

The logo for the Centre for Digital Music features a stylized banner with a green-to-blue gradient and vertical lines. The text "centre for digital music" is written in a white, lowercase, sans-serif font across the banner.

centre for digital music

An Adaptive Stereo Basis Method for Convolutional Blind Audio Source Separation

MARIA G. JAFARI, EMMANUEL VINCENT, SAMER A. ABDALLAH,
MARK D. PLUMBLEY AND MIKE E. DAVIES

Technical Report C4DM-TR-06-04
9 august 2006

An Adaptive Stereo Basis Method for Convolutional Blind Audio Source Separation

Maria G. Jafari¹, Emmanuel Vincent¹, Samer A. Abdallah¹, Mark D. Plumbley¹ and Mike E. Davies²

¹Centre for Digital Music
Queen Mary, University of London
Mile End Road – London E1 4NS – United Kingdom
mark.plumbley@elec.qmul.ac.uk

²Digital Communications Research Institute
University of Edinburgh
Mayfield Road – Edinburgh EH9 3JL – United Kingdom

Abstract: We consider the problem of convolutional blind source separation of stereo mixtures. This is often tackled using frequency-domain independent component analysis (FD-ICA), or time-frequency masking methods such as DUET. In these methods, the short-term Fourier transform (STFT) is used to transform the signal into the time-frequency domain. Instead of using a fixed time-frequency transform on each mixture channel, such as the STFT, we propose learning an adaptive transform from the stereo mixture pair. Many basis vector pairs of the resulting transform exhibit properties suggesting that they represent the components of individual sources, together with the filtering process from the sources to the microphone pair. A mask is then applied to the transformed signal, with the mask parameters determined by relative delays between the learned left and right basis vector pairs. The performance of the proposed adaptive stereo basis (ASB) algorithm is compared with FD-ICA and DUET under different reverberation and noise conditions, using both objective distortion measures and formal listening tests.

Keywords: blind source separation, audio source separation, independent component analysis, DUET algorithm, adaptive basis, sparse coding.

1 Introduction

Convolutional blind audio source separation is a problem that arises when an array of microphones records mixtures of sound sources that are convolved with the impulse response between each source and sensor.

Several methods have been proposed to tackle this problem, either in the time domain or in the frequency domain. Time domain methods mostly entail the extension of existing instantaneous blind source separation (BSS) algorithms to the convolutional case [1, 2, 3]. However, since they require the evaluation of convolutions, they can be computationally expensive [4].

An alternative and popular approach is the frequency domain independent independent component analysis (FD-ICA) method [4, 5, 6, 7, 8, 9]. This approach uses the short-time Fourier transform (STFT) to transform the convolved signal into the time-frequency domain, with instantaneous independent component analysis (ICA) performed separately in each frequency bin. This approach is typically simpler and computationally less complex than the time-domain approach, although it may require long STFT frames to successfully separate convolutionally mixed signals. There are other issues that need to be overcome, such as a tendency to flatten the estimated signal spectrum, due to a spectral shape ambiguity inherent in the problem. The use of separate ICA processes in each bin also introduces the well-known *permutation problem*, whereby the different frequency components of the signals become ‘swapped’ and require permutation to realign them. In common with other ICA-based separation methods, FD-ICA can only be used where the number of sources P to be separated is no more than the number of microphones Q .

Another approach that has been found to be successful in practical applications on stereo (two-microphone) anechoic mixtures is the degenerate unmixing estimation technique (DUET) [10, 11, 12]. Here the STFT is again used to transform the signal into the time-frequency domain, but then time-frequency masking is used to extract source components on the basis of which source dominates each time-frequency bin. Estimates of relative amplitude and delay between two microphones are used to identify the dominating source in each bin. In theory, DUET can perfectly separate sources that are *W-disjoint orthogonal*, *i.e.* with representations that are disjoint in the time-frequency domain. The use of time-frequency masking means that it is applicable to sources with more sources than microphones, and DUET has been successfully applied to separation of *e.g.* five speech source from stereo mixtures. However, performance of DUET has been observed to degrade with echoic mixtures, and large microphone spacing can also cause problems in estimating the relative delay used by the algorithm.

In this article, we propose an Adaptive Stereo Basis (ASB) source separation method for convolutional mixtures, based on the idea of masking in a transform domain. However, instead of using a fixed time-frequency transform such as the STFT, applied separately to each observation (microphone) channel, we learn an adaptive transform based on the observed stereo data that is applied to both channels together [13]. Many basis pairs of the resulting transform exhibit properties suggesting that they represent the components of individual sources, together with the filtering process from the sources to the microphone pair. We then calculate the relative time delays between left and right channels of the stereo basis pairs, corresponding to different directions of arrival (DOAs) of the sources, and use this information to separate the basis pairs into clusters, one for each source. This then gives us a time-invariant association of each source with a subset of the stereo basis pairs, allowing us to estimate the separated sources.

The structure of this paper is as follows: the convolutional BSS problem is described in Section 2, with the FD-ICA and DUET algorithms described in Section 3. Our proposed adaptive stereo basis method is introduced in section 4. The performance of the algorithm is evaluated in Section 5, followed by discussion and conclusions.

2 Convolutional Blind Source Separation

Consider the problem of linear convolutional mixing, for example microphones recording mixed sound sources in a room with delays and echoes. Here each microphone records a linear combination of the source signals s_p , at several times and levels, as well as multipath copies (echoes) of the sources. This scenario can be

modelled as a finite impulse response (FIR) convolutive mixture, given by [4]

$$x_q(n) = \sum_{p=1}^P \sum_{l=0}^{L_m-1} a_{qp}(l) s_p(n-l), \quad q = 1, \dots, Q \quad (1)$$

where $x_q(n)$ is the signal recorded at the q -th microphone at time sample n , $s_p(n)$ is the p -th source signal, $a_{qp}(l)$ denotes the impulse response of the mixing filter from source p to sensor q , and L_m is the maximum length of all impulse responses [14]. The aim of convolutive blind source separation is to estimate the original source signals $s_p(n)$ and the mixing process $a_{qp}(n)$ given only the mixtures $x_q(n)$.

2.1 Spectral shape ambiguity

We wish to estimate the original source $s_p(n)$ in (1), given only the observations $x_q(n)$. However, there is an inherent ambiguity in this problem, called the *spectral shape ambiguity*, as we can see if we transform (1) into the continuous frequency domain

$$x_q(\omega) = \sum_{p=1}^P a_{qp}(\omega) s_p(\omega) \quad (2)$$

where $x_q(\omega)$, $a_{qp}(\omega)$ and $s_p(\omega)$ are the continuous Fourier transforms of the respective quantities in (1). But we can also write

$$x_q(\omega) = \sum_{p=1}^P a'_{qp}(\omega) s'_p(\omega) = \sum_{p=1}^P a_{qp}(\omega) s_p(\omega) \quad (3)$$

where $a'_{qp}(\omega) = a_{qp}(\omega) g_p(\omega)$ and $s'_p(\omega) = s_p(\omega) / g_p(\omega)$. Unless some other information is used, the source signals s_p can only be identified up to some arbitrary filtering g_p [9].

