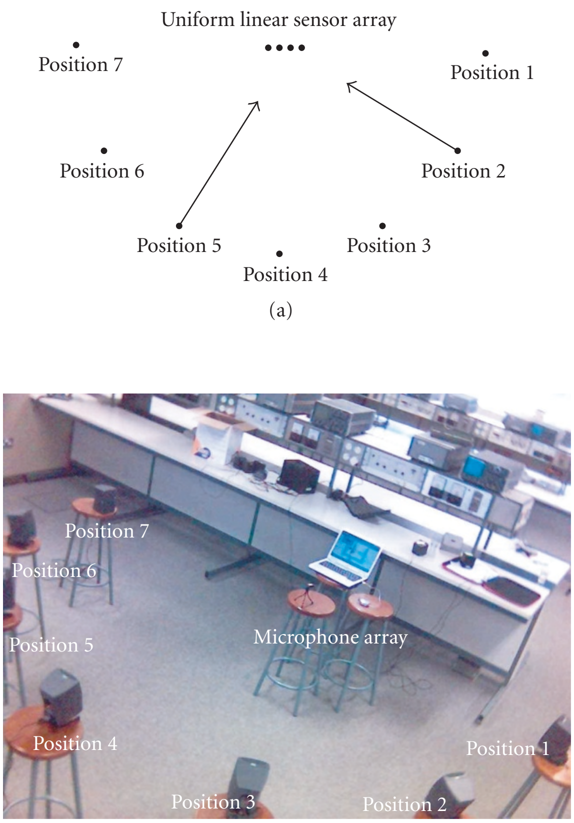
(a)

(b)

(c)

(d)

Figure 8: Underdetermined blind source separation in a simple echoic environment via echoic DESPRIT: (a) 2 speech sources travelling upon 3 and 2 paths, respectively, (b) 6 echoic mixtures of the 2 signals, (c) the two-dimensional power weighted histogram shows 5 peaks from which (d) 5 demixtures are recovered, 3 of which correspond to the first source and 2 of which correspond to the second source.



(a)

(b)

Figure 9: Real-room blind source separation experiment.

parameter locations and each of the 5 individual echoes is recovered as a scaled version of one of the original sources.

4.2. Real-room experiments

4.2.1. Impulse response measurement

As a demonstration of the usefulness of the echoic DESPRIT algorithm in a real world environment the following experiment was performed. As in Figure 9 seven speakers were positioned at locations on a semicircle and four microphones were positioned to form a uniform linear array with spacing of $D = 2.5\,\text{cm}$ located at the locus of the semicircle. A known white noise signal was played at each speaker position and then recorded at each of the 4 microphone positions, subsequently the impulse response between each speaker position and each microphone position was determined by deconvolving white noise signal from the recorded signal, that is,

$$\hat{a}_{m,n}(t) = \text{DFT}^{-1}\left\{\frac{\text{DFT}\{x_m(t)\}}{\text{DFT}\{s_n(t)\}}\right\}, \qquad (87)$$

where DFT$\{\cdot\}$ denotes the discrete Fourier transform, $\hat{a}_{m,n}(t)$ is the estimated impulse response between the nth

