INTERNATIONAL JOURNAL OF ADAPTIVE CONTROL AND SIGNAL PROCESSING
*Int. J. Adapt. Control Signal Process.* 2002; **00**:1–6 *Prepared using* acsauth.cls *[Version: 2002/11/11 v1.00]*

# Audio Source Separation: Solutions and Problems

Nikolaos Mitianoudis*, Mike Davies

*Electronic Engineering Department, Queen Mary, University of London, Mile End Road, London E1 4NS, UK*

SUMMARY

The problem of separating out a number of audio sources observed from an array of microphones in a real room environment has received a great deal of attention in the past decade. While there are now a number of workable methods that can even deal with relatively high reverberation [18], a number of interesting problems still remain. In this paper, the authors review the methods based around *Independent Component Analysis* (ICA), discussing the various choices available in algorithm design. We then explore the issue of sensitivity to speaker movement which appears to impose fundamental limitations on BSS performance.

KEY WORDS: Audio source separation, subband filtering, array processing, Independent Component Analysis, beamforming

## 1. Introduction

*Audio source separation* is the problem of automated separation of audio sources present in a room, using a set of differently placed microphones, capturing the auditory scene. The whole problem resembles the task a human can solve in a cocktail party situation, where using two sensors (ears), the brain can focus on a specific source of interest, suppressing all other sources present (*cocktail party problem*). Unmixing the auditory scene into its audio objects can have many applications. First of all, it will be possible to re-render the auditory scene using more speakers or different source placements. For audio recordings, having the audio objects unmixed, we will be able to remix the recording. For audio coding, we will be able to encode and compress each instrument seperately, according to the MPEG4 philosophy of audiovisual objects, increasing compression. It might also act as a denoising tool in the case of mobile phones, or other capturing devices. Last but not least, it will be a very valuable tool for music transcribers or automated music transcription and other monitoring devices.

The problem we will consider falls into the broad category of *Blind Source Separation* (BSS) and can be described as follows. Suppose there are $N$ audio sources in a room $\underline{s}(n) = [s_1(n), \ldots, s_N(n)]^T$ and the auditory scene is captured by $M$ microphones $\underline{x}(n) =$

*Correspondence to: Electronic Engineering Department, Queen Mary, University of London, Mile End Road, London E1 4NS, UK

*Received*

*Revised*

$[x_1(n), \ldots, x_M(n)]^T$. We can model the outputs of the microphones each as a filtered mixture of the sources with the possible addition of background noise $\epsilon_i(n)$, as follows:

$$x_i(n) = \sum_{j=1}^{N} \alpha_{ij}(n) * s_j(n) + \epsilon_i(n) \qquad i = 1, \ldots, M \tag{1}$$

where $\alpha_{ij}$ is a high-order FIR filter that models the *room transfer functions* between the $i^{th}$ sensor and the $j^{th}$ source. An alternative formulation would be to use IIR filters to model the room acoustics. However this introduces stability issues and restricts us to considering only minimum-phase mixing [17]. We will therefore concentrate on FIR channel modelling which is predominantly used in adaptive audio processing [8].

The blind source separation problem aims to produce source signal estimates, $\underline{\tilde{s}}(n)$, as a function of the observed mixtures $\underline{x}(n)$. The term *blind* stresses that little is known about the underlying signal and channel properties. We will see however that many of the choices to be made in audio source separation algorithm design involve introducing additional information about the structure of the problem. Such methods are often termed *semi-blind*.

In this study we will consider source estimates that are linear combinations of the mixture signals:

$$\tilde{s}_i(n) = \sum_{j=1}^{M} w_{ij}(n) * x_j(n) \qquad i = 1, \ldots, N \tag{2}$$

where $w_{ij}$ represent an unmixing FIR filter structure. The basic tool that will be used to adapt the unmixing filters will be *Independent Component Analysis* (ICA) [12]. ICA is a method for performing blind source separation from linear (instantaneous) mixtures. The technique assumes statistical independence between the sources and allows at most one Gaussian component. Similar methods for ICA have been developed from a number of different view points: minimising *Kullback-Leibler* (KL) divergence [1], *Infomax* [4] or *Maximum Likelihood* estimation [5, 22]. Other methods look for the directions of the most nonGaussian components, using *kurtosis* or *negentropy* as nonGaussianity measures [12]. Other approaches estimate the unmixing matrix by performing approximate diagonalization of a *cumulant tensor* of the mixtures [7], or even nonlinear decorrelation [2]. Finally, there are methods that exploit only second-order statistics (and nonstationarity) and perform separation by multiple decorrelation at the receiver [21, 14].

Applications of ICA to the audio source separation problem involve a number of options associated with the structure of the adaptive filter and the domain of the source models. Here we will review some of these choices along with various problems that can arise. We will restrict ourselves to the choices of the general framework rather than go into the detailed choice of different ICA algorithms. While we believe that there is still much work to be done in this area, there now exist many applicable ICA algorithms with comparable performance (e.g.[1, 25, 21, 18, 12, 7]). For the experimental results in this paper, we have used the fast ICA algorithm presented in [18]. The computational complexity of this frequency domain approach is evaluated in [18] and compared to the original natural gradient approach.

