Audio Source Separation of convolutive mixtures

Nikolaos Mitianoudis, Mike Davies

Abstract— **The problem of separation of audio sources recorded in a real world situation is well established in modern literature. A method to solve this problem is Blind Source Separation (BSS) using Independent Component Analysis (ICA). The recording environment is usually modelled as *convolutive*. Previous research on ICA of instantaneous mixtures provided solid background for the separation of convolved mixtures. The authors revise current approaches on the subject and propose a fast frequency domain ICA framework, providing a solution for the apparent *permutation problem* encountered in these methods.**

Keywords— **Audio Source separation, frequency domain Independent Component Analysis, convolutive mixtures.**

I. Introduction

SUPPOSE there are N audio sources in a room $\underline{s}[n] = [s_1[n], s_2[n], \ldots, s_N[n]]^T$ and M microphones capturing the auditory scene through the *observation signals* (*mixtures*) $\underline{x}[n] = [x_1[n], x_2[n], \ldots, x_M[n]]^T$. Generally, the *Blind Source Separation* (BSS) problem is defined as the procedure of estimating the original signals $\underline{s}[n]$ through the observed signals $\underline{x}[n]$. The term *blind* stresses that these methods employ no prior information about the source signals. In practice, all BSS methods are *semi-blind*, as some assumptions about the sources' statistics are often made. However, these models tend to be quite general, thus preserving the versatility of the method.

One can model the recording environment and form an expression connecting the observed signals and the original signals. A first crude approximation can be that each microphone captures a portion of each source. This seems to be a rather simplified model. Nonetheless, if we are referring to studio recordings, where audio signals are mixed using a mixing desk, the mixed signals can be modelled as linear combinations of the original sources, i.e. *instantaneous mixtures* of the sources.

$$\underline{x} = A\underline{s} + \underline{\epsilon} \tag{1}$$

where $\underline{\epsilon}$ models the additive noise captured during recording. The additive noise is still a very difficult and open problem for the BSS framework. The impact of noise on the estimator's performance depends on the type and level of noise. Cardoso [4] has pointed out that the benefits of noise modelling for BSS are not so clear. In cases of high SNR, the bias on estimates for A are small and some noise reduction can be achieved using robust denoising techniques [11] as a pre- or post-separation task. In cases of low SNR, the problem is very difficult to solve anyway. For the rest of the analysis, we will ignore the additive noise term for simplicity. Moreover, we will assume that the number of sources is equal to the number of sensors, i.e. $N = M$. In the instantaneous mixtures case, the blind source separation problem is concentrated on estimating the *unmixing matrix* $W \approx A^{-1}$ that can separate the audio sources.

A method proposed to solve this problem is *Independent Component Analysis* (ICA). ICA exploits the *nonGaussianity* of source signals and assumes *statistical independence* of the separated signals to perform separation. Similar methods for ICA have been developed from a number of different view points: minimising *Kullback-Leibler* (KL) divergence [1], *Infomax* [2] or *Maximum Likelihood* estimation [5], [23]. Other methods look for the directions of the most nonGaussian components, using *kurtosis* or *negentropy* as nonGaussianity measures [3], [16], [14]. Finally, other approaches estimate the unmixing matrix by performing approximate diagonalization of a *cumulant tensor* of the mixtures [7].

Unfortunately, the instantaneous mixtures model is rather incomplete in the case of sources recorded in a real acoustic room environment. Assume the case of a sound source and a microphone in a room. Previous research has shown that the signal captured by the microphone can be well represented by a *convolution* of the source signal with a high-order FIR filter, modelling the room acoustics between the source and the sensor [8]. To increase accuracy, we could use lower-order IIR filters to model room acoustics. However, as IIR filters are less stable and require minimum-phase mixing, we will model the channel using FIR filters. In the case of many sources and sensors, the signal at each sensor can be modelled by the following equation:

$$x_i[n] = \sum_{j=1}^{N} \sum_{k=1}^{L} \alpha_{jk} s_j[n-k] \qquad i = 1, \ldots, N \tag{2}$$

where L denotes the maximum room filter length. Equation (2) actually represents the superposition of the *convolutions* of the N sources with N filters of maximum length L. These mixtures are referred to as *convolutive mixtures*.

Many methods have been proposed to solve the convolutional ICA problem. Some of them suggested working directly in the time-domain [19], [28]. Working in the time domain has the disadvantage of being rather computational expensive, due to calculating many convolutions. Other approaches suggested moving to the STFT domain in order to transform the convolution into multiplication and apply ICA methods for instantaneous mixtures for each frequency bin [26]. Furthermore, audio signals are more nonGaussian in the frequency domain than in the time domain, making the STFT a suitable ICA framework. However, there is an inherent *permutation problem* in all FD-ICA methods, which does not exist in time-domain methods.

The authors are with the Electronic Engineering Department, Queen Mary, University of London, Mile End Road, London E1 4NS, UK (e-mail: nikolaos.mitianoudis@ elec.qmul.ac.uk).

The rest of the paper is organised as follows. In section II, we set a mathematical formulation of the problem, and revise some proposed time-domain and frequency-domain solutions. We introduce the permutation problem in frequency domain methods and review some proposed solutions. In section III, we propose a new framework for Frequency-domain ICA. First of all, we propose modelling the sources in the STFT domain. Secondly, we incorporate a time-varying scaling parameter in the proposed *time-frequency source model* to impose efficient frequency coupling to solve the permutation problem along with a *Likelihood Ratio* jump. In addition, a *fast frequency domain ICA* algorithm is proposed to improve the speed and stability of the FD-ICA framework. In section IV, we evaluate the performance of the proposed framework.