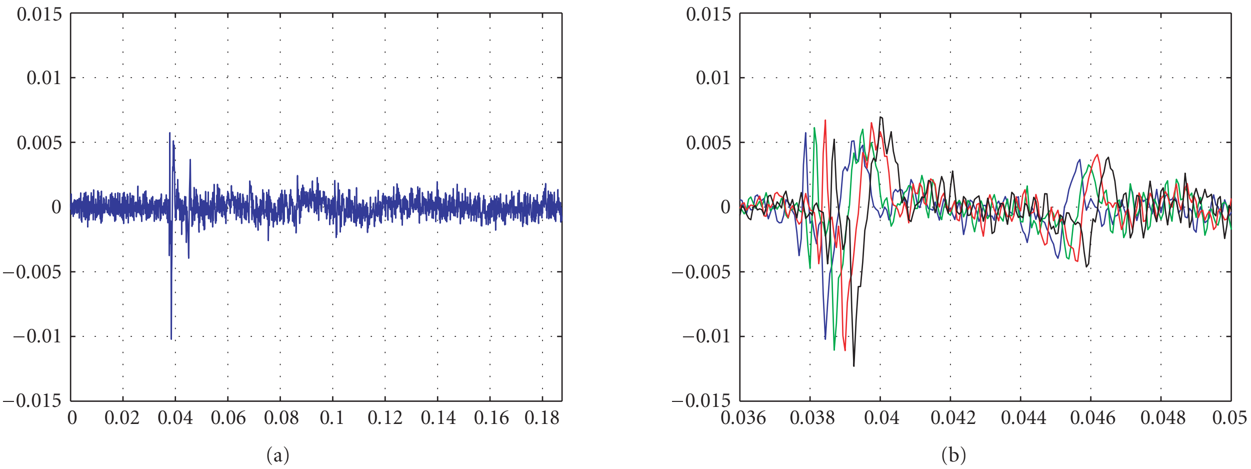


Figure 10: The measured impulse response in a real room between source position 1 and sensor 1 (a) and the measured impulse response in a real room between source position 1 and the 4 sensors of the microphone array (b).

source position and the mth sensor, $s_n(t)$ is the white noise signal played at the nth position, and $x_m(t)$ is its recording by the mth sensor. The top plot of Figure 10 shows the measured impulse response between the 1st speaker position and the 1st microphone of the sensor array. Although measurement noise dominates our impulse response estimate, there are clearly visible spikes indicating echoic paths and these spikes are few in number $P_1 \approx 5$. The bottom plot of Figure 10 zooms in and overlays the impulse responses for the 4 microphones, a superficial inspection may suggest that the impulse responses are merely delayed and attenuated versions of each other, indicating that an anechoic mixing model may be useful in this environment. Measurement of the reverberation time of the room using our impulse response measurements is a difficult task, we consider it to be approximately 10 milliseconds long. The speaker-microphone-room geometry in our experiment is such that the main propagation paths were direct, ceiling and floor. These propagation paths had the same relative attenuation and delay across the microphone array which was parallel to the floor and ceiling and so a near-anechoic mixing environment resulted. This geometry would not be atypical of large rooms such as lecture theatres, conference rooms, or laboratories such as the one used in our experiment.

4.2.2. *Real-room underdetermined blind source separation experiment*

Using the four measured room impulse responses corresponding to each of the source positions 1, 3, 4, 5, and 6, four real-room mixtures were created from five speech signals. The DUET, hard DESPRIT, soft DESPRIT, and echoic DESPRIT blind source separation algorithms were performed on the 4 real-room mixtures (2 of the 4 mixtures for DUET). The weighted histograms obtained are plotted in Figure 11, one-dimensional histograms are plotted because the peaks are located on or very near to the $\alpha = 1$ line. The plots are marked with ticks on the x-axis indicating the histogram peak locations obtained when only one source is located at the positions 1, 3, 4, 5, and 6. DUET produces many spurious peaks and only some may be considered to be in the correct location. Hard DESPRIT produces 5 distinct peaks near and around the correct locations. Soft DESPRIT produces 5 similar peaks but the peak near $\tilde{\delta} = 4.5$ samples dominates the histogram. Echoic DESPRIT produces 7 distinct peaks, the dominant 5 are localized around the correct peak locations and the other two lie between the 3rd, 4th, and 5th peaks and do not appear in the other histograms indicating that they may correspond to other propagation paths undetected by the other algorithms. Section 4.1.3 demonstrates the ability of hard DESPRIT to outperform soft DESPRIT in noisy environments, such as the one created using our impulse response measurements, this ability is evident when comparing the plots of Figure 11.

5. CONCLUSION

In this work, we explored possible extensions to the DUET blind source separation algorithm to the case when more than 2 mixtures are available. Three extensions were proposed, all of which combine the DUET method with the ESPRIT direction of arrival estimation technique. The first, called *hard DESPRIT*, is perhaps the natural extension of DUET to $M > 2$ mixtures and still assumes that only one source is active at any time-frequency point. The second, *soft DESPRIT*, allows for up to $M - 1$ sources to be active at any time-frequency point provided that no additional sources are active for the specific frequency over a window of $M - 1$ adjacent time points. The third, *echoic DESPRIT*, allows for the separation of echoic mixtures provided the mixing impulse response is sparse, each source signal travels on a small number (up to $\lfloor M/2 \rfloor$) of paths from source to sensor. In echoic ESPRIT, the constraint is that the number of source