Specifically, in this study, we will consider the following aspects of the audio source separation problem. The first issue that arises is the choice of the *unmixing domain*. One solution is to perform the unmixing in the time domain [26], by direct adaptation of the unmixing FIR coefficients as stated by equation 2. However a more efficient solution is to work within a

*subband framework* (e.g. using the Discrete Fourier Transform), allowing us to approximate convolution as multiplication in the subbands [17, 25, 19]. As with more traditional adaptive subband filtering the choice of the subband structure will affect the level of aliasing present in the solution.

The next key decision is the choice of the *source modelling domain*. Most, though not all (e.g. [21]), audio source separation algorithms have exploited the non-Gaussianity of the sources' sample statistics. To do this we need to initially select a domain within which to apply our nonGaussian assumptions. Typically this would either be the time domain [17, 19], or the frequency domain [25, 10, 18]: see Figures 1 and 2. Working in the frequency domain is computationally simpler and typically produces a sparser representation that enables better separation performance [18]. We discuss this choice in more detail in Section 3 and identify the various trade-offs involved.

One consequence of modeling the sources in the frequency domain is that it introduces an additional complication: *the permutation problem*. That is: when separating the mixtures in each subband independently we cannot guarantee that the ICA outputs will have the same permutation of separated sources along the frequency axis. We therefore need to introduce an additional step or constraint to correctly align the permutations. As the frequency domain modelling approach to ICA is a popular one there has been a great deal of effort put into solving this. We will examine the options available in Section 4.

It should be stressed that theoretically the permutation problem is not an inherent problem in covolutional blind signal separation (see for example [9] in this issue). It is an artefact of source modelling in the frequency domain. That said, there is also some numerical evidence, [20], that time domain cumulant criteria are not robust to incorrect permutations when applied to audio source separation. However for the remainder of this paper we will only consider the permutation problem as it manifests itself in the frequency domain.

It should be stressed that the permutation problem is not an inherent problem of convolutive source separation. It is an artefact of source modelling in the frequency domain. As such it does not occur when using time domain source models (e.g. [17]).

One of the interesting insights that has emerged from exploring the solutions of the permutation problem is the interpretation of BSS as a broadband adaptive beamformer. In the final section of this paper we explore this analogy and identify a key limitation of blind audio source separation problems in general. Namely the effect on separation quality of parameter mis-alignment. This sensitivity is a function of frequency and is likely to limit the mid to high frequency performance in any real world environment.

## 2. Unmixing domain choice

One of the first considerations of the source separation problem is the choice of the unmixing domain, i.e. the domain of adaptive filtering. A great part of the proposed solutions prefer not to work in the *time-domain*, the main reason being the computational cost of the convolution. As a result, filtering adaptation and implementation are performed in the *frequency domain*, mainly due to the following property of the Fourier Transform:

$$x(n) = \alpha(n) * s(n) \rightleftharpoons X(f) = A(f)S(f) \tag{3}$$

where $n$ represents the time index, $f$ frequency index and $*$ represents *linear convolution*.

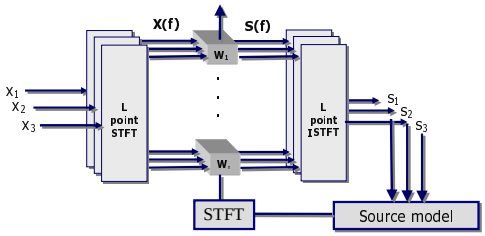


Figure 1. Unmixing in the frequency domain, but source modeling in the time domain (Lee et al solution).

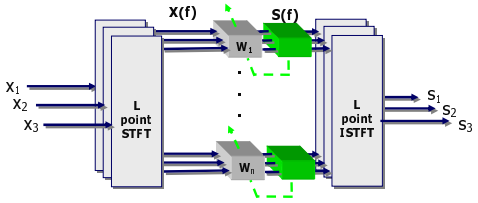


Figure 2. Unmixing and source modeling in the frequency domain (Smaragdis solution).

One can approximate the linear convolution, as a circular convolution and therefore approximate this property on equation 1 by applying the short time Fourier transform (STFT) giving:

$$X_i(f,t) \approx \sum_{j=1}^{N} A_{ij}(f) S_j(f,t) + E_i(f,t) \qquad i = 1, \ldots, M \tag{4}$$

However the fact that the Discrete Fourier Transform (DFT) domain performs cicular convolution can introduce erros. Furthermore, while it is well known that linear and cicular convolution become equivalent if the DFT length, $T$, is at least twice the length of the room impulse response function, $T \geq 2P$, this model is not applicable when the filter weights are being adapted in the DFT domain.

To understand the nature of this error in this case it is instructive to consider the DFT as a bank of critically sampled narrow-band filters. The approximation error then manifests itself in the form of aliasing between neighbouring frequency bins [27, 11]. In Figure 3, one can see the frequency response of a 16-point DFT filterbank. We can see that aliasing between neighbouring bins starts at relatively high signal levels (-4dB), which can cause distortion in the analysis and reconstruction part of the BSS algorithm.