II. Previous Work on Convolutive Mixtures

A. Problem Formulation

Let $\underline{x}[n] = [x_1[n], x_2[n], \ldots, x_N[n]]^T$ be N observed signals that are convolutive mixtures of N statistically independent sources $\underline{s}[n] = [s_1[n], s_2[n], \ldots, s_N[n]]^T$ such that:

$$x_i[n] = \sum_k \mathbf{A}(k)\underline{s}[n-k] \tag{3}$$

where $\mathbf{A}(k)$ is the FIR mixing matrix [26], modelling the room acoustics between each source and sensor. The problem is to estimate the unknown source signals given the observations $\underline{x}[n]$. There are a number of ambiguities present in this approach: *permutation* and *spectral shape* (including *sign*). Methods to overcome these ambiguities will be described later on.

For the rest of the analysis, we will assume that the room acoustics can be inverted, a common assumption made by all convolutional ICA methods. However, this is not always valid. It is shown [12] that the mixing matrix A can have poor conditioning for a range of source-sensor placements, in addition to symmetric source-sensor geometries. If the sensors are placed at or near these ill-conditioned locations, then blind source separation methods either fail or have degraded performance, as the problem is literally reduced to the *less sensors than sources* case.

B. Time-domain Methods

The first efforts on source separation of convolved mixtures were made in the time domain, mainly inspired by *blind deconvolution* methods. Torkolla [28] modelled the unmixing procedure as FIR filtering of the mixtures.

$$u_i[n] = \sum_{j=1}^{N}\sum_{k=1}^{L} w_{jk}x_j[n-k] \tag{4}$$

In order to estimate the coefficients w_{jk}, he used an information maximisation approach, similar to [2].

Lee et al [19] looked at modelling the unmixing procedure as an IIR filter and derived a solution for this problem, noting that the recording environment had to be *minimum phase*, which is not always valid. Therefore, he proposed a FIR approach calculating the updates in the frequency domain but the non-linearity (i.e. the signal model) in the time domain.

$$\Delta W_f = \eta(I - \mathcal{E}\{FFT(f, \phi(\underline{u}(t)))\underline{u}^H(f,t)\})W_f \tag{5}$$

where η is the learning rate and f denotes the corresponding frequency bin. This does not incur a permutation problem but does involve repeated mapping between the frequency domain and the time domain during optimisation.

C. Frequency-domain methods

Smaragdis [26] proposed working solely in the frequency domain. Transforming equation (3) to the STFT domain (using sufficiently long frames), we can transform the convolution to multiplication.

$$\underline{x}(f,t) = A_f\underline{s}(f,t) \tag{6}$$

where A_f is a $N \times N$ complex matrix. The problem is defined as estimating an unmixing matrix $W_f \approx A_f^{-1}$ for each frequency bin. Assuming frequency independent priors for W_f and the separated sources $\underline{u}(f,t)$, one can find a ML solution separately for each frequency bin by maximising the following log-likelihood:

$$\log p(\underline{x}(f,t)|W_f) = \mathcal{E}\{\log p(\underline{u}(f,t))\} + \log\det W_f \tag{7}$$

The *natural gradient* algorithm [1], although derived from a different principle, can be used to solve this problem [5]. We can perform unimixing by running independent natural gradient algorithms for every frequency bin, i.e.

$$\Delta W_f = \eta(I - \mathcal{E}\{\phi(\underline{u}(f,t))\underline{u}^H(f,t)\})W_f \tag{8}$$

D. The permutation problem in frequency-domain methods

Unfortunately, solving the problem independently for each frequency bin generates the *permutation problem*, since there is the inherent permutation ambiguity in the rows of W_f [22], [26]. This is more complicated than the *ordering ambiguity* in instantaneous mixtures ICA, since the ordering of the sources must remain the same along the frequency axis. As a result, the algorithm produces different permutations of separated sources along the frequency axis, and therefore the sources remain mixed. To solve this problem, we need to impose some *frequency coupling*.

Smaragdis proposed an adaptive scheme to apply some frequency coupling between neighboring frequency bins.

$$\Delta W_f \leftarrow \Delta W_f + k\Delta W_{f-1} \tag{9}$$

where $0 < k < 1$. This heuristic adaptive solution can be interpreted as placing weakly coupled priors on W_f of the form:

$$p(W_f|W_{f-1}) \propto \exp(-\frac{1}{2\sigma^2}||W_f - W_{f-1}||_F) \tag{10}$$

This imposes some weak smoothness constraint across frequency. However, it had limited effect.

Parra et al [22] also worked in the frequency domain using non-stationarity to perform separation. Their solution to the problem was to impose a constraint on the unmixing filter length q. This is achieved by applying a projection operator P to the filter estimates at each iteration, where $P = FZF^{-1}$, F is the Fourier transform and Z is a diagonal operator that projects on the first q terms. In other words, it imposes a *smooth* constraint on the unmixing filters, as they are modelled as FIR filters. Again mixed success has been reported for this method.

Another solution is to use *beamforming* to align the permutations along the frequency axis. All BSS methods make no assumptions about the position of the sources in the 3D space. However, *beamforming* estimates the directions of signal's arrival (DOA) in order to steer the beam of an array of sensors to focus on a specific source. The BSS setup can be regarded as a N-microphone beamformer. Saruwatari et al [24] proposed a combined method, where the sources are separated by a FD-ICA approach using the Ikeda [17] algorithm, while the correct permutations are lined up to give a consistent DOA for the sources.

All these solutions try to solve the permutation problem imposing *frequency coupling in the channel*. An alternative approach can be to impose *frequency coupling in the source model*, as explained in the next section.

III. A fast frequency domain ICA framework

A. A time-frequency source model

In Lee's approach [19], the signals are modelled in the time-domain (the tanh nonlinearity is applied in the time-domain). On the other hand, Smaragdis proposed modelling the signals in the frequency-domain. Below, we argue that a *time-frequency model* is more appropriate.