The spectral shape ambiguity is probably most clearly demonstrated for FD-ICA methods, whereby the use of standard pre-whitened ICA for each frequency bin typically leads to source estimates that are whitened, *i.e.* have the same signal power in each frequency bin. This ambiguity can also be a problem for other methods, although its effect may be reduced if specific signal models are used, or if assumptions on the filter order are employed. Nevertheless, we can sidestep the spectral shape ambiguity entirely by mapping the estimated sources back to the observation (microphone) space [9]. Suppose that a blind source separation algorithm produces \hat{s}_p as the estimate of the p th source signal and \hat{a}_{qp} as the estimate of the mixing process. Then the *image* \hat{x}_{qp} of the p th source signal at the q th microphone is given in the frequency domain by

$$\hat{x}_{qp}(\omega) = \hat{a}_{qp}(\omega) \hat{s}_p(\omega) \quad (4)$$

where we note that the right hand side is *not* summed over p . In the time domain this is

$$\hat{x}_{qp}(n) = \sum_{l=0}^{L_m-1} \hat{a}_{qp}(l) \hat{s}_p(n-l). \quad (5)$$

For each of the P sources, Equation (4) results in Q estimates, one at each microphone corresponding to what that source would sound like at that microphone. It is straightforward to verify that the filtering ambiguity $g_p(\omega)$ cancels out in Equation (4), so the spectral shape ambiguity does not arise in the source images $\hat{x}_{qp}(\omega)$ [9].

In this article, we use source image estimates \hat{x}_{qp} to avoid the spectral shape ambiguity while comparing the performance of the algorithms.

3 Algorithms for Convolutive BSS

3.1 Time-Domain Approach

The convolutive source separation problem can be approached directly in the time domain, by searching for a set of deconvolving filters $w_{pq}(k)$ to produce an output

$$y_p(n) = \sum_{q=1}^Q \sum_k w_{pq}(k) x_q(n-k) \quad (6)$$

where the filter coefficients $w_{pq}(k)$ are adapted to optimize certain statistics of the outputs $y_p(n)$ [1, 2, 3].

However, the frequency domain interpretation (2) suggests that with a transformation into the frequency domain we could replace a complex deconvolution problem with many parallel source separation problems, each at each different frequency [4].

3.2 Frequency-domain ICA

The mixing model (2) in the continuous frequency domain suggests that we could tackle the convolutive blind source separation problem by searching for a suitable unmixing filter $w_{pq}(\omega)$ yielding

$$y_p(\omega) = \sum_{q=1}^Q w_{pq}(\omega) x_q(\omega) \quad (7)$$

where $y_p(\omega)$ is our estimate of the source $s_p(\omega)$, $p = 1, \dots, P$ and $P = Q$. However, to use ICA on the frequency components in an FD-ICA algorithm we need to gather statistics about the mixtures x_q . We therefore divide the input sequence into frames, and use the short-time Fourier transform (STFT)

$$x_q(f, t) = \sum_{m=0}^{N-1} x_q(m+t) \gamma(m) e^{-j2\pi f m} \quad (8)$$

where N is the frame length, $x_q(n)$ is the time-domain signal sampled at frequency f_s , $\gamma(m)$ is a window function, that typically decays smoothly to zero at each end, t is the STFT block index, and $f \in \{0, \frac{f_s}{N}, \dots, \frac{(N-1)f_s}{N}\}$ denotes the frequency bin.

In this time-frequency domain (8), the mixing model in (1) is reasonably approximated in matrix notation by

$$\mathbf{x}(f, t) = \mathbf{A}(f) \mathbf{s}(f, t) \quad (9)$$

where $\mathbf{A}(f)$ and $\mathbf{s}(f, t)$ are the time-frequency domain representations of the mixing filters and the original sources respectively. The separating model then becomes

$$\mathbf{y}(f, t) = \mathbf{W}(f) \mathbf{x}(f, t) \quad (10)$$

where $\mathbf{y}(f, t)$ are the recovered source estimates in the frequency domain, and $\mathbf{W}(f)$ are the separating filters to be estimated. The convolutive BSS problem is thus transformed into multiple complex valued ICA problems in the time-frequency domain, with a suitable ICA algorithm (*e.g.* [15, 16, 17]) used to estimate $\mathbf{W}(f)$ separately in each frequency bin.

Once we have the separated source estimates, we can calculate the source images $\hat{x}_{qp}(f, t)$ as suggested by (4), using the estimate $\hat{\mathbf{A}}(f) = \mathbf{W}^{-1}(f)$ for the mixing process.

3.2.1 The permutation problem

The use of separate ICA algorithms for each frequency bin f in (10) leads to the well-known *permutation problem*. Due to the inherent ambiguity in the identification of the sources, any ICA algorithm can only find a set of original sources relative to some unknown permutation. Since these are applied independently

to each frequency bin, a further process is required to match the source estimates $\mathbf{y}(f, t)$ at a particular frequency bin f with those at other frequency bins.

Several different methods have been proposed to perform this source matching process across frequency bins. For example, Smaragdis [5] proposed a frequency coupling between the separating matrices for adjacent frequency bins; Parra [7] proposed a constraint on the length of the separating filter; Ikeda and Murata [6] suggested that for a particular source, the time envelopes across the frequency bins are similar, and therefore they are matched in order to align the permutations; while Davies [8] proposed a time-frequency source model which couples the frequency bins by measuring the signal envelope along the frequencies.

3.2.2 Solving the permutation problem using beamforming

An alternative approach to solving the permutation problem is to consider the spatial arrangement of the source and microphones: a *beamforming* approach [18]. If most of the signal observed at the microphones arrives from the direction of the direct path from the source, the time delay between the microphones will correspond to the *direction of arrival* (DOA) of the source. The permutation problem is addressed using this beamforming approach by Kurita et al [19, 20], who use an FD-ICA method to separate the sources at each frequency bin, and then permute these source estimates so that their DOAs are aligned.