If the length of the room transfer functions $P$ is substantially shorter than the DFT length $P \ll T$ then the filter will not change much between frequency bins and the aliasing effect will be supressed. However, room acoustics tend to have long transfer functions such that we might expect $P > T$. Thus this argument does not apply.

Instead, aliasing can be reduced by simply oversampling the filterbank [27, 11] (Lambert [16] also showed this in terms of the filter's Laurent series expansion ). Further improvements

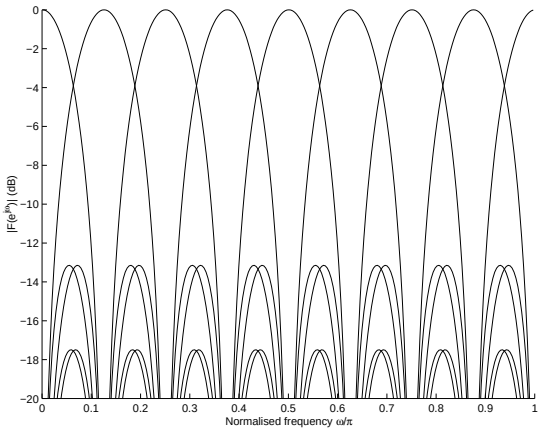


Figure 3. Filter bank characteristic of a 16-point DFT.

can also be obtained through the careful selection of subband filter parameters.

Another possible approach to aliasing reduction is to include adaptive cross terms between neighbouring frequency bins [11], however this introduces additional complexity and is not a favoured technique.

To see the effect of aliasing we applied a standard convolutive BSS algorithm [18] using a 4096 point windowed DFT filterbank with differing amounts of oversampling to a mixture of speech signals recorded (separately) in a real room environment. The average $P$ for the four room transfer functions was $\sim 450ms$. Performance is measured in terms of *Distortion* as introduced by Schobben et al [24].

$$D_{i,j}(f) = 10 \log \frac{\mathcal{E}\{\mathrm{STFT}\{(s_{i,x_j}(t) - \lambda_{ij}\tilde{s}_{i,x_j}(t))^2\}\}}{\mathcal{E}\{\mathrm{STFT}\{s_{i,x_j}(t)^2\}\}} \tag{5}$$

where $\lambda_{ij} = \mathcal{E}\{s_{i,x_j}(t)^2\}/\mathcal{E}\{\tilde{s}_{i,x_j}(t)^2\}$.

Our observations are broadly similar to those in [19]. The distortion, was significantly improved with oversampling. Figure 4, shows the *increase* in distortion introduced with 50% frame overlap in comparison to that with 90% overlap as a function of frequency. It is clear that oversampling predominantly benefits the high frequency distortion. The overall distortion values are summarised in Table I.

## 3. Source modeling domain choice

In order to estimate the unmixing filters using an ICA framework, we have to exploit the observed signal's statistics. One can perform ICA for instantaneous mixtures, by either minimising *Kullback-Leibler* (KL) divergence [1], *Infomax* [4] or *Maximum Likelihood* estimation [5] or even look for directions of the most nonGaussian components, using *kurtosis* or *negentropy* as nonGaussianity measures [12]. All these approaches end up (either implicitly or explicitly) imposing source models on the sample statistics. For example a common approach

Table I. Average along frequency Distortion ($dB$) performance for differing amounts of oversampling.

| | $\mathtt{D}_{1,1}$ | $\mathtt{D}_{2,1}$ | $\mathtt{D}_{1,2}$ | $\mathtt{D}_{2,2}$ |
|---|---|---|---|---|
| Mixing 50% overlap | -3.78 | -4.45 | -6.59 | -2.69 |
| Mixing 75% overlap | -4.56 | -4.90 | -7.21 | -3.43 |
| Mixing 90% overlap | -5.86 | -6.26 | -8.80 | -4.99 |

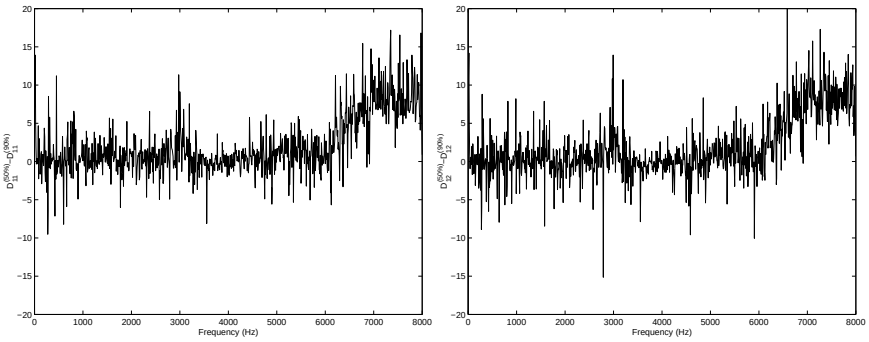


Figure 4. Difference in distortion between the case of 50% overlap and 90% overlap for source 1 at microphone 1 (left plot), and microphone 2 (right plot).

to blind speech separation is to adopt superGaussian models for the time domain sample statistics.

