Sparse Coding for Convolutive Blind Audio Source Separation*

Maria G. Jafari, Samer A. Abdallah, Mark D. Plumbley, and Mike E. Davies

Centre for Digital Music,
Queen Mary University of London, UK
{maria.jafari, samer.abdallah}@elec.qmul.ac.uk
http://www.elec.qmul.ac.uk

Abstract. In this paper, we address the convolutive blind source separation (BSS) problem with a sparse independent component analysis (ICA) method, which uses ICA to find a set of basis vectors from the observed data, followed by clustering to identify the original sources. We show that, thanks to the temporally localised basis vectors that result, phase information is easily exploited to determine the clusters, using an unsupervised clustering method. Experimental results show that good performance is obtained with the proposed approach, even for short basis vectors.

1 Introduction

The convolutive blind audio source separation problem arises when an array of microphones records mixtures of a set of sound sources that are convolved with the impulse response between each source and sensor. The problem is often addressed in the frequency domain, through the short-time fourier transform (STFT), where the statistics of the sources are sparser, so that ICA algorithms achieve better performance [1], and the approximations of convolutions by multiplications yield reduced computational complexity. Source separation is then performed separately at each frequency bin, resulting in the introduction of the well-known problem of frequency permutations [2], whose solution amounts to clustering the frequency components of the recovered sources, using additional information about the mixing system or the sources. The most successful methods in this context have perhaps been beamforming approaches [2-5], which exploit phase information contained in the de-mixing filters identified by the source separation algorithm, but suffer from phase ambiguities in the upper frequencies, since phase is defined exclusively up to 2π. An alternative approach to convolutive BSS was proposed in [7], and is based on the use of sparse coding to identify the mixing matrix from the observed data. No assumptions are required on the number of microphones, or the type of mixing (eg. instantaneous or convolutive) in the underlying model, but the recovered matrix implicitly encodes

* This work was funded by EPSRC grants GR/S85900/01, GR/R54620/01, and GR/S82213/01.

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 132–139, 2006.

these characteristics of the system. Thus, it could even potentially deal with the more sources than sensors case. The subspaces corresponding to the original sources are then identified using clustering techniques. In this paper we investigate the performance of the frequency domain ICA (FD-ICA) and sparse coding approaches. We find that the latter yields mostly temporally localised basis vectors, that do not suffer from the phase ambiguity encountered in the frequency domain. Hence, in contrast to the approach in [7], which uses manual clustering, we propose an unsupervised clustering method that exploits phase information to separate the sources. The structure of this paper is as follows: the convolutive BSS problem is described in section 2, together with an overview of FD-ICA; the sparse coding method is summarised in section 3. The clustering technique proposed is discussed in section 4, where the performance of the sparse coding and FD-ICA methods are also compared. Conclusions are drawn in section 5.

2 Problem Formulation

We consider the problem of separating 2 sampled real-valued speech signals, $\mathbf{s}(n)$, from 2 convolutive mixtures, $\mathbf{x}(n)$, recorded from an array of microphones, so that the signal recorded at the q-th microphone, $x_q(n)$, is

$$x_q(n) = \sum_{p=1}^{2} \sum_{l=1}^{L} a_{qp}(l) s_p(n-l), \quad q = 1, 2 \tag{1}$$

where $s_p(n)$ is the p-th source signal, $a_{qp}(l)$ denotes the impulse response from source p to sensor q, and L is the maximum length of all impulse responses [2]. The aim of blind source separation is to find estimates for the unmixing filters $w_{qp}(l)$, using only the sensor measurements, and to reconstruct the sources from

$$y_p(n) = \sum_{q=1}^{2} \sum_{l=1}^{L} w_{qp}(l) x_q(n-l), \quad p = 1, 2 \tag{2}$$

where $y_p(n)$ is the p-th recovered source. Typically, the N-point STFT is evaluated, and the mixing and separating models in (1) and (2) become, respectively $\mathbf{X}(f,t) = \mathbf{A}(f)\mathbf{S}(f,t)$ and $\mathbf{Y}(f,t) = \mathbf{W}(f)\mathbf{X}(f,t)$ where t denotes the STFT block index. The resulting N instantaneous BSS problems, are addressed independently in each subband with an ICA algorithm, and the problem of frequency permutations that is introduced is solved essentially by clustering the frequency components of the recovered sources. This is often done using beamforming techniques, such as in [2-5], where the direction of arrival (DOA) of the sources are evaluated from the beamformer directivity patterns