To ensure the direction of arrival calculation is unique, the inter-microphone spacing must satisfy $d < \lambda_{\min}/2 = c/(2f_{\max})$ so that there is less than one wavelength difference between two sources with DOA of $+\pi/2$ and $-\pi/2$ from the equal-delay direction. For example, with $f_{\max} = 8$ kHz and $c = 340$ m/s we get $d \leq (340/16000)\text{m} \approx 2.1$ cm [14]. If uniqueness is not satisfied, for example when the microphone spacing is too large (*e.g.* $d \approx 1$ m), then several DOAs may correspond to a given delay, and permutation errors may occur. This problem is known as *spatial aliasing*.

To avoid the spatial aliasing problem, we can perform DOA estimation using only the lower band of frequencies $f < f_{\max}$ [21] and use another method to solve the permutation problem for higher frequencies. Alternatively, Mitianoudis and Davies [22] proposed the use of a ‘peakier’ directivity pattern method based on the MuSIC algorithm [23]. While the spatial ambiguity problem still exists, the more ‘spiky’ directivity patterns mean that the DOAs can be usefully aligned over a wider frequency range $0 \leq f < 12f_{\max}$ [22]. We will use this MuSIC-based method in our comparative evaluation later.

3.3 DUET algorithm

In FD-ICA the STFT was used to transform the mixture signal into the time-frequency domain to approximate the convolutive mixing process (1) by a set of parallel instantaneous narrowband mixing processes (9). A side-effect of the STFT is that many signals are *sparse* in the time-frequency domain: *i.e.* signals are zero or very small more often than it might be expected from their variances [24]. It has been noted that many ICA algorithms have improved performance when sources are sparse [25].

In addition, this sparsity property also means that in most time-frequency bins, all but at most one source will have a time-frequency coefficient $s_p(f, t)$ of zero or close to zero. So rather than performing ICA at each frequency bin f , we could instead attempt to simply identify the dominating source in each time-frequency bin (f, t) and separate by *time-frequency masking*. This is the essence of the degenerate unmixing estimation technique (DUET), which has the additional advantage that it can be applied to the underdetermined case, where there are more sources than mixtures [10].

In the DUET algorithm we assume that in the TF domain each time-frequency point of a mixture signal is due only to one of the sources, a property denoted as *W-disjoint orthogonality* [24]. Two sources s_l and s_m are W-disjoint orthogonal if the supports of the STFT representations of the sources are disjoint, *i.e.* [10]

$$s_l(f, t)s_m(f, t) = 0, \quad \forall l \neq m, \quad \forall t, f. \quad (11)$$

Therefore in place of the permutation problem present in FD-ICA, in the DUET algorithm we have to identify which source is present in each time-frequency point.

DUET assumes an anechoic mixing model with $Q = 2$ microphones:

$$x_q(n) = \sum_{p=1}^P \alpha_{qp} s_p(n - \delta_{qp}), \quad q = 1, 2 \quad (12)$$

where α_{qp} and δ_{qp} are the attenuation and time delay coefficients associated with the path between the p -th source and the q -th sensor [12]. In the time-frequency domain, equation (12) can be written as in (9) but where $\mathbf{A}(f)$ will take a special form since it will represent a set of 1-tap filters each with a scaling and single time delay. Due to the spectral shape ambiguity (and scaling ambiguity) we can absorb the filtering on the first mixture channel into the sources. Setting the parameters of the first mixture ($q = 1$) to $\alpha_{qp} = 1$ and $\delta_{qp} = 0$ for all $p = 1, 2$, we drop the subscript $q = 2$ for the remaining parameters, writing α_p and δ_p in place of α_{2p} and δ_{2p} respectively. The mixing matrix, $\mathbf{A}(f)$, is then of the form [10]

$$\mathbf{A}(f) = \begin{pmatrix} 1 & \dots & 1 \\ \alpha_1 e^{-j2\pi f \delta_1} & \dots & \alpha_P e^{-j2\pi f \delta_P} \end{pmatrix}. \quad (13)$$

To identify which source is present in each time-frequency bin (f, t) , the relative amplitude and delay between the left and right channels for each frequency bin are calculated. These are then clustered, so that time-frequency bins within the same cluster correspond to the same source.

Suppose that one source dominates in a given time-frequency bin (f, t) . Then from Equations (9), (11), and (13) the relative amplitude and delay parameters for the source active in that time-frequency bin can be estimated by [12]

$$\hat{\alpha}(f, t) = \frac{|x_2(f, t)|}{|x_1(f, t)|} \quad (14)$$

$$\hat{\delta}(f, t) = -\frac{1}{2\pi f} \angle \left(\frac{x_2(f, t)}{x_1(f, t)} \right) \quad (15)$$

where $\angle(\cdot)$ denotes the phase of a complex number taken between $-\pi$ and π .

In evaluating the delay parameter δ the DUET algorithm suffers from the same spatial ambiguity experienced in frequency domain ICA (Section 3.2.2). To avoid this, we normally assume that the sensor spacing is small enough not to introduce these ambiguities.

Following the computation of $\hat{\alpha}(f, t)$ and $\hat{\delta}(f, t)$, these estimated mixing parameters are used to produce a weighted two-dimensional histogram. The number of histogram peaks are used to estimate the number P of source signals present in the mixture, with the relative amplitude $\hat{\alpha}_p$ and delay $\hat{\delta}_p$ of the p th estimated source given by the values of the mixing parameters at the peak of the p th cluster in the histogram.

Once the histogram peaks have been found, a set of binary time-frequency masks $M_p(f, t)$, $p = 1, \dots, P$ is then constructed to perform separation of the sources s_p . To build the mask, each time-frequency point in proximity of a peak is assigned to the source corresponding to that peak. The time-frequency representation of the p -th estimated source is then constructed from the masked observations, either masked from one observation channel x_1 , or remixed from both channels. Finally, the time-domain source estimates are obtained by inverting the STFT, for example using the overlap-add method.

In theory, perfect separation can be achieved with these masks, provided the sources do not overlap in the TF domain, *i.e.* that W-disjoint orthogonality holds [12]. Empirical evidence is presented in [12] that speech signals are approximately pairwise W-disjoint orthogonal.

3.3.1 Extending DUET for source images

There are several ways to extend DUET to produce source image estimates at the microphones, to allow us to measure separation performance on source images at the microphones (Equation (5)). For example, it would be possible to use maximum likelihood (ML) source estimation [12] and create the source image using (4). However, we have observed that for echoic convolutive mixtures this can produce poor results, apparently due to inaccurate estimates of the mixing delays δ_p .

For the evaluation in this article, we directly calculate the image $\hat{x}_{qp}(f, t)$ of the p -th estimated source observed at the q -th microphone using

$$\hat{x}_{qp}(f, t) = M_p(f, t)x_q(f, t), \quad \forall f, t. \quad (16)$$

The time-domain estimate $\hat{x}_{qp}(n)$ is obtained by inverting the STFT for each source/microphone pair. Conceptually this approach uses DUET time-frequency masking to directly calculate an estimate of the image x_{qp} of source s_p at the q th microphone, without calculating a single source estimate as an intermediate stage.