It has also been shown [6] that the *Cramer-Rao bound* for ICA is related to how close the source distributions are to the Gaussian. Therefore to maximise the statistical performance from a finite amount of data we should in principle choose a modelling domain that maximises non-Gaussianity. The question is: in which domain should we model our sources?

### *3.1. Time-domain source modeling*

Source modeling in the time-domain is the obvious choice for an audio source separation algorithm. Speech signals are superGaussian in the time-domain, though this is predominantly due to the 'bursty' nature of the vocalisation in speech. Indeed short term instances of speech (over approximately 20msec.) can appear as either superGaussian or subGaussian depending on the speech segment [10]. This is illustrated in Figure 5.

Musical signals on the other hand are less easily categorized in the time domain (single instrument pieces tend to be superGaussian while polyphonic music is often claimed to be subGaussian, however this really depends on the instruments and the genre of the piece). As such, separating out musical signals using time domain modeling is more difficult.

While modelling the sources in the time domain it is still possible to use subband methods

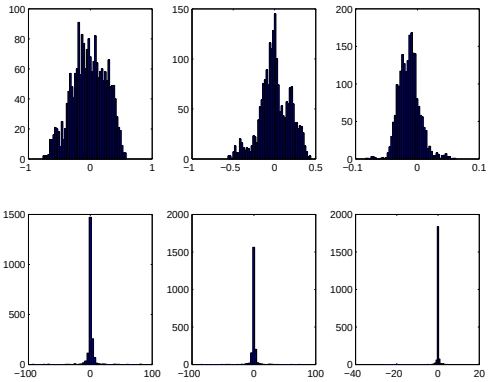

Figure 5. Statistical properties of short audio segments. Histograms in the time domain (first row) and in the frequency domain (second row).

for implementing the unmixing system (e.g. see [17, 19]). However as shown in Figure 1 this requires iteratively mapping between time and frequency domains. This added complexity, along with the fact that in the time domain the degree of nonGaussianity is quite weak has led a number of researchers to look at performing the statistical analysis in the frequency domain.

In contrast to this, one major advantage of working in the time domain is that, at least theoretically, the *permutation problem* does not exist. In the ICA framework, we have to assume that the estimated sources are statistically independent, i.e. $p(\underline{\tilde{s}}) = p(\tilde{s}_1)p(\tilde{s}_2)\dots p(\tilde{s}_N)$. This will cause a single ordering ambiguity in the time domain that is not problematic. We will see that a much more severe ordering ambiguity occurs when modelling in the frequency domain.

### *3.2. Frequency-domain source modeling*

The idea is to use the sample statistics in the frequency domain, using an STFT or other subband filterbank. This allows us to treat each subband as an independent BSS problem with instantaneous (complex) mixing giving the following advantages. First of all, for pretty much all audio signals, the statistics in the frequency domain are far sparser, even in short frames, than their time domain counterparts as can be seen in Figure 5 and they are always superGaussian. This simplifies the separation algorithm since we only need a single nonlinearity to model such statistics. Care has to be taken, however, in defining this nonlinearity since we need an algorithm that can cope with complex signals and mixing [25, 10]. A family of such models has been proposed in [10, 18] to overcome this problem. As mentioned above, the high degree of nonGaussianity should also provide superior statistical performance (however see Section 3.3 below). Finally, since all the computation is performed independently within each subband we have reduced the computational cost (compare Figures 1 and 2).

The drawback of working independently in each subband is the *permutation problem*. While standard ICA has an inherent ambiguity in the ordering of the sources we are now faced with this problem in each subband. This means that we need a way to correctly align the components from each subband. To tackle this problem, we need to devise a coupling mechanism that can sort the permutation of the sources to form a consistent spectrum along the frequency axis.

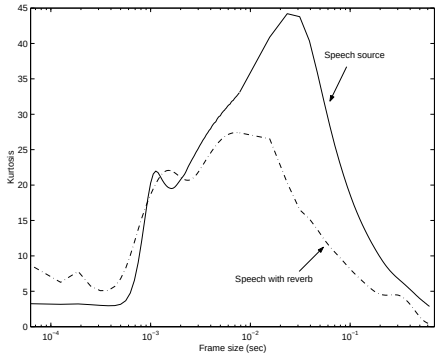


Figure 6. Frame size effect on signal's statistical properties (i.e. estimated kurtosis/sample).

Due to the popularity of the frequency domain approach a number of methods have been proposed to solve the permutation problem. We will discuss these in detail in Section 4.

### *3.3. Effect of frame size*

In audio source separation of signals recorded in a real room, the room transfer functions are usually quite long (i.e. $>$ 100ms). Therefore, to adequately model these, we need to make the frame size large ( $>$ 2048 at 16kHz sampling). However, as the frame size increases, the signal captured by the frame tends to be less stationary and we end up averaging over different quasi-stationary segments of audio. The result is the signal tends to be more Gaussian (roughly via the *central limit theorem*). Even without large frame sizes the presence of the reverberation itself will tend to make the signal more Gaussian for the same reasons. As one of the arguments for working in the frequency domain was to increase the nonGaussianity of the signals we see that there will be a trade-off between large frame sizes that can fully describe the room acoustics and small frame sizes where nonGaussianity is greatest.