$$F_p(f,\theta) = \sum_{q=1}^{2} W_{qp}^{\mathrm{ICA}}(f) e^{j2\pi f d \sin\theta_p / c}, \quad p = 1, 2 \tag{3}$$

where W_{qp}^{ICA} is the ICA de-mixing filter from the q-th sensor to the p-th output, d is the spacing between two sensors, θ_p is the angle of arrival of the p-th source

signal, and $c \approx 340$m/s is the speed of sound in air. The frequency permutations are then determined by ensuring that the directivity pattern for each beamformer is approximately aligned along the frequency axis. There exists, however, an ambiguity in the DOA estimation, due to the restriction on the phase difference to lie between $-\pi$ and π, which results in the creation of additional nulls in the directivity pattern of magnitude similar to that corresponding to the angle of arrival [5]. The distance between two microphones should satisfy $d \leq c/2f_{\max}$, in order to avoid spatial aliasing [2]; when this condition is not met, ambiguities in the position of the nulls are introduced, resulting in inaccurate DOA estimates, and the frequency, f_{M}, above which multiple nulls are expected is $f_{\mathrm{M}} = c/2d$.

3 Overview of Sparse ICA

The aim of sparse coding is to find sparse dictionaries from the mixtures, so that only a small number of coefficients, $\mathbf{s}(n)$, are needed to encode the observed data, $\mathbf{x}(n)$ [6]. The convolutive BSS problem was first addressed within this framework in [7], by finding a set of basis vectors for the observed data, followed by clustering to identify the subspaces corresponding to the original sources. The approach does not explicitly model the mixing process nor the number of mixtures, but is based on the assumption that the recordings are generated by signals that are sparse in the dictionary domain. Prior to estimating the basis vectors, the observed vector is reshaped into a $K \times k_{max}$ matrix on which learning is performed. A frame 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

$$\tilde{\mathbf{X}}_{i,k} = \begin{cases} x_1\left[(k-1)Z + \frac{i+1}{2}\right] & : i \text{ odd} \\ x_2\left[(k-1)Z + \frac{i}{2}\right] & : i \text{ even} \end{cases} \tag{4}$$

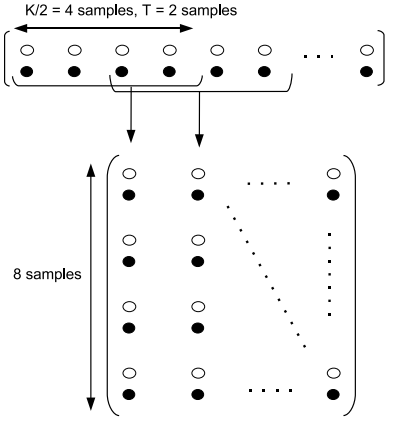


Fig. 1. Reshaping of the sensor vector prior to training with ICA

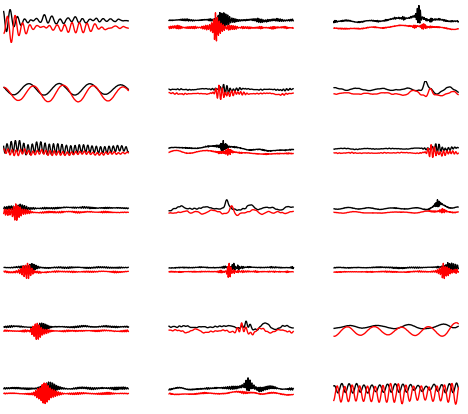


Fig. 2. Examples of basis vectors extracted with the sparse ICA algorithm

where $Z = K/2-T$, and $i \in \{1, \ldots, K/2\}$, and $k \in \{1, \ldots, k_{max}\}$. The reshaping of the sensor vector $\mathbf{x}(n)$ is illustrated in figure 1. The basis vectors are learned from the resulting matrix, and with any ICA algorithm using a sparse prior. Here we use [7]

$$\mathbf{\Delta W} = \eta \left(\mathbf{I} - E\{\mathbf{f}(\mathbf{y})\mathbf{y}^T\} \right) \mathbf{W} \tag{5}$$