4 Adaptive Stereo Basis method

Both the FD-ICA and DUET methods use the STFT separately for each microphone to transform the time-domain signal into a time-frequency representation. In the case of FD-ICA, the STFT is used to approximate convolution in time by multiplication in frequency, allowing separation via parallel ICA algorithms at each frequency. In the case of DUET, the STFT is used to transform the signal into a representation where the signals are approximately disjoint, allowing separation via binary masking in the time-frequency domain, although other transforms are possible [12].

The method that we proposed is based on the search for a transform that will directly allow us to partition the transform components into subsets corresponding to each source. If we could achieve this with the single-channel STFT, this would be a simple filtering operation, assigning frequency bands (subsets of frequency bins) to each source. However, since the sources we are considering do not occupy disjoint frequency bands, we use an *adaptive* transform.

In fact, we can use ICA to learn such an adaptive transform, but instead of using it across mixtures to separate sources, we use it across time samples to search for interesting structure in the data. In an early application of this method, Bell and Sejnowski [26] found that ICA trained on time-frames of monophonic recordings of ‘tooth taps’ discovered features (basis vectors) exhibiting localized time and phase structure, while those learned by *e.g.* principal components analysis (PCA) did not. Other studies on monophonic audio signals have reported that the basis vectors learned by ICA from speech signals are mostly well localised in time and frequency, yielding a representation that exhibits wavelet-like bases [27, 28]. The resulting representation of the sounds transformed into this learned basis are sparse, *i.e.* with most coefficients close to zero, giving a representation reminiscent of that of auditory nerve fibres [28].

In a preliminary study [13], we investigated an extension of this technique to stereo signals, applying an ICA algorithm to sequences of stereo time frames. We found that many of the resulting basis vectors typically exhibited the wavelet-like localized time and frequency representation as for the monophonic case. However, while the frequency representation of a typical basis vector is *localized* around a particular centre frequency, it is not *narrowband* as is the case for STFT basis vectors, and a time-domain centre is normally observed. Furthermore, many bases also displayed relative amplitude differences and time delays between the two channels, suggesting that the basis vectors discovered by the algorithm represent the components of individual sources and the filtering process from the sources to each of the microphones. If this is the case, then by partitioning these bases into subsets corresponding to each of the sources, it should be possible to separate the original source signals from each other. This is the principle behind the proposed Adaptive Stereo Basis (ASB) method.

4.1 Learning the stereo basis set

To learn the stereo basis set, the observed vector sequence $\mathbf{x}(n)$ is first reshaped into a $K \times k_{\max}$ matrix on which learning is performed. Successive frames of $K/2$ samples is taken from each mixture, with an overlap of T samples. Thus, the (i, k) -th element of the new matrix, $\tilde{\mathbf{X}}$, is

$$[\tilde{\mathbf{X}}]_{i,k} = \begin{cases} x_1((k-1)Z + (i+1)/2) & : i \text{ odd} \\ x_2((k-1)Z + i/2) & : i \text{ even} \end{cases} \quad (17)$$

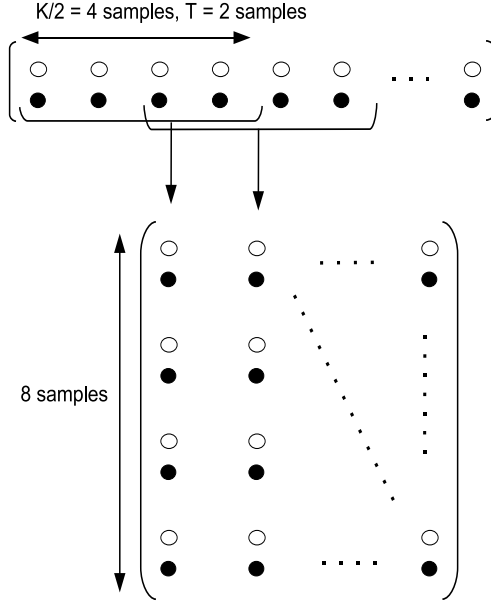


Figure 1: Reshaping of the sensor vector prior to training with ICA. In this illustration, we have $K/2 = 4$ sample pairs per frame, with an overlap of $T = 2$ samples.

where $Z = K/2 - T$, and $i \in \{1, \dots, K/2\}$, and $k \in \{1, \dots, k_{\max}\}$. The reshaping of the sensor vector $\mathbf{x}(n)$ is illustrated in figure 1.

For each column $\tilde{\mathbf{x}}$ of $\tilde{\mathbf{X}}$ we construct the representation coefficients $\mathbf{u} = \mathbf{B}\tilde{\mathbf{x}}$. The coding matrix $\mathbf{B} \in \mathbb{R}^{K \times K}$ is learned using an ICA algorithm, such as [13]

$$\Delta \mathbf{B} = \eta (\mathbf{I} - E\{\mathbf{f}(\mathbf{u})\mathbf{u}^T\}) \mathbf{B} \quad (18)$$

where η is the learning rate, and $\mathbf{f}(\mathbf{u}) = -\nabla_{\mathbf{u}} \log p(\mathbf{u})$ is the activation function, using $p(\mathbf{u}) = \prod_{p=1}^P p(u_p)$ for some prior $p(c_p)$. We use the generalized exponential prior [29] $p(c_p) \propto \exp(-|c_p|^\alpha)$ where the exponent α is estimated through maximum likelihood.

The reshaping of $\mathbf{x}(n)$ into the matrix $\tilde{\mathbf{X}}$ emphasises the correlations between the sources at the two microphones. Stacking the columns of $\mathbf{x}(n)$ allows features relating to temporally correlated signals from each recording to be extracted. Basis vector pairs are then extracted from the columns of the inverse matrix \mathbf{B}^{-1} .

Figure 2 shows some of the feature vector (*basis vector*) pairs obtained from two mixtures generated when two male speech signals were synthetically mixed using a source image technique, in low noise and low reverberation conditions (see Section 5).

This figure illustrates that the basis vector pairs encode how the extracted features are received at the microphones. Many of the basis vectors are localised in time, and they seem to capture information about time-delay and amplitude differences that characterise the mixing channel. This observation, together with measurements of the relative time delay (see Fig. 3 below), suggests that the convolutive nature of the mixing process has been captured by the algorithm, and that each basis vector pair relates to a particular source.