To explore this trade-off we examined the statistics of a single frequency bin of a windowed DFT as a function of frame size. Specifically we filtered a speech signal with the following filter:

$$h(n) = w(n)e^{-j\omega_0 n} \qquad n = 1, \ldots, T \tag{6}$$

where $w(n)$ represents the window, $\omega_0 \in [-\pi, \pi]$ represents the frequency at which we want to study our signal's statistics and finally $T$ is the analysis frame length. We used a Hamming window and observed the signal at 1kHz. We measured the signal's nonGaussianity at frame lengths varying from $6ms$ to $625ms$. *Kurtosis* is used as a significant measure of nonGaussianity. As our data are complex, kurtosis is given by the following expression:

$$kurt(x) = \frac{\mathcal{E}\{\mid x \mid^4\}}{\mathcal{E}\{\mid x \mid^2\}^2} - 2 \tag{7}$$

We then repeated the measurement for a reverberant version of the same speech signal (using an estimated room transfer function [28]). Figure 6 shows the level of estimated kurtosis of the signals as a function of frame size.

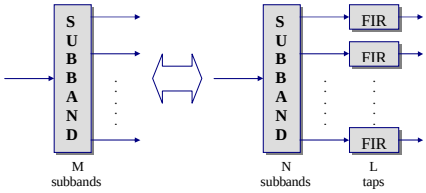


Figure 7. A possible framework to solve the frame size problem

The results follow our intuition. For very small frame sizes the estimated kurtosis tends to that for the time domain source model. Although the signals still have positive kurtosis they are only weakly nonGaussian. As the frame size increases we are able to exploit the sparsity of the sources in the STFT domain.

For the speech signal with no reverberation the estimated kurtosis is maximum for a frame size of about 20-30ms (this is the frame size that is commonly used for efficient speech coding). Note at this value the estimated kurtosis is 10 times that for the time domain model. Finally, as the frame size grows very large we begin to average over a sufficient number of different stationary segments and the estimated signal kurtosis tends towards zero. The effect of reverberation (dashed line) is to reduce the peak value of the estimated kurtosis. However the other general trends persist.

One conclusion is that when source modelling in the frequency domain we cannot afford to choose long frame sizes with $T \gg P$ since this will merely render our signal statistics Gaussian. A similar conlcusion was drawn by Araki et al [3].

Instead, we can choose $P \approx T$ and use oversampling as explained in Section 2. However, when in a highly reverberant environment even this condition may lead to poor performance. In this situation a possible solution, often adopted in subband adaptive filtering algorithms is to use a mixture of subband and convolutional filtering, as depicted in Figure 7, where an $M$ subband filterbank is replaced by a smaller $N$ subband structure and a *short* unmixing filter of size $L$ (such that $M = NL$). This would enable us to work with highly sparse signals while also reducing to some extent the permutation problem (see below).

## 4. Permutation problem

As explained in Section 3, the *permutation problem* arises from assuming statistical independence along the frequency axis when modelling the sources in the frequency-domain. Therefore the ordering of estimated sources along frequency is uncertain. To solve this problem a mechanism must be introduced to couple the signals along frequency. A number of solutions for the permutation problem have been proposed. They can be categorized into two basic groups:

### *4.1. Source modeling Solutions*

In *source modeling* solutions, the aim is to exploit the coherence and the information between frequency bands in order to identify the correct alignment between the subbands. This can be done by imposing time-frequency source models, as proposed by Ikeda [13], Mitianoudis and Davies [18], exploiting the fact that audio signals are rarely independent between frequency bands (due to harmonics and transients). Ikeda used signal envelopes in the time-frequency representation to impose coupling, whereas Mitianoudis and Davies proposed a generative time-frequency model along with a likelihood ratio jump solution in order to force subband coupling and align the sources after unmixing. Both these approaches can be categorised as *flipping solutions*. This takes account of the fact that the permutation problem is a discrete problem embedded within a continuous one. Davies [10] showed that attempting to incorporate these solutions solely within the gradient update tends to cause the solution to get trapped in local minima, resulting in poor separation.

### *4.2. Channel modeling Solutions*

In channel modeling solutions, the aim is to exploit additional information about the room transfer functions in order to select the correct permutations.

An early approach was to assume *smooth* filters, as a constraint to the unmixing algorithm. Smaragdis [25] used a heuristic approach to achieve that. Parra and Spence [21] aligned permutations using a length-constrained FIR filter model, which has been reported to get trapped in local minima [14]. Both approaches can be characterized as *gradient solutions*, and problems similar to those noted in [10] tend to occur.

A more successful approach is to consider the BSS setup as a $N$-sensor beamformer and employ its directivity pattern to resolve the permutations, as investigated by Saruwatari et al [23], Ikram and Morgan [15], Parra and Alvino [20]. We will analyse the application of beamforming in the BSS concept in detail next.