If we examine the statistical properties of an audio signal over shorter quasi-stationary periods in the time-domain (frames of the STFT), the signal is not always well modelled as supergaussian. Looking at the statistical properties of these segments in the frequency domain, they can be better modelled as supergaussian, as these sections have very heavy tailed distributions [10]. Figure 1 exhibits the histograms of some audio signal segments in the time-domain and the histograms of the real part of prewhitened FFT of these segments.

This implies the frequency domain is a better candidate for source modelling. This will provide a better achievable performance, since as noted by various authors (e.g. [4]), the *Cramer-Rao bound* is related to how close the source distributions are to the Gaussian. That is that the more nonGaussian the distributions are, the better the achievable performance.

In addition, most of the supergaussianity measured in the time domain comes from the fluctuating amplitude of the audio signal. The *slowly varying amplitude* profile also gives us valuable information that can be exploited for source separation and is not affected by the permutation problem. Therefore, we can exploit this property to introduce frequency coupling within the STFT structure.

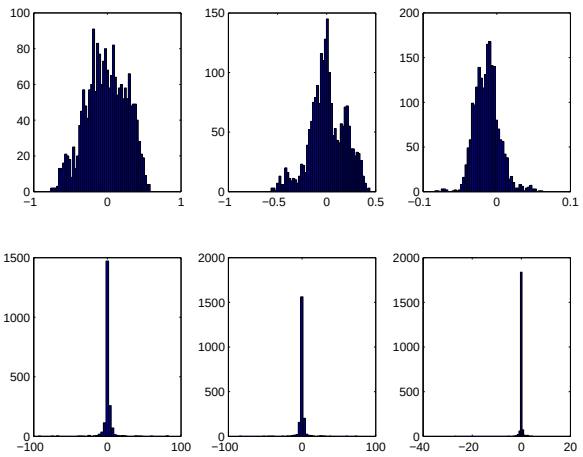


Fig. 1. Exploring the statistical properties of short audio segments. Histograms in the time domain (first row) and in the prewhitened Frequency domain (second row).

Motivated by this, we introduce the following *time-frequency model*. We will generally assume that the STFT coefficients of the separated sources follow an exponential non-gaussian distribution. However, we want our model to incorporate some information about the scaling of the signal with time (i.e. the signal envelope), assuming that it is approximately constant over the analysis window. This can be modelled by a *nonstationary time varying* scale parameter β_k.

$$p(u_k(f,t)) \propto \beta_k(t)^{-1} exp(-h(u_k(f,t)/\beta_k(t))) \qquad (11)$$

where $h(u)$ defines the density's form, the index t represents the time-frame index, f the frequency bin and k is the source index. The key feature is that the β_k term *is not a function of frequency*. This restriction provides us with sufficient coupling between frequency bins to break the *permutation ambiguity*. The β_k term can be interpreted as a volume measurement. Literally, it measures the overall signal amplitude along the frequency axis, emphasising the fact that one source is louder at a certain time slot. This loudness indication can force alignment of the permutations along the frequency axis.

The next step would be to see how the proposed time-frequency model alters the natural gradient algorithm in (8). Effectively, the source model is represented by the *activation function* $\phi(u)$. More specifically, we have :

$$\phi(u) = -\frac{\partial}{\partial u} \log p(u) = -\frac{p'(u)}{p(u)} \qquad (12)$$

The proposed model gives the following activation function:

$$\phi(u_k(f,t)) \propto \beta_k(t)^{-1} h'(u_k(f,t)/\beta_k(t)) \qquad (13)$$

The natural gradient algorithm is altered as follows:

$$\Delta W_f = \eta(I - \beta(t)^{-1}\mathcal{E}\{g(\underline{u}(f,t))\underline{u}^H(f,t)\})W_f \qquad (14)$$

where $\beta(t) = diag(\beta_1(t), \beta_2(t), \ldots, \beta_N(t))$, $g(u) = h'(u)$ and η is the learning rate. The value for $\beta_k(t)$ is estimated adaptively from the separated signals $\underline{u}(f,t)$.

We note that care needs to be taken in defining activation functions for complex variables. Below, we will consider activation functions of the form $(u/|u|)f(|u|)$. Although a variety of other activation functions are valid, such as $g(u) = \tanh(\Re\{u\}) + j\tanh(\Im\{u\})$, proposed by Smaragdis [26], it seems more intuitive to impose no preference on the phase angles. That is to introduce circularly symmetric priors on complex variables. This is essentially the same as the priors on subspaces as proposed by Hyvarinen et al in *Independent Subspace Analysis* (ISA) [15]. Assuming complex Laplacian priors in the form of $p(u) \propto exp(-|u|) \Rightarrow h(u) = |u|$, we set $f(|u|) = 1$. The activation function in (14) is then the following:

$$g(u) = u/|u|, \qquad \forall |u| \neq 0 \tag{15}$$

Although the discontinuity due to $|u|$ implies the cost function will not be smooth at certain points, in practice, the performance of the algorithm appears to be unaffected. MacKay [20] also supported that the function in (15) can have the same robustness property as the tanh function. Alternatively, we could use a "smoothed" Laplacian prior $p(u) \propto \exp(-|u| + \log|u|)$, as proposed by Zibulevsky [29].

Assuming complex Laplacian priors, we can use the following estimate for $\beta_k(t)$:

$$\beta_k(t) = \frac{1}{N}\sum_f |u_k(f,t)| \tag{16}$$

B. Permutation Problem Revisited

Let us now investigate the effect of this time-frequency model upon the permutation symmetries. Without the $\beta(t)$ term the log likelihood function has an identical minimum for every permutation of the sources at each frequency. Incorporating β, we weight the likelihood of an unmixing matrix at a given frequency with the time envelope induced by the components at other frequencies. Thus, β allows the matching of time envelopes, providing us with a discriminator for the different permutations.