4.2 Clustering using time delays

Having identified the basis vectors, we perform clustering to identify the subspaces corresponding to the original sources. For each basis pair k we find the time delay τ_k between the vectors in the pair. The time

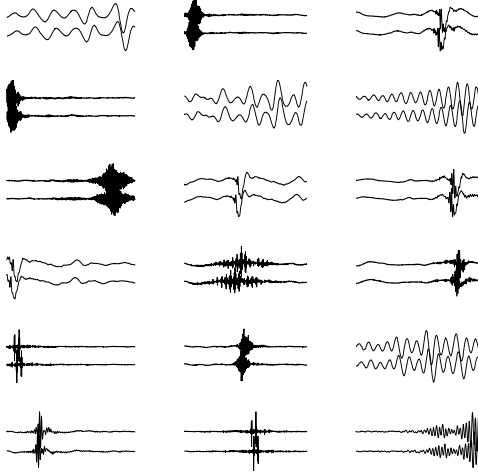


Figure 2: Examples of basis vectors extracted with the adaptive stereo basis algorithm.

delay τ_k is evaluated using the well known generalised cross-correlation with phase transform (GCC-PHAT) algorithm [30]

$$R_{a_1 a_2}(\tau) = \int_{-\infty}^{\infty} A_1(\omega) A_2^*(\omega) / (|A_1(\omega) A_2^*(\omega)|) e^{j\omega\tau} d\omega \quad (19)$$

where $A_1(\omega)$, $A_2(\omega)$ are the Fourier transforms of the basis vectors, which are taken from the columns of $\mathbf{A} = \mathbf{B}^{-1}$. Since GCC-PHAT considers all frequency bins together, we reduce the problem of phase ambiguities (spatial aliasing) that can occur with the FD-ICA and DUET algorithms. We have observed that the function $R_{a_1 a_2}(\tau)$ typically exhibits a sharp peak at the lag corresponding to the time delay between the two signals.

The upper plot in Figure 3 depicts the time-delay estimates obtained with GCC-PHAT, for all basis vector pairs shown in figure 2. The histogram of the estimated time-delays is shown in the lower plot of Figure 3. The figure shows that the directions of the two sources (corresponding to a delay of about 9 and -9 samples) are correctly identified, and most basis functions are associated with one of the directions of arrival.

Once we have the delays τ_k for each basis vector k , we then construct the histogram of τ_k across the different values of k , and use the K-means clustering algorithm to find the peaks T_p , $p = 1, \dots, P$ corresponding to each of the P sources.

We construct a set of mask matrices $\mathbf{H}^{(p)} = \text{diag}(h_1^{(p)}, \dots, h_K^{(p)})$ for $p = 1, \dots, P$, with the mask values given by

$$h_k^{(p)} = \begin{cases} 1 & \text{if } (T_p - \Delta) \leq \tau_k \leq (T_p + \Delta) \\ 0 & \text{otherwise} \end{cases}$$

for $p = 1, \dots, P$, $k = 1, \dots, K$, and where the threshold Δ determines the tolerance allowed on the delay estimation for a basis vector to be considered to represent part of a particular source. Thus the diagonal elements of $\mathbf{H}^{(p)}$ are one or zero depending on whether or not a transform component belongs to the cluster corresponding to the p -th source.

Finally the the image \hat{x}_{qp} of the p -th source estimate at the q -th microphone, is calculated using

$$\hat{\mathbf{x}}_p = \mathbf{B}^{-1} \mathbf{H}^{(p)} \mathbf{u} \quad (20)$$

where $\hat{\mathbf{x}}_p = [\hat{x}_{1p}, \hat{x}_{2p}, \dots, \hat{x}_{Qp}]^T$ is the vector of images of the p -th source at all Q microphones and $\mathbf{H}^{(p)}$ is the p -th diagonal masking matrix given above. Note that, in contrast to the mask $M_p(f, t)$ used in the DUET algorithm, which depends both on the frequency bin index f and the time frame index t , the ASB masking matrix $\mathbf{H}^{(p)}$ operates across basis pair indices k only and is independent of the time frame.

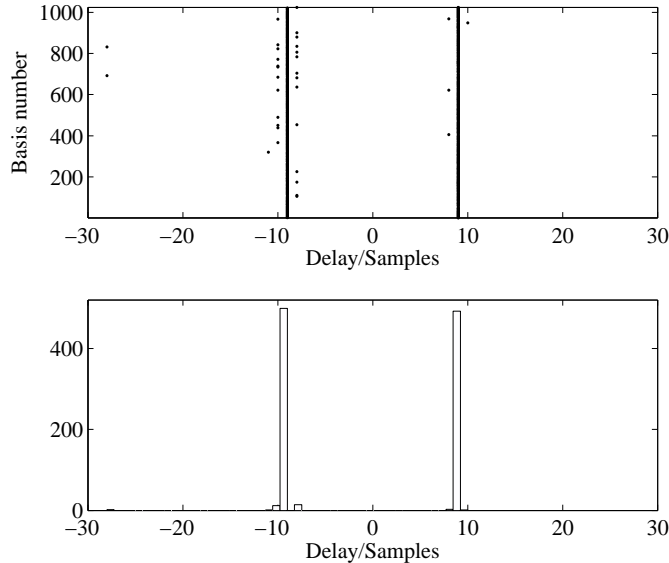


Figure 3: Plot of the time delays estimated for all basis vectors (upper plot), and its histogram (lower plot).

5 Evaluation

We evaluated FD-ICA, DUET and the proposed ASB algorithm on several mixtures of two male speech sources. The speech sources were sampled at 16 kHz with a duration of 1 minute each.

To allow us to control the room Reverberation Time (RT) and the Input Signal-to-Noise Ratio (ISNR), the sources were mixed using simulated room impulse responses, determined by the image technique [31] using McGovern’s RIR Matlab function¹. The positions of the microphones and the loudspeakers are illustrated in figure 4. Six different mixing conditions were obtained by varying RT between 20 ms (320 samples), 80 ms (1300 samples) and 320 ms (5100 samples), and adding white noise to the mixture with ISNRs of 40 dB and 20 dB.

We chose the STFT frame lengths separately for each algorithm, but fixed for all the reverberation times tested. We used the FD-ICA algorithm with the MuSIC-based permutation alignment algorithm described by Mitianoudis and Davies [9], setting the STFT frame size to 2048 samples, which was previously found to be appropriate for this algorithm at a 16kHz sampling rate [9, 32]. For the DUET algorithm we used an STFT frame size of 1024 samples, which was found by Yilmaz and Rickard [12] to give the best separation performance at 16 kHz. For the proposed adaptive stereo basis algorithm, we used an adaptive basis frame size of 512 samples, to be consistent with preliminary experiments which indicated that this would be sufficient for separation at a 16 kHz sampling rate with reasonable room reverberation times [32].

Excerpts of the original mixture and source signals and of the estimated source signals are available for listening on our demo web page².