*4.2.1. Beamforming solution* Recently the relationship between convolutive blind source separation and beamforming has been highlighted. In a given frequency bin the unmixing matrix can be interpreted as a null-steering beamformer that uses a blind algorithm (ICA) to place nulls on the interfering sources. However, the source separation framework, as described so far, does not utilize any information concerning the geometry of the auditory scene (e.g. *Directions Of Arrival* (DOA) of source signals, microphone array configuration). Inclusion of this additional information can help align the permutations using the sources' estimated DOA to align the permutations along the frequency axis. In a similar manner to the flipping solutions described above, the permutations of the unmixing matrices are flipped so that the *directivity pattern* of each beamformer is approximately aligned.

Assume that $W_{ik}(f)$ is the unmixing filter coefficient between the $k^{th}$ sensor and the $i^{th}$ source at frequency $f$. The directivity pattern of the $i^{th}$ source at frequency bin $f$ can be defined as follows:

$$F_i(f,\theta) = \sum_{k=1}^{N} W_{ik}^{phase}(f) e^{j2\pi f d_k \sin\theta_i / c} \tag{8}$$

where $W_{ik}^{phase}(f) = W_{ik}(f)/|W_{ik}(f)|$ and accounts for sensor characteristics, $d_k$ is the distance of the $k^{th}$ sensor from the origin, $\theta_i$ is the DOA of the $i^{th}$ source and $c$ is the velocity of sound.

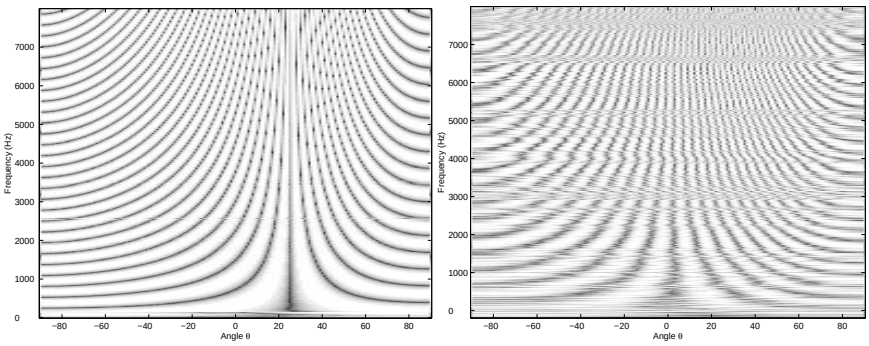


Figure 8. Frequency dependent directivity pattern estimated for (a) a single-delay transfer function and (b) a real reverberant room.

The ideal directivity pattern along frequency for a single delay system is depicted in Figure 8(a), where we can clearly see a null at around $25^o$, which is the actual DOA of the source. Aligning the nulls across frequency is therefore equivalent to assuming that the direct acoustic path dominates the room transfer function (i.e. is approximately anechoic). While the directivity pattern estimated for a real room transfer function (Figure 8(b)) broadly follows this trend it is not so smooth.

From Figures 8(a) and (b), we can also see that we get multiple nulls after a certain frequency ($f_0 \sim 300Hz$) in our beamforming pattern. This frequency is a function of the spacing between the sensors, i.e. $f_0 \sim c/2d_k$. Below this frequency, the nulls appear to have almost a consistent DOA, occasionally slightly shifted. Above this frequency, we start get multiple nulls, however the "main" null is still placed around the DOA.

Saruwatari et al [23] estimated the DOA by taking the statistics with respect to the direction of the nulls in all frequency bins and then tried to align the permutations by grouping the nulls that exist in the same DOA neighbourhood. On the other hand, Ikram and Morgan [15] proposed to estimate the sources' DOA in the lower frequencies, as it is less noisy than in higher frequencies. The next step is to align permutations, by looking for nulls in the neighbourhood of the estimated DOA. Parra and Alvino [20] used more sensors than sources along with *known* source locations and added this information as a geometric constraint to their unmixing algorithm. In other words, the geometric constraint is equivalent to enforcing the ICA algorithm to look for sources on specific DOAs. This *semi-blind* technique can help the ICA algorithm converge more quickly as it constrains the algorithm close to the correct solution. It also helps to reduce permutation errors.

Equally important in beamforming and source separation is the *choice of sensor spacing* $d_k$ in eq 8. Choosing smaller spacing will reduce the multiple nulls at mid-high frequencies. We know that the frequency, where multiple nulls start to appear, is $f_0 \sim c/2d_k$. Therefore, if we want to increase $f_0$, we have to keep $d_k$ small. However, the *Signal-to-Noise Ratio* will decrease, as microphone noise will tend to become more correlated and the problem reduces to the less sensors than sources case. Therefore, the choice of sensor spacing is a trade-off between *separation quality* and *beamforming pattern clarity.*