Nonetheless, a direct application of (14) and (16) does not guarantee that the correct permutation will be found. The β term will break the symmetry, however, it will not necessarily change the cost function enough to completely remove spurious minima. Thus, a gradient optimisation scheme is likely to get trapped in a local minimum. This may explain the poor performance of Parra's solution [22] in certain examples as observed by Ikram et al [18].

In the following analysis, we introduce a post processing mechanism in the algorithm by which the correct permutations are sorted. Fortunately, due to the symmetry of the problem if we know where one minimum is, we know where they all are. It is therefore possible to introduce a jump step into the update that chooses the permutation that is most likely.

Here we describe a solution for $N = 2$, using the Laplacian prior. Suppose that for a given set of known W_f, $\beta_k(t)$ and $\underline{u}(f,t) = W_f\underline{x}(f,t)$, we wish to compare two possible choices for source estimates of u:

$$1.\begin{bmatrix} \gamma_{11} & 0 \\ 0 & \gamma_{22} \end{bmatrix}\tilde{u}(f,t) = u(f,t) \tag{17}$$

$$2.\begin{bmatrix} 0 & \gamma_{12} \\ \gamma_{21} & 0 \end{bmatrix}\tilde{u}(f,t) = u(f,t) \tag{18}$$

where γ_{ij} are rescaling parameters that account of incorrect scaling. To compare these two possibilities, we will evaluate their likelihood over T time frames.

$$1.\log p(u|\gamma_{11},\gamma_{22}) = -T\log(\gamma_{11},\gamma_{22}) + \log p(\tilde{u}) \tag{19}$$

$$2.\log p(u|\gamma_{12},\gamma_{21}) = -T\log(\gamma_{12},\gamma_{21}) + \log p(\tilde{u}) \tag{20}$$

with the values of γ_{ij} chosen to maximise the likelihood. For the Laplacian model these are:

$$\gamma_{ij} = \frac{1}{T}\sum_t \frac{|u_i(f,t)|}{\beta_j(t)} \tag{21}$$

We can now evaluate the likelihood of the estimated $u(f,t)$ in terms of the known quantities $u(f,t)$ and γ. For case 1, we have:

$$\log p(\tilde{u}) \propto -\gamma_{11}^{-1}\sum_t \frac{|u_1(f,t)|}{\beta_1(t)} - \gamma_{22}^{-1}\sum_t \frac{|u_2(f,t)|}{\beta_2(t)} \tag{22}$$

which reduces to $\log p(\tilde{u}) \propto -2T$. The analysis for case 2 is identical. Therefore, we get:

$$\log\frac{p(\text{“}case1''})}{p(\text{“}case2''})} = -T\log(\gamma_{11}\gamma_{22}) + T\log(\gamma_{12}\gamma_{21}) \tag{23}$$

and we can form the following *likelihood ratio test* (LR):

$$LR = \frac{p(\text{“}case1''})}{p(\text{“}case2''})} = \frac{\gamma_{12}\gamma_{21}}{\gamma_{11}\gamma_{22}} \tag{24}$$

If $LR < 1$, we permute W_f before proceeding. This likelihood ratio test is performed after calculating the update ΔW_f, lining up permutations not sorted by the gradient step.

There are basically *two drawbacks* in this approach. Firstly, this becomes *more complicated for more than 2 sources*, although one possible solution would be to consider the sources in a pairwise fashion. Secondly, the algorithm has to work *only in batch mode*, as usage of a one-sample likelihood is not possible. On the other hand, the algorithm seems to perform well in the majority of cases.

A similar approach, using the time-frequency envelope to remove the permutation ambiguity, was proposed by Ikeda et al [17]. However, the approach presented in this section is more probabilistically justified and ends up proposing a different solution.

C. A fast Frequency Domain ICA algorithm

So far, we have only considered a gradient-based optimisation scheme to produce maximum likelihood (or MAP) estimates of the original audio sources. However, all

gradient-based optimisation methods have two major drawbacks. First of all, they *converge relatively slowly*. For a common frequency domain ICA scenario, we found that the natural gradient would require around 500 updates to each W_f (iterations) on average for some decent separation quality. Secondly, *their stability* depends on the *choice of the learning rate*. In addition, natural signals have greater low frequency values; therefore the time-frequency values tend to have different signal levels for every frequency bin. Inevitably, keeping a constant learning rate for all learning procedures may inhibit the separation quality at certain frequency bands. This may also give a reason why the natural gradient approach does not perform well at high frequencies, as observed by Smaragdis [26].

For these reasons, we want to replace the natural gradient scheme in the FD-ICA framework with a Newton-type optimisation scheme. Their basic feature is that they converge much faster than gradient algorithms with the same separation quality and while they are more computationally expensive, the number of iterations for convergence is sufficiently decreased. In addition, they tend to be much more stable. Hyvarinen et al [3], [16], [14] introduced several types of Newton-type "fixed-point" algorithms in ICA of instantaneous mixtures, using kurtosis or negentropy.

In [14], Hyvarinen explored the relation between a generalised "fixed-point" (approximate Newton method) ICA algorithm with the maximum likelihood ICA approach on instantaneous mixtures. In the following analysis, we show that it is elementary to extend the algorithm proposed in [14] to be applicable to the proposed time-frequency framework.

In the ML-ICA approach for instantaneous mixtures, we form and try to maximise the following likelihood with respect to the unmixing matrix W:

$$F = \log L(\underline{x}|W) = \mathcal{E}\{\log p(\underline{u})\} + \log|det(W)| \qquad (25)$$

Performing *gradient ascent*, we can derive the Bell-Sejnowski [2] algorithm.