where η is the learning rate, and $\mathbf{f}(\mathbf{y})$ is the activation function. Details for its choice can be found in [7]. The algorithm (5) operates upon $\mathbf{y} = \mathbf{W}\tilde{\mathbf{X}}$, where the time index n has been dropped for the sake of clarity, and $\mathbf{W} \in \mathbb{R}^{K \times K}$. The reshaping of $\mathbf{x}(n)$ into the matrix $\tilde{\mathbf{X}}(n)$ emphasises the correlations between the sources at the two microphones. Stacking the columns of $\mathbf{x}(n)$ ensures that features relating to temporally correlated signals from each recording are extracted, leading to basis pairs that encode information about the mixing channel, as can be seen from the basis pairs plotted in figure 2, where a time-delay is clearly visible in several of the vector pairs. The strong directionality observed indicates that each basis pair relates to a particular source, and thus the proposed method is based on the property of spatial diversity. However, should the sources be aligned along the same DOA, the technique cannot be used.

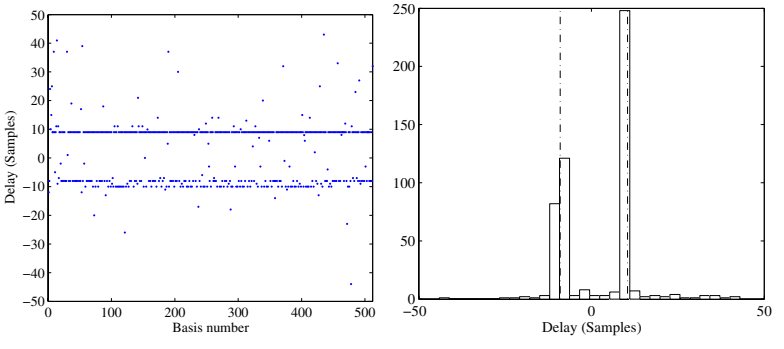
4 Frequency Domain Versus Sparse ICA

In this section, we consider the separation of two speech signals, one each from a male and a female speaker, from two mixtures recorded in a university lecture room, and sampled at 16kHz. Further details of the experimental set up can be found in [10]. The sources were also recorded separately, so that they could be used for performance evaluation. The performance of the sparse ICA approach is compared to a representative FD-ICA method [10] (MD2003) since, due to their inherent similarities, we expect other FD-ICA algorithms to have comparable performance. The sparse ICA approach was first used to learn the basis vectors from the real data; the mixtures were buffered into frames of 512 samples, so

that 256 samples from each mixture were taken, as shown in figure 1, and the algorithm (5) was used for training. The learned basis vector pairs are found in the columns of $\mathbf{W}^{-1}$, examples of which are shown in figure 2. This figure illustrates that the basis vector pairs encode how the extracted features are received at the microphones and, therefore, they capture information about time-delay and amplitude differences that characterise the mixing channel. Moreover, most of the basis vectors have the additional property of being localised in time, which implies that time delays can be estimated more accurately. Reconstruction of the two original signals is achieved with $\hat{\mathbf{s}}_1 = \mathbf{W}^{-1}\mathbf{H}^{(1)}\mathbf{y}$, and $\hat{\mathbf{s}}_2 = \mathbf{W}^{-1}\mathbf{H}^{(2)}\mathbf{y}$ where $\mathbf{H}^{(1)}$ is a diagonal matrix whose diagonal elements are ones or zeros depending on whether a component belongs to the first source, and similarly for $\mathbf{H}^{(2)}$ [7]. We propose clustering the basis vector pairs, and therefore determine the diagonal elements of $\mathbf{H}^{(1)}$ and $\mathbf{H}^{(2)}$, according to the following algorithm:

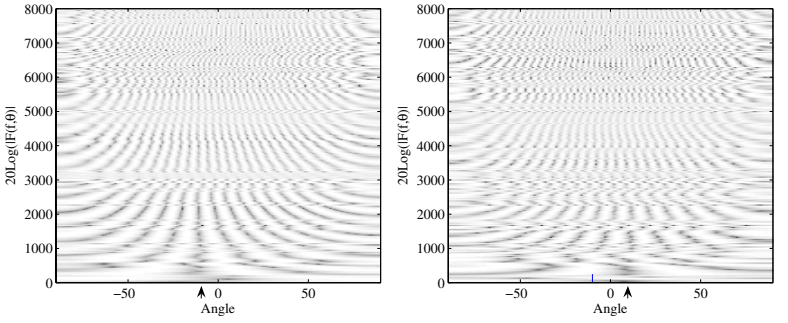
1 For each basis pair k find the time delay τ_k between the vectors
2 Form the histogram of τ_k, and use k-means to find the peaks, τ_{k_1} and τ_{k_2} corresponding to the sources
3 $h_{kk}^{(p)} = \begin{cases} 1, & \text{if } (\tau_{k_p} - \tau_\delta) \leq \tau_k \leq (\tau_{k_p} + \tau_\delta), \\ 0, & \text{otherwise} \end{cases}$