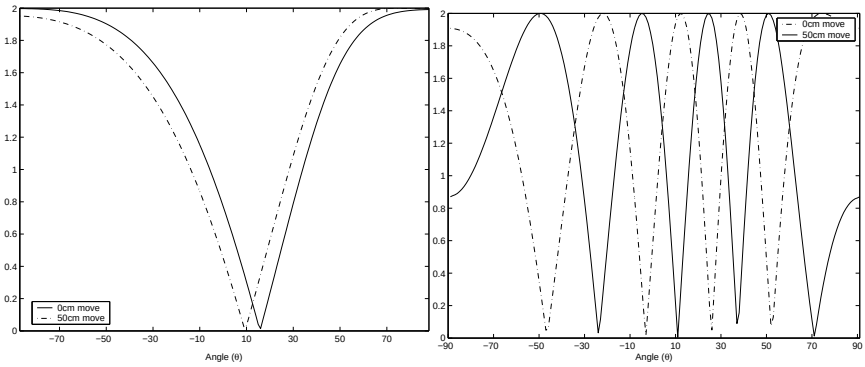


Figure 9. Comparing beamforming patterns at (a) 160Hz and (b) 750Hz.

## 5. Effect of speaker movement on source separation

In this section we consider the effect on the source separation solution of a misalignment of one of the source locations. This will have an impact on any adaptive algorithm that is working in a nonstationary environment. Given that the BSS solutions are statistical in nature there will always be a finite amount of training data necessary to obtain a good separation performance. Therefore in any nonstationary environment we can anticipate that the adaptive filter estimates will lag behind the ideal unmixing filters. We will now investigate the impact of such a misalignment on the source separation algorithm.

We initially performed the following experiment. Using a reasonably reverberant university lecture room ($\sim 7.5m \times 6m$) we performed a simple 2 sources - 2 sensors experiment. We placed two speakers (source 1 and source 2, 1m apart) and two cardiod microphones (mic 1 and mic 2, also separated by $1m$) in the centre of the room. The distance between the speakers and the sensors was $2m$. As before source separation was performed using the approach presented in [18].

Following the initial experiment, we made a new recording with exactly the same setup except that source 2 was moved $50cm$ to the left. We unmixed these sources using first the unmixing filters estimated from the previous experiment and then subsequently estimating the unmixing filters from the new setup. Below we analyse the performance degradation due to the source misalignment and explain our findings in terms of beamforming patterns.

### *5.1. Beamformer's sensitivity to movement*

Using the estimated unmixing filters we can produce their beamforming patterns along frequency (see equation 8) for the original and the displaced source case. Comparing beamforming patterns along frequency, we see that the beamformer's sensitivity to movement is a function of frequency. At low frequencies the beamformer's null has been slightly shifted, due to movement. Figure 9(a) shows the directivity patterns at 160Hz for both the original and displaced experiments. Whilst there will be some degradation due to the misalignment the original beamformer can still suppress the source at this frequency quite effectively.

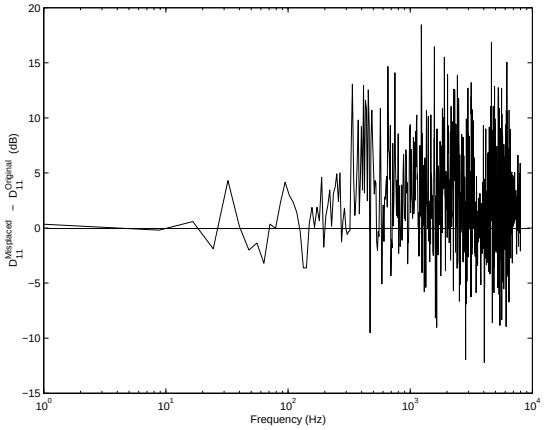


Figure 10. Distortion increases as a function of frequency in the case of a misaligned beamformer.

In contrast to this, even at moderate frequencies the directivity pattern becomes more oscillatory: due to the shorter wavelength. Thus as the frequency increases the source separation algorithm is unable to suppress the interfering source. Figure 9(b) shows that, even at 750Hz, we have a situation where the null is almost replaced by a peak in the misaligned patterns. In this situation, our beamformer is rendered useless. Hence, in mid-higher frequencies source displacement can degrade the performance dramatically.

To quantify this we evaluated the change in distortion (equation 5) due to the movement as a function of frequency. This is shown in Figure 10 (we have used a log frequency scale to highlight the behaviour at low frequancies). As predicted, distortion is not significantly affected at low frequencies. However above about 200Hz the distortion increases by more than 5dB. The result is that all practical source separation above about 300Hz has been lost.

### *5.2. Distortion introduced due to movement*

We conclude this section by examining how the distortion introduced by the displacement of one source is manifested in the separated signals. From listening to the separated sources it was clear that source 1 contained a considerable amount of crosstalk. Source 2 however contained no crosstalk but sounded substantially more "echoic".

These observations can again be explained by considering the directivity patterns associated with the unmixing filters. From our above arguments we see that displacing source 2 will introduce the observed crosstalk. However, since source 1 was not displaced the unmixing filters adapted to *cancel* this source will still place a correct null in the direction of source 1. As a consequence despite the fact that source 2 was moved it is still correctly separated.

The added reverberation in source 2 comes about due to mapping the signals back to the microphone domain (this is an effective way to avoid signal whitening: see [18] for details).