In [14], Hyvarinen tries to solve the following optimisation problem:

$$\max_W G(W\underline{x}) \qquad (26)$$
$$\text{subject to } \mathcal{E}\{\underline{u}\underline{u}^T\} = I$$

where $G(u)$ is a non-quadratic function. The solution for this problem can be estimated by finding the maximum of the following function:

$$K(W) = G(W\underline{x}) - \alpha(\mathcal{E}\{\underline{u}\underline{u}^T\} - I) \qquad (27)$$

where α is the *Lagrange multiplier*. Performing a gradient ascent on $K(W)$, we get:

$$\nabla K = \mathcal{E}\{G'(W\underline{x})\underline{x}^T\} - \alpha CW \qquad (28)$$

where $C = \mathcal{E}\{\underline{x}\underline{x}^T\}$. If we choose $G(\underline{u}) = \log p(\underline{u})$, then this update law is almost identical to the Bell-Sejnowski law with a different term controlling the scaling of the unmixing matrix W. This implies that the algorithm in (28) can be viewed as solving a *constrained Maximum Likelihood problem*. After a series of steps (see [14]) and using $G(\underline{u}) = \log p(\underline{u})$, we end up to the following learning rule:

$$\Delta W = D[diag(-\alpha_i) + \mathcal{E}\{\phi(\underline{u})\underline{u}^T\}]W \qquad (29)$$

where $\alpha_i = \mathcal{E}\{u_i\phi(u_i)\}$, $D = diag(1/(\alpha_i - \mathcal{E}\{\phi'(u_i)\}))$. This algorithm converges at a substantially faster rate than the gradient based update rules.

Comparing the update rule in (29) with the original natural gradient law, we can see that they are similar. Instead of a constant learning rate, there is a learning rate (the D matrix) that adapts to the signal. Hence, the algorithm is less dependent on signal levels and therefore more stable. Hyvarinen states that replacing I with the adaptive term $diag(-\alpha_i)$ is also beneficial for convergence speed. If we use pre-whitened data $\underline{x}$, then the formula in (29) is equivalent to the original fixed-point algorithm [16], while it is still expressed in terms of the natural gradient algorithm. The most important consequence for us, however, is that the nonlinear activation function $\phi(u)$ in (29) has exactly the same interpretation as in the ML-approach, as mentioned in section IIB.

D. A unifying framework

We can now use all the previous analysis to form a unifying framework for the convolutive mixtures problem.

First of all, we prewhiten the time-frequency STFT coefficients of the mixtures $\underline{x}(f,t)$ and store the prewhitening matrices V_f for each frequency bin. Prewhitening is an essential step for the algorithm, which cannot be omitted.

The next step is to estimate the unmixing matrix for each frequency bin. We will use the "fixed-point" approach, as described in (30), using random initialisation for W_f. Moreover, he have to keep the rows of W_f orthogonal with unit norm, as described in (31).

$$\Delta W_f = D[diag(-\alpha_i) + \mathcal{E}\{\phi(\underline{u}(f,t))\underline{u}^H(f,t)\}]W_f \qquad (30)$$

$$W_f \leftarrow W_f(W_f^H W_f)^{-0.5} \qquad (31)$$

The parameters in this update rule are calculated as previously. In addition, we will use the proposed time-frequency source model, as described earlier on, to impose frequency coupling. Therefore, the activation function $\phi(u)$ in (30) is:

$$\phi(\underline{u}) = \beta^{-1}(t)\underline{u}/\underline{u}| \qquad \forall u \neq 0 \qquad (32)$$

The derivative $\phi'(u)$ used in the calculation of D can be approximated by:

$$\phi'(\underline{u}) = \beta^{-1}(t)(|\underline{u}|^{-1} - \underline{u}^2|\underline{u}|^{-3}) \qquad \forall u \neq 0 \qquad (33)$$

The next step is to remove the *permutation ambiguity* by applying the *likelihood ratio jump* solution.

An important issue is the *spectral shape ambiguity*. In [6], Cardoso shows that we can remove this ambiguity, by focusing on the observation spaces containing each source rather on the columns of the mixing matrix. Therefore,

he proposed an operator that projects the components to the observation space to solve the problem. In our case, we will use this operator to return the separated sources to the observation space. Thus, we will have N estimates of each source. Denoting the estimated unmixing matrix as W_f, the prewhitening matrix as V_f for each frequency bin f, then the separated sources, *observed at each microphone*, are given by:

$$\tilde{s}_{i,x_j}(f,t) = [V_f^{-1}W_f^{-1}]_{ji}u_i(f,t),\ \forall i,j = 1,\ldots,N \quad (34)$$

where $\tilde{s}_{i,x_j}$ is the *i-th* estimated source observed at the *j-th* microphone.

IV. Evaluation

It is not our intention to provide an exhaustive comparison of the many different approaches to BSS with convolutive mixtures. Instead, we present two experiments to demonstrate that the proposed fast FD-ICA framework can produce fast and good quality separation, providing a robust solution for the permutation problem. Other results have been reported elsewhere [21].

A. Experiment 1

In our initial experiment, we created a synthetic convolutive mixture of two speech sources (around 3secs at 16KHz) that illustrates the permutation problem in the Smaragdis algorithm. The synthesised acoustic paths consisted of an initial delay followed by single echo. The echo times were between 1 and 5 milliseconds and echo strengths between 0.1 and 0.5 of the direct path signal.

Spectrograms of the separated sources are given in figure 2 along with equivalent separations for the Smaragdis algorithm. It is clear that the permutation inconsistencies that occurred in the Smaragdis case are no longer present. Omitting the LR step in our algorithm seems to produce the same permutation errors as in Smaragdis's case. In both cases, the frame size was 2048 samples (approximately 150ms) with a frame overlap of 50%. However, while Smaragdis's algorithm required about 500 iterations to reach convergence, the fast FD-ICA framework required only 50 (both algorithms processed the data in batch form). This is very typical of the dramatic improvement in efficiency that can be achieved using Fast ICA techniques.