for $p = 1, 2$, where $h_{kk}^{(p)}$ is the kk-th element of $\mathbf{H}^{(p)}$. The inclusion of the τ_δ allows the algorithm to perform a degree of de-noising. We estimate the time delay between sensor pairs using the popular generalised cross-correlation with phase transform (GCC-PHAT) algorithm, originally proposed in [9], $R_{a_1 a_2}(\tau) = \int_{-\infty}^{\infty} A_1(\omega) A_2^*(\omega) / (|A_1(\omega) A_2^*(\omega)|) e^{j\omega\tau} d\omega$, where $A_1(\omega)$, $A_2(\omega)$ are the Fourier transforms of the basis vectors. 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. Figure 3(a) depicts the time-delay estimates obtained with GCC-PHAT, for all basis vector pairs, and figure 3(b) shows their histogram; values of τ_{k_1} and τ_{k_2} were obtained with k-means as 10.04 and -9.03 samples, and τ_δ was set to 2 samples.



(a) Delays for all basis vectors (b) Histogram of delays

Fig. 3. Plot of the time delays estimated for all basis vectors, and its histogram



(a) Directivity pattern for source $\hat{s}_1$. (b) Directivity pattern for source $\hat{s}_2$.

Fig. 4. Directivity patterns for the outputs of FD-ICA, after permutation alignment

Frequency domain separation was performed with the algorithm in [10] using the 256 and 2048-point STFT, and permutations were aligned as in [3]. MD2003 with the latter frame length has been shown to successfully achieve separation on this data in [10]. Figure 4 shows a plot of the directivity pattern of the outputs evaluated at all frequencies with (3), following permutation alignment. The plots show that permutations are correctly aligned in the low frequency bands, while the behaviour of the algorithm is less clear in the higher frequencies, where time delay estimation is less accurate. The DOAs estimated from the plots were found to be 12° and −11°, corresponding to time delays of approximately 10 and −9 samples. The sample delay at a frequency f is estimated from the directions of arrival by $\tau = 2\pi f \sin\theta f_s/c$, where f_s is the sampling frequency.

Tables 1 and 2 show the global performance of the two methods, as evaluated from [11]. The evaluation criteria allows for the recovered sources to be modified by a permitted distortion. In Table 1, we consider a time-invariant gain distortion, and the sources recovered at the two channels are compared to the

Table 1. Global performance measures when a gain distortion is allowed. SDR, SIR, and SAR measures are respectively the signal-to-distortion, signal-to-interference, and signal-to-artifact ratios.

Method	Channel 1					
	SDR (dB)		SIR (dB)		SAR (dB)	
	$\hat{s}^{\mathrm{M}}$	$\hat{s}^{\mathrm{F}}$	$\hat{s}^{\mathrm{M}}$	$\hat{s}^{\mathrm{F}}$	$\hat{s}^{\mathrm{M}}$	$\hat{s}^{\mathrm{F}}$
$\mathrm{MD2003}_{256}$	−5.13	−8.61	2.73	1.83	−2.49	−6.01
$\mathrm{MD2003}_{2048}$	−5.24	−6.26	4.69	6.17	−3.50	−5.07
Sparse ICA	−8.69	−10.09	1.59	2.74	−5.98	−8.00
	Channel 2					
$\mathrm{MD2003}_{256}$	−8.29	−6.09	−0.81	2.64	−4.00	−3.58
$\mathrm{MD2003}_{2048}$	−6.94	−3.42	3.86	7.24	−5.06	−2.28
Sparse ICA	−6.76	−11.61	−0.41	7.40	−2.40	−10.83