5.1 Objective evaluation

We evaluated the performance of each method using the objective criteria of Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Noise Ratio (SNR) and Signal-to-Artifacts Ratio (SAR) as defined in [33]. SDR measures the difference between an estimated source and a target source allowing for possible linear filtering between the estimated and target source: we allowed for time-invariant filtering of filter length 1024 samples when calculating SDR. SIR, SNR and SAR provide a more detailed diagnosis of the performance by distinguishing between the elements of the total distortion which are due due to unwanted interfering sources (SIR), remaining mixing noise (SNR) and other artefacts (SAR).

¹<http://2pi.us/code/rir.m>

²http://www.elec.qmul.ac.uk/people/mariaj/asb_demo/

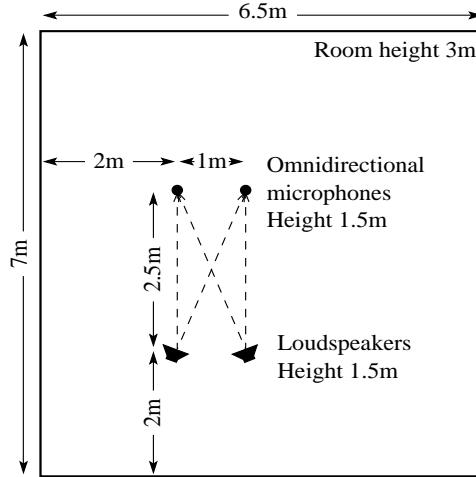


Figure 4: Experimental setup for simulated speech recordings. The reverberation times were set to either 20 ms, 80 ms or 320 ms.

These SDR, SIR, SNR and SAR criteria are defined in [33] on a per-source basis. To gain a single figure for all sources, we averaged the criteria across all microphones and all sources. The results are presented in Table 1.

We see that with short reverberation times (RT=20 ms) our proposed method outperforms both FD-ICA and DUET by more than 7 dB SDR in relatively clean conditions (ISNR=40 dB) and by about 2 dB SDR in more noisy conditions (ISNR=20 dB). As the reverberation time increases, the performance of the proposed method degrades faster than the other algorithms, with FD-ICA (which has the longest frame size) providing best performance at long reverberation times.

In noisy reverberant conditions (ISNR=20 dB, RT=320 ms), we found that the unsupervised K-means clustering algorithm used in the proposed algorithm failed to find both source directions, resulting in a negative SDR. Supervised clustering based on the true source directions improved the SDR to -0.6 dB, but this remained lower than with FD-ICA and DUET in this case. Supervised clustering did not change the performance of the proposed algorithm significantly in other conditions.

Table 1: Objective performance of FD-ICA, DUET and ASB with default frame sizes on simulated speech recordings. All values are expressed in decibels (dB). Bold numbers indicate the best SDR for each mixing condition. See text for comments.

Mixing conditions	ISNR RT	40 dB			20 dB		
		20 ms	80 ms	320 ms	20 ms	80 ms	320 ms
FD-ICA	SDR	7.0	11.2	6.3	6.2	6.5	4.2
	SIR	10.4	16.1	9.1	12.3	14.0	9.1
	SNR	19.1	19.9	28.9	26.7	10.7	25.8
	SAR	11.1	14.2	10.3	7.7	11.4	7.0
DUET	SDR	7.9	8.2	5.3	6.3	5.7	3.5
	SIR	13.4	13.8	10.0	14.7	12.7	8.9
	SNR	21.0	21.0	20.3	11.8	11.8	11.5
	SAR	10.3	10.2	7.9	9.3	9.0	7.3
ASB	SDR	15.4	7.7	1.3	8.3	6.8	-4.2
	SIR	25.7	16.3	8.9	19.7	17.8	7.4
	SNR	20.2	28.0	22.9	12.5	26.3	16.9
	SAR	18.2	9.8	4.2	12.6	7.5	-2.1

5.2 Evaluation using listening tests

When we listened to the result of an earlier preliminary investigation [32], we found that the objective SDR measures did not always correspond to our perceived quality of the separation. We therefore decided to perform a formal subjective listening test to give an alternative comparison of the relative performance of the three algorithms. Such tests are common in audio coding, with standardized test procedures such as MUSHRA (MUltiple Stimulus test with Hidden Reference and Anchors) [34], but have not yet found widespread use in the source separation community.

For our listening test, we adapted the MUSHRA standard and built a Matlab graphical interface to allow subjects to listen to the stimuli and input their scores [35]. Subjects were asked to assess the *basic quality* of each stimulus, a term used to mean the overall perceived quality of the sound, including all possible types of distortion. Each subject was asked to grade the basic quality of the estimated sources compared to a given target source on a scale between 0 and 100, where 100 corresponded to the target source and 0 to the worst estimated source over all conditions. For more details on the listening test procedure, see [35].

Eight subjects took part in the listening tests, and each complete listening test took between about 1 and 2 hours, including breaks. The algorithm developers who had already heard the stimuli were excluded from the listening test. The test results are shown in Figures 5 and 6.

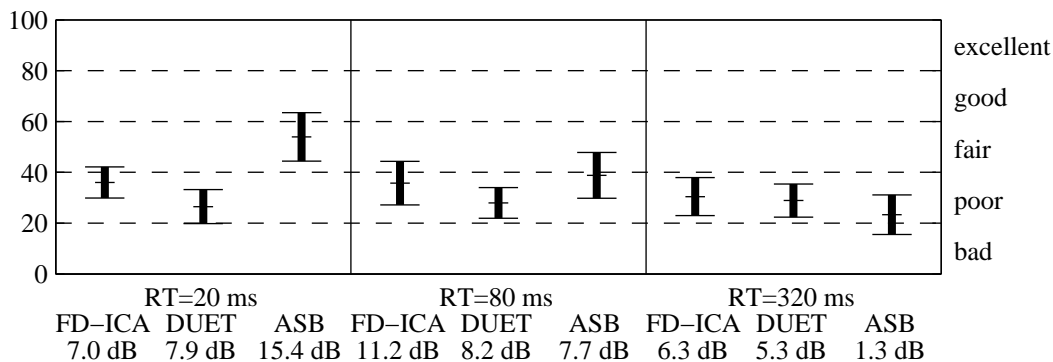


Figure 5: Subjective performance of FD-ICA, DUET and ASB with default frame sizes on simulated speech recordings with ISNR=40 dB. Bars indicate 95% confidence intervals. SDR values are displayed below for comparison. See text for comments.