B. Experiment 2

The second experiment was chosen to test the algorithm's ability in highly reverberant conditions. To do this, we used Westner's room acoustic data. Westner [13] placed a number of microphones and loudspeakers in a conference room with bad acoustics and measured the transfer function between each speaker and microphone position. Using his *roommix* function, one can simulate any of the measured speaker-microphones configurations in that conference room, generating a very challenging data set. For our experiment, we placed our sources to speaker positions 1 and 2 and we used microphones 2 and 1 to capture the auditory scene, according to Westner's configuration [13].

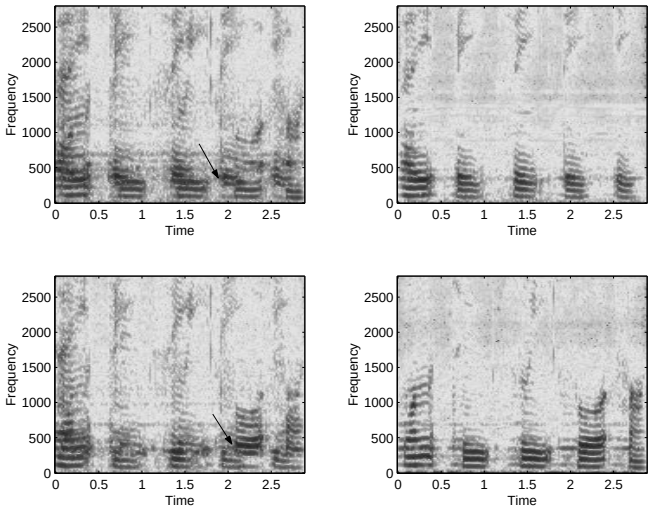


Fig. 2. Permutation problem illustrated. Separated sources using the Smaragdis algorithm (left) and the algorithm proposed in section III (right). Permutation inconsistencies are highlighted with arrows.

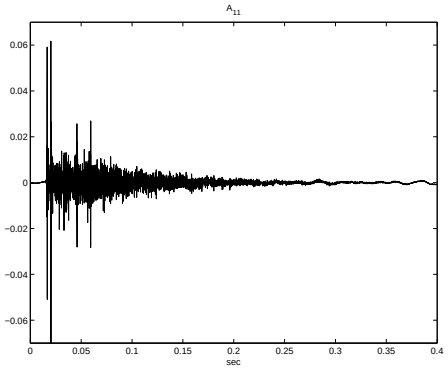


Fig. 3. One of the filters modelling the room acoustics created by Westner's *roommix* function.

An example of the simulated room impulse responses used in this experiment is depicted in figure 3. The room acoustics have substantial reverberation for several hundred milliseconds and therefore this experiment is expected to be very challenging.

We applied the algorithm to speech data (around 7secs at 16KHz), using a STFT frame size of around 500 msecs with 75% overlapping and a hamming window. The fast FD-ICA algorithm managed to reduce the crosstalk by a considerable amount. Choosing a long frame length is inevitable, as it needs to be much greater than the mixing filters length, so that the convolution is actually transformed into multiplication in the frequency domain. The fact that reverberation continued beyond the frame length means that the transfer function cannot be perfectly modelled.

One drawback of our current approach is that we are attempting to reconstruct the signals at the microphones. Thus, the reverberation is still present on the separated sources. One possible solution to this problem has recently been proposed in [27].

C. Performance Measurements

To quantify the performance of our fast implementation and compare it against a natural gradient update scheme, we measured the *Improvement in Signal-to-Noise Ratio* (ISNR) achieved at each microphone. This metric is also referred to as *Noise Reduction Rate* (NRR) in [24]. Note that ISNR can be used as a performance metric, as the sources are observed at the microphones, therefore there is no scale ambiguity.

$$ISNR_{i,j} = 10 \log \frac{\mathcal{E}\{(s_{i,x_j}(t) - x_j(t))^2\}}{\mathcal{E}\{(s_{i,x_j}(t) - \tilde{s}_{i,x_j}(t))^2\}} \qquad (35)$$

where x_j is the mixed signal at the *j-th* microphone, $\tilde{s}_{i,x_j}$ is the *i-th estimated* source observed at the *j-th* microphone and s_{i,x_j} is the *i-th original* source observed at the *j-th* microphone. The ISNR results for the experiments described above are presented in table I. These clearly demonstrate the superiority of the fast learning algorithm when faced with a challenging acoustical environment.

In figure 4, we compare the performance of the fast FD-ICA framework with the natural gradient (NG) algorithm in the Westner case. We can see the improvement in convergence speed and separation quality. In this plot, we can also see that the actual speed of the proposed framework, as it converges in around 20 iterations.

We can also measure the *distortion* along the frequency axis, as proposed by Schobben et al [25].

$$D_{i,j}(f) = 10 \log \frac{\mathcal{E}\{\mathrm{STFT}\{(s_{i,x_j}(t) - \lambda_{ij}\tilde{s}_{i,x_j}(t))^2\}\}}{\mathcal{E}\{\mathrm{STFT}\{s_{i,x_j}(t)^2\}\}} \qquad (36)$$

where $\lambda_{ij} = \mathcal{E}\{s_{i,x_j}(t)^2\}/\mathcal{E}\{\tilde{s}_{i,x_j}(t)^2\}$. In figure 5, we plot $D_{1,1}$ and $D_{1,2}$ for Exp. 2 using fast FD-ICA along frequency. We can see that the distortion remains negative along the greatest part of the spectrum (significantly lower compared to the NG approach), except from some specific frequency bands, where the signal energy levels are low (high frequencies), or the BSS problem is ill-determined.