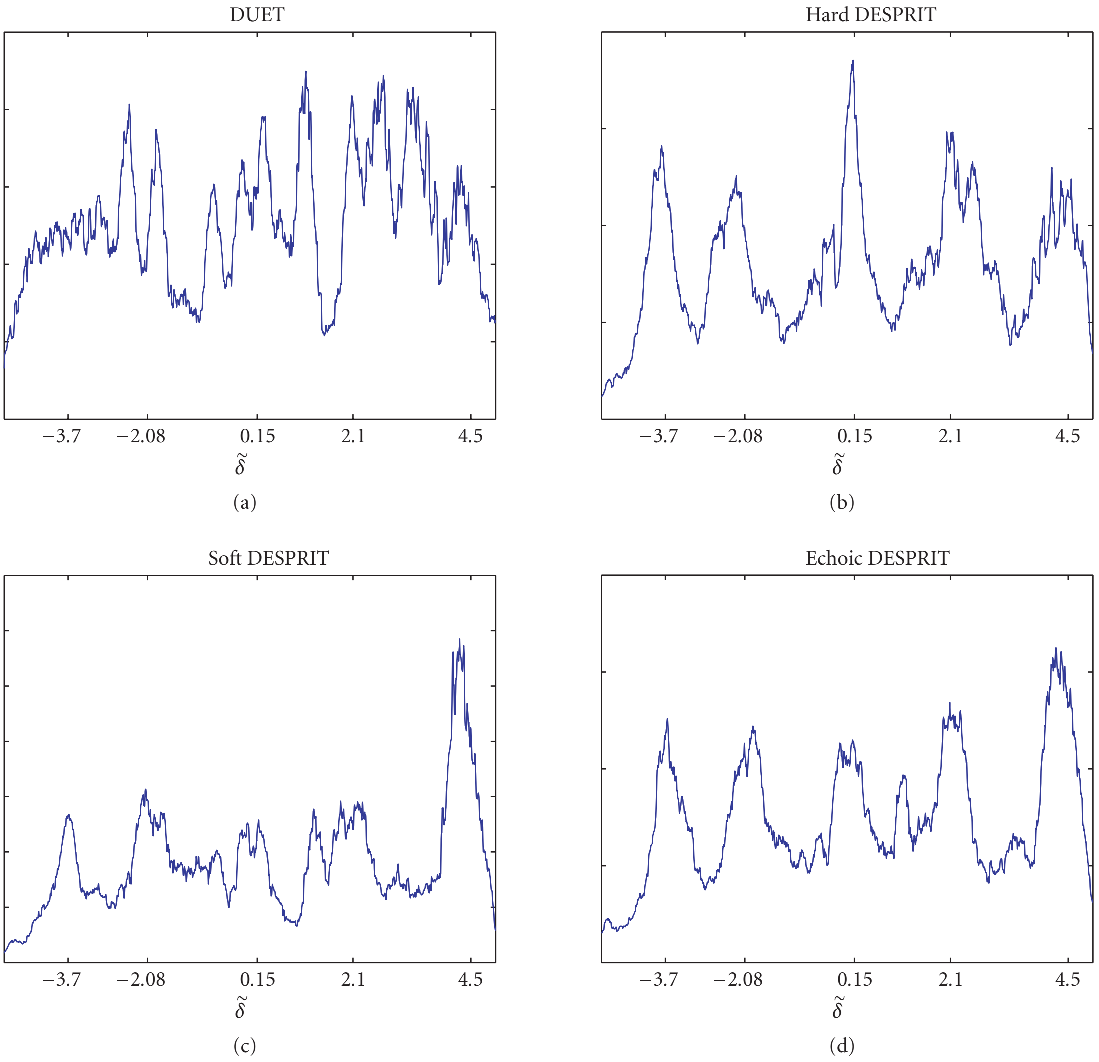


Figure 11: The histogram obtained by DUET algorithm when applied to 2 of the 4 real-room mixtures and the histograms obtained by the hard DESPRIT, soft DESPRIT, and echoic DESPRIT algorithms when they are applied to the 4 real-room mixtures.

arrivals active at a given time-frequency point cannot exceed $\lfloor M/2 \rfloor$. Results of tests on simulated and real-world mixtures demonstrate the capability of the extensions.

ACKNOWLEDGMENTS

This material is based upon work supported by an Irish Research Council for Science, Engineering and Technology (IRCSET) Grant and by Science Foundation Ireland under the PIYRA Programme.

REFERENCES

[1] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Anaylsis*, Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications and Control, John Wiley & Sons, New York, NY, USA, 2001.