Table 2. Global performance measures when a filter distortion is allowed

Method	SDR (dB)		SIR (dB)		SAR (dB)	
	$\hat{s}^{\mathrm{M}}$	$\hat{s}^{\mathrm{F}}$	$\hat{s}^{\mathrm{M}}$	$\hat{s}^{\mathrm{F}}$	$\hat{s}^{\mathrm{M}}$	$\hat{s}^{\mathrm{F}}$
	Channel 1					
$\mathrm{MD2003}_{256}$	−7.12	−8.25	2.47	2.01	−4.66	−5.69
$\mathrm{MD2003}_{2048}$	−6.27	−7.07	7.82	9.26	−5.43	−6.48
Sparse ICA	−7.77	−10.50	3.94	3.55	−6.00	−8.74
	Channel 2					
$\mathrm{MD2003}_{256}$	−10.33	−8.03	−0.37	2.21	−6.67	−5.55
$\mathrm{MD2003}_{2048}$	−8.15	−5.62	6.52	9.99	−7.12	−5.08
Sparse ICA	−8.27	−9.40	1.29	12.00	−5.35	−9.10

original sources recorded at the microphones. Negative SIR values indicate that the interfering source is larger than the target source, and the algorithm has failed to recover the target. Large negative SAR values, with SAR ≈ SDR, indicate that large artifacts are present, and dominate distortion [11]. The results suggest that sparse ICA and $\mathrm{MD2003}_{256}$ have similar performance, and both fail to recover the source $\hat{s}^{\mathrm{M}}$ at channel 2. An informal listening test indicates, however, that the objective assessment in Table 1 is not a good guide to the audible performance. The test reveals that MD2003 separates the sources with a frame of 2048 samples, while it fails with a short frame of 256 samples. This is in contrast to sparse ICA which uses a frame of 256 samples, and whose outputs are clearly separated, although the interfering source is still audible. Interestingly, the algorithm seems also to have performed some de-reverberation, which is particularly audible for the female source, $\hat{s}^{\mathrm{F}}$, at the second channel. Moreover, the outputs sound quite natural and large artifacts do not appear to be present. This is in disagreement with the large negative SAR values which suggest that sparse ICA introduces large artifacts. To obtain a more meaningful objective assessment, a time-invariant filter distortion is allowed, with a 64 taps filter. The results are shown in Table 2, where the recovered sources are compared to the original signals at the speakers. In this case, it was found that the objective assessment is more closely in agreement with the informal listening test, but still overcritical of sparse ICA. The results in this section also show how the STFT length is a crucial parameter for FD-ICA. Since modeling of real room transfer functions typically requires long frame sizes, better separation is achieved with a frame size of 2048 samples. Sparse ICA, on the other hand, provides good separation even with a very short frame size.

5 Conclusions

In this paper, we have shown that most of the basis vectors extracted with sparse coding are temporally localised functions that do not suffer from phase ambi-

guities encountered in the frequency domain. A simple unsupervised clustering technique that exploits this property has been proposed. The performance of the algorithm with real data has been investigated, and informal listening tests have suggested that it separates the signals with short basis vectors, in contrast to FD-ICA, which requires long basis vectors. Currently available objective testing methods fail to verify this, so further subjective listening tests are planned to formally substantiate this performance.

References

1. J.F. Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, pp. 2009–2025, 1998.
2. 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.
3. N. Mitianoudis and M. Davies, "Permutation alignment for frequency domain ICA using subspace beamforming methods," in *Proc. ICA*, 2004, pp. 669–676.
4. H. Saruwatari, S. Kurita, and K. Takeda, "Blind source separation combining frequency-domain ICA and beamforming," in *Proc. ICASSP*, 2001, vol. 5, pp. 2733–2736.
5. M. Ikram and D. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. ICASSP*, 2002, vol. 1, pp. 881–884.
6. J.-H. Lee, T.-W. Lee, H.-Y. Jung, and S.-Y. Lee, "On the efficient speech feature extraction based on independent component analysis," *Neural Processing Letters*, vol. 15, pp. 235–245, 2002.
7. S. Adballah and M. Plumbley, "Application of geometric dependency analysis to the separation of convolved mixtures," in *Proc. ICA*, 2004, pp. 22–24.
8. J.F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 44, pp. 3017–3030, 1996.
9. C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustic, Speech, and Signal Processing*, vol. 24, pp. 320–327, 1976.
10. N. Mitianoudis and M. Davies, "Audio source separation of convolutive mixtures," *IEEE Trans. on Audio and Speech Processing*, vol. 11, pp. 489–497, 2003.
11. C. Févotte, R. Gribonval and E. Vincent, "BSS_EVAL Toolbox User Guide," *IRISA Technical Report 1706*, April 2005. http://www.irisa.fr/metiss/bss_eval/.