TABLE I

ISNR (dB) MEASUREMENTS FOR THE FAST FD-ICA METHOD (AFTER 50 ITERATIONS) AND THE NATURAL GRADIENT ALGORITHM (AFTER 500 ITERATIONS).

	$ISNR_{1,1}$	$ISNR_{2,1}$	$ISNR_{1,2}$	$ISNR_{2,2}$
Exp.1 FastFD-ICA	8.02	3.92	6.79	4.85
Exp.1 Nat.Grad.	5.33	1.21	4.92	2.40
Exp.2 FastFD-ICA	4.19	3.09	4.18	3.40
Exp.2 Nat.Grad.	3.18	2.34	3.87	2.17

D. Computational Cost

The computational cost of the Fast FD-ICA framework is slightly increased, compared to the NG framework. We have to consider the extra cost introduced by the fast algorithm and the Likelihood-Ratio jump. In terms of *floating point operations*, the "fixed-point" algorithm requires 1.45 times more *flops* per iteration than the NG algorithm. Including the LR jump, it requires 2.02 times more flops per iteration. Considering that the new framework converges in 10-30 times less iterations, we can see the overall gain in computational cost and convergence speed. However, the computational cost of the LR jump increases significantly with N sources. Working on a pairwise basis, the cost of the LR jump will scale quadratically with N. Prewhitening also increases the computational cost, however, it is preformed only once before the main algorithm and its cost is 0.3 times the cost of an iteration.

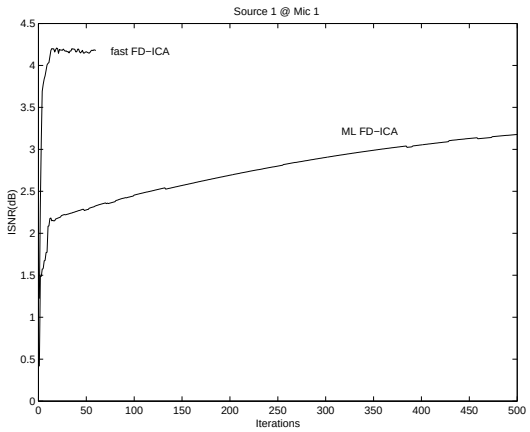


Fig. 4. Comparison of the fast FD-ICA algorithm with the natural gradient approach in the Westner case.

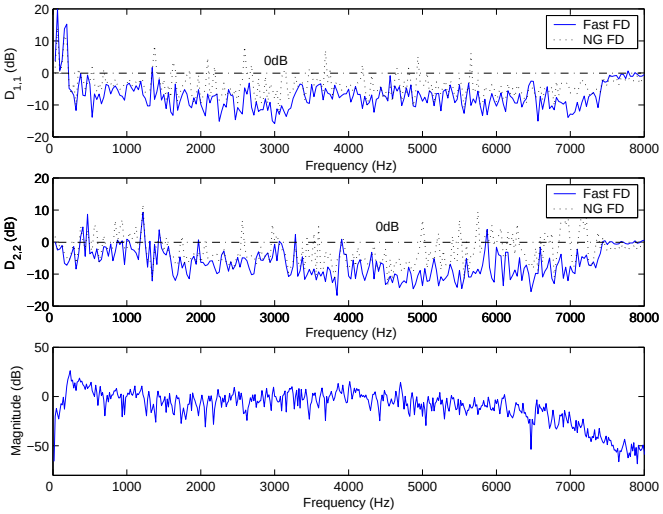


Fig. 5. Measuring distortion along frequency for the NG FD-ICA and Fast FD-ICA case, together with signal energy levels along frequency (bottom).

V. Conclusions - Future Work

In this study, we explored the Blind Source Separation problem of convolved mixtures, in the case of equal number of sources and sensors, proposing a fast frequency-domain

solution. The key points of this solution were:

Firstly, we proposed a new *time-frequency source model* for a ML-ICA approach, incorporating a *time-varying* parameter, aiming to model the audio signals more effectively and impose *frequency coupling* between neighboring frequency bins. In addition, a method to tackle the permutation problem was proposed, by incorporating a *likelihood ratio* test at each iteration.

In addition, a fast Newton-type ICA algorithm was adapted in the frequency domain framework, replacing the natural gradient ICA algorithm. As a result, the speed of the FD-ICA framework increased by an order of magnitude. Incorporating the proposed solution for the permutation problem in the "fast ICA" implementation produced good separation results. However, this is a *batch* algorithm and is not directly applied to real-time systems.

In the future, the authors would like to investigate a more complete Bayesian solution, incorporating strong priors on the channel model. Recent work exploiting DOA estimates to solve the permutation problem [24] suggests that such a model could be very effective. Furthermore, introducing higher-level priors on the source model can improve separation quality, or even achieve separation of specific sources. This could include incorporating source identity or harmonic models similar to those used in *Computational Auditory Scene Analysis* (CASA).

Acknowledgments

N. Mitianoudis is supported by the Elec. Eng. Department, QMUL. The authors would also like to thank the reviewers for their insightful comments that helped improve the paper.

References

[1] Amari S., Cichocki A., Yang H. H., *A new learning algorithm for blind source separation*, Advances in Neural Information Processing Systems, pp. 757-763, MIT Press, Cambridge MA, 1996.

[2] Bell A.J., Sejnowski T.J., *An information-maximization approach to blind separation and blind deconvolution* , Neural Computation, 7, pp. 1129-1159, 1995.

[3] Bingham E. and Hyvarinen A., *A fast fixed-point algorithm for independent component analysis of complex-valued signals*, Int. J. of Neural Systems, 10(1):1-8, 2000.