[2] A. J. Bell and T. J. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[3] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley & Sons, New York, NY, USA, 2003.

[4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1846, 2004.

[5] T. Melia, S. Rickard, and C. Fearon, "Histogram-based blind source separation of more sources than sensors using a DUET-ESPRIT technique," in *Proceedings of the 13th European Signal Processing Conference (EUSIPCO '05)*, Antalya, Turkey, September 2005.

[6] S. Rickard, T. Melia, and C. Fearon, "DESPRIT - histogram based blind source separation of more sources than sensors using subspace methods," in *Proceedings of the IEEE Workshop on Applications of Signal Processing in Audio and Acoustics*, pp. 5–8, New Paltz, NY, USA, October 2005.

[7] T. Melia, S. Rickard, and C. Fearon, "Extending the DUET blind source separation technique," in *Proceedings of Signal Processing with Adaptative Sparse Structured Representations Workshop (SPARS '05)*, Rennes, France, November 2005.

[8] R. Balan, J. Rosca, and S. Rickard, "Scalable non-square blind source separation in the presence of noise," in *Proceedings of*

the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03), vol. 5, pp. 293–296, Hong Kong, April 2003.

[9] R. Balan and J. Rosca, "Sparse source separation using discrete prior models," in *Proceedings of Signal Processing with Adaptative Sparse Structured Representations Workshop (SPARS '05)*, Rennes, France, November 2005.

[10] Y. Li, S.-I. Amari, A. Cichocki, D. W. C. Ho, and S. Xie, "Underdetermined blind source separation based on sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 423–437, 2006.

[11] R. Saab, O. Yilmaz, M. McKeown, and R. Abugharbieb, "Underdetermined sparse blind source separation with delays," in *Proceedings of Signal Processing with Adaptative Sparse Structured Representations Workshop (SPARS '05)*, Rennes, France, November 2005.

[12] P. D. O'Grady and B. A. Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," in *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA '04)*, pp. 430–436, Granada, Spain, September 2004.

[13] F. Abrard and Y. Deville, "A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Processing*, vol. 85, no. 7, pp. 1389–1403, 2005.

[14] P. Georgiev, F. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 992–996, 2005.

[15] A. Blin, S. Araki, and S. Makino, "A sparseness-mixing matrix estimation (SMME) solving the underdetermined BSS for convolutive mixtures," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 4, pp. 85–88, Montreal, Quebec, Canada, May 2004.

[16] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," in *Proceedings of the International Workshop on Independence and Artificial Neural Networks*, Tenerife, Spain, February 1998.

[17] K. Torkkola, "Blind separation of convolved sources based on information maximisation," in *IEEE Workshop on Neural Networks and Signal Processing*, pp. 423–432, Kyoto, Japan, September 1996.

[18] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.

[19] 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 Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[20] T.-W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Processing Letters*, vol. 6, no. 4, pp. 87–90, 1999.

[21] Y. Li, A. Cichocki, and S.-I. Amari, "Analysis of sparse representation and blind source separation," *Neural Computation*, vol. 16, no. 6, pp. 1193–1234, 2004.

[22] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[23] 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, 1989.

[24] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, 1996.

[25] B. Ottersten, M. Viberg, and T. Kailath, "Performance analysis of the total least squares ESPRIT algorithm," *IEEE Transactions on Signal Processing*, vol. 39, no. 5, pp. 1122–1135, 1991.

[26] 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, no. 5, pp. 1232–1242, 1995.

Thomas Melia was born in Dublin, Ireland, in 1982. He received a Bachelor's degree in electronic engineering from the University College Dublin, Ireland, in 2003. He is currently a Ph.D. student with the Sparse Signal Processing Group at the University College Dublin, Ireland. His research interests include sparse signal processing and blind source separation.

Scott Rickard received the S.B. degree in mathematics in 1992, the S.B. degree in computer science and engineering in 1993, and the S.M. degree in electrical engineering and computer science, also in 1993, all from MIT. He received the M.A. and Ph.D. degrees in applied and computational mathematics from Princeton University, Princeton, NJ, in 2000 and 2003, respectively. He is currently a Senior Lecturer in the School of Electrical, Electronic, and Mechanical Engineering at the University College Dublin, Ireland. His research for the past several years has focused on the application of time-frequency methods and sparse signal processing for the blind separation of more sources than sensors.