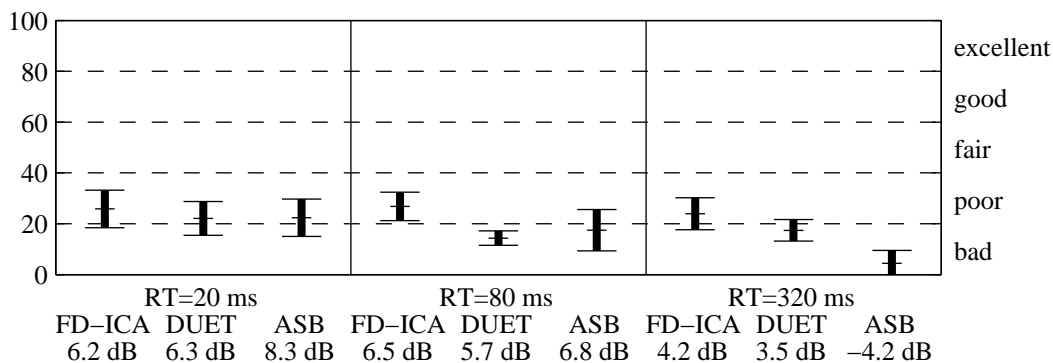


Figure 6: Subjective performance of FD-ICA, DUET and ASB with default frame sizes on simulated speech recordings with ISNR=20 dB. Bars indicate 95% confidence intervals. SDR values are displayed below for comparison. See text for comments.

The results are generally consistent with the objective criteria. In particular, the proposed ASB algorithm performs significantly better than FD-ICA and DUET in clean, less reverberant, conditions (ISNR=40 dB, RT=20 ms). Performance differences are less significant in other conditions, except in the noisy reverberant case (ISNR=20 dB, RT=320 ms) mentioned above.

6 Discussion

6.1 Algorithm comparison

FD-ICA with beamforming-based source matching, DUET and the proposed adaptive stereo basis (ASB) algorithms are based on an essentially similar approach. A transformation is applied on the observed data in order to find a set of basis vectors, followed by direction-based clustering to associate each vector with a source. However, they exhibit some differences that become important when applied to realistic mixtures. We summarize their respective advantages and limitations below.

The main characteristic of ASB is that it is based on an adaptive transform of the observed data, where the basis vectors are estimated from the data. Conversely, FD-ICA and the DUET algorithm use the STFT, a fixed time-frequency transform. Thus we believe that ASB has the potential to provide a sparser representation of the data, and hence improve performance.

DUET and ASB achieve separation by clustering the dictionary elements, the former according to phase (delay) and amplitude information, and the latter according to phase only. FD-ICA with beamforming also uses phase information to align the permutations across all frequencies. Both FD-ICA and DUET suffer from phase ambiguities in the upper frequencies. To avoid this problem, DUET was designed under the assumption that the microphone separation, d , is small enough so that phase ambiguities do not arise [12]. Clearly, this assumption cannot always be satisfied, particularly when the problem is truly blind (i.e. the microphone separation is not known, and cannot be controlled), or for certain applications, such as for CD recordings where phase ambiguities would arise with a sensor spacing of less than 1cm at 44.1 kHz [36]. To help select the correct phase difference between the two sensors where phase ambiguities are possible, a modified version of DUET has been proposed which uses amplitude differences in the high frequency range [36]. In the ASB algorithm we found experimentally that the basis vectors learned by the algorithm are typically time-localized rather than narrowband. It is therefore possible to identify a unique time delay between the left and right channels, using in our case the GCC-PHAT algorithm, and the phase ambiguity problem does not arise.

DUET was developed for anechoic mixing, and can have difficulties dealing with echoic (convolutive) mixing. Histograms obtained from anechoic mixtures are typically well localised, with distinct peak regions corresponding to the sources, while they are more spread out for echoic mixtures [12]. Conversely, ASB does not make any specific assumptions regarding the mixing channel. The learned basis pairs should automatically capture the nature of the channel, so we would expect the method to be able to deal with reverberation. However, the performance of the ASB algorithm does degrade with longer reverberation times (RT = 80 ms and above), perhaps due to the current frame size limit: 80 ms is equivalent to 1280 samples, compared to the currently feasible frame size of 512 samples in the ASB algorithm.

6.2 Training the basis set

In comparison to methods that used a fixed basis, the adaptive stereo basis algorithm requires increased computational expense of fitting an ICA model to the frames of stereo data. There is also a potential problem of *overfitting* due to the large effective dimensionality of the model.

The first problem, that of computational expense, is partly due to the use of a stochastic gradient optimisation. The use of the ‘natural’ or covariant gradient [1] instead of a direct application of a steepest-descent algorithm does improve matters somewhat. We expect that further reduction in computation time would be possible through the use of second-order derivatives (*i.e.* curvature) to improve the convergence of ICA [17].

The second problem, that of overfitting, is potentially more serious as it is an intrinsic limitation of the model in its present form. For example, in our experiments, the ICA weight matrix had 512×512 entries and thus required the optimisation of 262144 parameters. At 16 kHz, a two-channel signal requires approximately 8.2s to deliver this many samples. Our one-minute signals supplied less than 8 times as much data as there were parameters to be optimised, which is rather low and may lead to overfitting.

In applications where the mixing system is known to be stable for long periods, sufficient training data

could be collected to avoid overfitting, though of course this would bring us back to the computational expense of fitting an ICA model to such a large amount of data.

Alternatively, there are several structural aspects of the system that could potentially be exploited to regularise or constrain the ICA model [37]. For a further possibility, since the frames used to train the model are extracted from a longer signal which is assumed to be stationary, there should be no privileged times within the frame. This type of shift invariance has been exploited in single-channel sparse coding [38] and could possibly be adapted for use here.

7 Conclusions

We have considered the problem of convolutive blind audio source separation, and we have presented a stereo coding method. The method is based on the identification of stereo basis vectors adapted to the data. The basis functions are mostly temporally localized, and can be clustered according to directions of arrival (DOA). Separation can then be performed using binary masking on the resulting basis components.

The performance of the algorithm was compared to that of frequency domain ICA (FD-ICA) and the DUET algorithm, using speech signals mixed in a simulated room. Evaluation was performed using both objective measures and subjective listening tests.

The results of both the objective SDR comparison and the formal listening tests indicate that the proposed stereo coding method is competitive with both FD-ICA and the DUET algorithm, and significantly outperforms either of the other algorithms with low noise and short reverberation times ($RT = 20$ ms or 320 samples) of the same order as the frame size used in the ASB algorithm (512 samples). However, the performance of ASB on more echoic rooms ($RT = 80$ ms and above) indicates there is still more work to be done.