[4] Cardoso J.F., *Blind signal separation: statistical principles*, Proc. of the IEEE, vol. 9, no 10, pp. 2009-2025, Oct. 1998.

[5] Cardoso J.F., *Infomax and maximum likelihood for source separation*, IEEE Signal Processing Letters, no 4, pp. 112-114, 1997.

[6] Cardoso J.F., *Multidimensional independent component analysis*, Proc. ICASSP, Seattle, WA, 1998.

[7] Cardoso J.F., Souloumiac A., *Blind beamforming for nonGaussian signals*, IEE Proc., 140(6):362-370, 1993.

[8] Christensen K.B., *The application of Digital Signal Processing to Large-Scale Simulation of Room Acoustics: Frequency Response Modelling and Optimization Software for a Multichannel DSP Engine*, Journal of the AES, 40, pp. 260-276 April 1992.

[9] Comon P., *Independent Component Analysis, a new concept?* , Signal Processing, Elsevier, 36(3):287–314, April 1994

[10] Davies M., *Audio Source Separation*, Mathematics in Signal Processing V, 2000.

[11] Godsill S.J., Rayner P. J. W., and O. Cappe, *Digital audio restoration*, Applications of Digital Signal Processing to Audio and Acoustics, pp. 133-193, Kluwer Academic Publishers, 1998.

[12] Hopgood J.R., Rayner P.J.W., Yuen P.W.T., *The effect of sensor placement in Blind Source Separation*, IEEE WASPAA, New Paltz, New York, October 2001.

[13] http://www.media.mit.edu/∼westner/

[14] Hyvarinen A., *The Fixed-Point Algorithm and Maximum Likelihood Estimation for Independent Component Analysis*, Neural Processing Letters, 10(1):1-5.

[15] Hyvarinen A., Hoyer P., *Independent Subspace Analysis shows emergence of phase and shift invariant features from natural images*, Int. Joint conference on Neural Networks, 1999.

[16] Hyvarinen A., Oja E. *A fast fixed-point algorithm for independent component analysis*, Neural Comp., 9(7):1483-1492, 1997.

[17] Ikeda S., Murata N., *A method of ICA in Time-Frequency Domain*, Int. Conf. on ICA and Signal Separation, pp 365-371, Aussois, France, Jan 1999.

[18] Ikram M., Morgan D., *Exploring permutation inconsistency in blind separation fo speech signals in a reverberant environment*, ICASSP, Instanbul, 2000

[19] Lee T.W. , Bell A.J., Lambert R., *Blind separation of delayed and convolved sources*, Advances in Neural Information Processing Systems 9, MIT Press, Cambridge MA, pp 758-764, 1997.

[20] MacKay D.J.C., *Maximum Likelihood and Covariant Algorithms for Independent Component Analysis*, Draft paper available from *http://www.inference.phy.cam.ac.uk/mackay/*.

[21] Mitianoudis N., Davies M., *New fixed-point solutions for convolved mixtures*, 3rd Int. Conf. on ICA and Source Separation, San Diego, December 2001.

[22] Parra L., Spence C., *Convolutive blind source separation of non-stationary sources*, IEEE Trans. on Speech and Audio Processing, pp. 320-327, May 2000.

[23] Pearlmutter B.A., Parra L.C., *Maximum Likelihood Blind Source Separation: A Context-Sensitive Generalization of ICA*, in Advances in Neural Information Processing Systems, vol. 9, pp. 613-619, 1997.

[24] Saruwatari H., Kawamura T., Shikano K., *Fast-convergence algorithm for ICA-based blind source separation using array signal processing*, IEEE WASPAA, New Paltz, New York, October 2001.

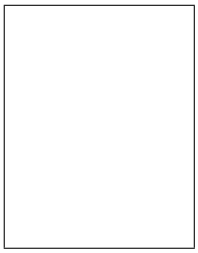
[25] Schobben D.,Torkkola K., Smaragdis P. *Evaluation of Blind Singal Separation Methods*, Inter. Conf. on ICA and Source Separation, Aussois, France, pp. 261-266, 1999.

[26] Smaragdis P., *Information Theoretic Approaches to Source Separation*, MSc thesis, MIT Media Lab, June 1997.

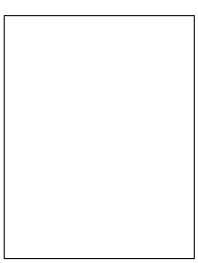
[27] Sun X., Douglas S. *A natural gradient convolutive Blind Source Separation algorithm for speech mixtures*, 3rd Inter. Conf. on ICA and Source Separation, San Diego, December 2001.

[28] Torkkola K., *Blind separation of convolved sources based on information maximisation*, IEEE workshop on Neural Networks and Signal Processing, Kyota, Japan, 1996.

[29] Zibulevsky M, Pearlmutter B., *Blind source separation by sparse decomposition*, 2nd Int. Conf. on ICA and Source Separation, Helsinki, August 2000.



Nikolaos Mitianoudis is a PhD student at Queen Mary, University of London, UK. In 1998, he received a diploma in Electronic and Computer Engineering from the Aristotle University of Thessaloniki, Greece. In 2000, he received a MSc in Communications and Signal Processing from Imperial College, University of London, UK. Currently, he is working towards his PhD in audio source separation using independent component analysis.



Mike Davies is a lecturer at Queen Mary , University of London, UK. He received a B.A. degree in Engineering from Cambridge University in 1989 and a Ph.D. in nonlinear dynamics and signal processing from University College London in 1993. In 1993 he was awarded a Royal Society University Research Fellowship, held at UCL and Cambridge University. In 2001, he co-founded the DSP research group in Queen Mary, University of London. His research interests include: nonlinear dynamics, audio processing, statistical signal processing, array signal processing and machine learning.