In future work, we plan to explore frame sizes longer than 512 samples. To ameliorate the increased computation time involved, we plan to investigate ways to partially structure the ICA bases to allow faster and more robust learning. Other methods may prove useful to learn the basis vector sets, such as the recent K-SVD algorithm [39]. We believe the proposed adaptive stereo basis method is interesting and promising, although further investigation is required in order to reduce the computation cost and improve its robustness to noise and reverberation.

Acknowledgements

This work was funded by EPSRC grants GR/S85900/01, GR/R54620/01, and GR/S82213/01. The authors wish to thank Scott Rickard and Nikolaos Mitianoudis for providing implementations of the DUET and FD-ICA algorithms, and all the subjects who participated in the listening tests.

References

- [1] S.-I. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang, “Multichannel blind deconvolution and equalization using the natural gradient,” in *Proceedings of the First IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Paris, France, April 1997, pp. 101–104.
- [2] K. Torkkola, “Blind separation of convolved sources based on information maximization,” in *Proc. of the IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, 1996, pp. 423–432.
- [3] S. C. Douglas and X. Sun, “Convolutive blind separation of speech mixtures using the natural gradient,” *Speech Communication*, vol. 39, pp. 65–78, 2003.
- [4] S. Makino, H. Sawada, R. Mukai, and S. Araki, “Blind source separation of convolutive mixtures of speech in frequency domain,” *IEICE Trans. Fundamentals*, vol. E88, pp. 1640–1655, 2005.

- [5] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, pp. 21–34, 1998.
- [6] S. Ikeda and N. Murata, “A method of ICA in time-frequency domain,” in *Proc. of the International Conference on Independent Component Analysis and Blind Source Separation (ICA99)*, Aussois, France, 1999, pp. 365–371.
- [7] L. C. Parra and C. Spence, “Convolutional blind separation of non-stationary sources,” *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 320–327, 2000.
- [8] M. E. Davies, “Audio source separation,” in *Mathematics of Signal Processing*, J. G. McWhirter and I. K. Proudler, Eds. Oxford University Press, 2002, pp. 57–68.
- [9] N. Mitianoudis and M. E. Davies, “Audio source separation of convolutional mixtures,” *IEEE Trans. on Speech and Audio Processing*, vol. 11, pp. 489–497, 2003.
- [10] A. Jourjine, S. Rickard, and Ö. Yilmaz, “Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures,” in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, 2000, pp. 2985–2988.
- [11] S. Rickard and F. Dietrich, “DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET,” in *Proceedings of the IEEE Workshop on Statistical Signal and Array Processing (SSAP)*, 2000, pp. 311–314.
- [12] Ö. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. on Signal Processing*, vol. 52, pp. 1830–1847, 2004.
- [13] S. A. Abdallah and M. D. Plumbley, “Application of geometric dependency analysis to the separation of convolved mixtures,” in *Proc. of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, Granada, Spain, 2004, pp. 22–24.
- [14] 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. on Speech and Audio Processing*, vol. 12, pp. 530–538, 2004.
- [15] J.-F. Cardoso and B. Laheld, “Equivariant adaptive source separation,” *IEEE Trans. on Signal Processing*, vol. 44, pp. 3017–3030, 1996.
- [16] S. Amari and A. Cichocki, “Adaptive blind signal processing - neural network approaches,” *Proceedings of the IEEE*, vol. 86, pp. 2026–2048, 1998.
- [17] A. Hyvärinen, “Fast and robust fixed-point algorithms for independent component analysis,” *IEEE Trans. on Neural Networks*, vol. 86, pp. 626–634, 1999.
- [18] S. Araki, S. Makino, R. Mukai, Y. Hinamoto, T. Nishikawa, and H. Saruwatari, “Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming,” in *Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP’02)*, vol. 2, 2002, pp. 1785–1788.
- [19] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, “Evaluation of blind signal separation method using directivity pattern under reverberant conditions,” in *Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, vol. 5, 2000, pp. 3140–3143.
- [20] H. Saruwatari, S. Kurita, and K. Takeda, “Blind source separation combining frequency-domain ICA and beamforming,” in *Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, vol. 5, 2001, pp. 2733–2736.

- [21] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 881–884.
- [22] N. Mitianoudis and M. E. Davies, "Permutation alignment for frequency domain ICA using subspace beamforming methods," in *Proc. of the International Conference on Independent Component Analysis and Blind Source Separation (ICA 2004)*, Granada, Spain, 2004, pp. 669–676.
- [23] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, pp. 276–280, 1986.
- [24] P. D. O’Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology*, vol. 15, pp. 18–33, 2005.
- [25] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proceedings of the IEEE*, vol. 86, pp. 2009–2025, 1998.
- [26] A. Bell and T. Sejnowski, "Learning the higher-order structure of a natural sound," *Computation in Neural Systems*, vol. 7, pp. 261–266, 1996.
- [27] S. A. Abdallah and M. D. Plumbley, "If edges are the independent components of natural images, what are the independent components of natural sounds?" in *Proc. of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA2001)*, San Diego, California, 2001, pp. 534–539.
- [28] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, pp. 356–363, 2002.
- [29] L. Zhang, A. Cichocki, and S.-I. Amari, "Self-adaptive blind source separation based on activation functions adaptation," *IEEE Transactions on Neural Networks*, vol. 15, no. 2, pp. 233–244, 2004.
- [30] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustic, Speech, and Signal Processing*, vol. 24, pp. 320–327, 1976.
- [31] S. McGovern, "A model for room acoustics," 2003, Available at: <http://2pi.us/rir.html>.
- [32] M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Sparse coding for convolutive blind audio separation," in *Proc. of the International Conference on Independent Component Analysis and Blind Source Separation (ICA 2006)*, Charleston, SC, USA. Springer-Verlag, Berlin, 2006, pp. 132–139.
- [33] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.
- [34] *Recommendation ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems*, International Telecommunication Union, 2003.
- [35] E. Vincent, M. G. Jafari, and M. D. Plumbley, "Preliminary guidelines for subjective evaluation of audio source separation algorithms," in *Proc. of the ICA Research Network International Workshop, Liverpool, UK*, 2006, to appear.
- [36] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proceedings of the International Conference on Digital Audio Effects (DAFx-03)*, London, UK, 2003, pp. 209–213.
- [37] Y. Matsuda and K. Yamaguchi, "Linear multilayer ICA integrating small local modules," in *4th Intl. Symp. on Independent Component Analysis and Signal Separation (ICA2003)*, Nara, Japan, 2003, pp. 403–408.

- [38] T. Blumensath and M. Davies, “Unsupervised learning of sparse and shift-invariant decompositions of polyphonic music,” in *Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, vol. 5, Montreal, Canada, May 2004, pp. V-497–500.
- [39] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: Design of dictionaries for sparse representation,” in *Proceedings of the Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS’05), Rennes, France*, November 2005, pp. 9–12.