Specifically, if we assume that at a given frequency bin

$$\underline{X} = \begin{bmatrix} A_{11}A_{12} \\ A_{21}A_{22} \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} \tag{9}$$

represents the mixed signals. After separation, we map back to the microphone space, so we are trying to estimate the individual source signals observed at the microphones $\underline{X}_{s1}$, $\underline{X}_{s2}$:

$$\underline{X}_{s1} = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} S_1, \underline{X}_{s2} = \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} S_2 \tag{10}$$

this introduces the following constraint into our source estimates:

$$\underline{X} = \underline{\tilde{X}}_{s1} + \underline{\tilde{X}}_{s2} \tag{11}$$

However, due to misaligned beamforming, one source will get contamination from the other. That is:

$$\underline{\tilde{X}}_{s1} = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} S_1 + \delta \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} S_2 \tag{12}$$

where $\delta$ and $G_1, G_2$ model the error due to misaligned beamforming.

Due to the constraint imposed by equation 11, while the second source will recieve no contamination from source 1 we will get the following estimate for source 2.

$$\underline{\tilde{X}}_{s2} = \left( \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} - \delta \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \right) S_2 \tag{13}$$

Typically, we can expect the additional filtered component of $S_2$ to add reverberation even though there is no crosstalk from source 1.

## 6. Discussion

In this study we have discussed a number of design decisions that have to be made in the basic structure of any blind audio source separation algorithm. It has been particularly instructive to consider BSS algorithms as a set of broad band beamformers. In particular, it has highlighted a problem of performance degradation when one of the source signals moves. This may well prove to be an intrinsic limitation of all current BBS techniques. The source movement or displacement limits the quality of mid to high frequency separation that we can expect in a real nonstationary environment.

It is interesting to compare this with how humans appear to localize sounds. Auditory psychologists have observed that the human ear tends to localize low frequency sounds ($<$ 2000kHz) primarily through phase difference, while localization of high frequency sounds is performed using amplitude difference. Whether or not a similar strategy could be adopted for blind signal separation remains to be seen.

## REFERENCES

1. S.-I. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, 1996.
2. B. Ans, J. Hérault, and C. Jutten. Adaptive neural architectures: detection of primitives. In *Proc. of COGNITIVA'85*, pages 593–597, Paris, France, 1985.
3. S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari. Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech. In *ICASSP*, pages 2737–2740, 2001.

4. A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.
5. J.-F. Cardoso. Infomax and maximum likelihood for source separation. *IEEE Letters on Signal Processing*, 4:112–114, 1997.
6. J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 9(10):2009–2025, 1998.
7. J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, 1993.
8. K.B. Christensen. The application of digital signal processing to large-scale simulation of room acoustics: Frequency response modelling and optimization software for a multichannel dsp engine. *Audio Engineering Society*, 40:260–276, 1992.
9. P. Comon. Contrasts, independent component analysis, and blind deconvolution. *International Journal of Adaptive Control and Signal Processing, this issue*, 2003.
10. M. Davies. Audio source separation. *Mathematics in Signal Processing*, V, 2000.
11. A. Gilloire and M. Vetterli. Adaptive filtering in subbands with applications to acoustic echo cancellation. *IEEE Trans. Signal Processing*, 40(8):1862–1875, 1992.
12. A. Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.
13. S. Ikeda and N. Murata. A method of ICA in time-frequency domain. In *Proc. Int. Workshop on ICA and Signal Separation*, pages 365–370, Aussois, France, 1999.
14. M.Z. Ikram and D.R. Morgan. Exploring permutation inconsistency in blind separation of signals in a reverberant environment. In *ICASSP*, 2000.
15. M.Z. Ikram and D.R. Morgan. A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation. In *ICASSP*, 2002.
16. R. H. Lambert. *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures*. PhD thesis, Univ. of Southern California, 1996.
17. T.-W. Lee, A. J. Bell, and R. Lambert. Blind separation of delayed and convolved sources. In *Advances in Neural Information Processing Systems*, volume 9, pages 758–764. MIT Press, 1997.
18. N. Mitianoudis and M. Davies. Audio source separation of convolutive mixtures. *IEEE Trans. Audio and Speech Processing*, 11(5):489–497, 2003.
19. S.E. Nordholm N. Grbic, X-J Tao and I. Clesson. Blind signal separation using overcomplete subband representation. *IEEE Trans.on Speech and Audio Processing*, 9(5):524–533, 2001.
20. L. Parra and C. Alvino. Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, 10(6):352–362, 2002.
21. L. Parra and C. Spence. Convolutive blind source separation based on multiple decorrelation. In *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP'97)*, Cambridge, UK, 1998.
22. B. A. Pearlmutter and L. C. Parra. Maximum likelihood blind source separation: A context-sensitive generalization of ICA. In *Advances in Neural Information Processing Systems*, volume 9, pages 613–619, 1997.
23. H. Saruwatari, T. Kawamura, and K. Shikano. Fast-convergence algorithm for ica-based blind source separation using array signal processing. In *Proc. Int. IEEE WASPAA*, pages 91–94, New Paltz, New York, 2001.
24. D. Schobben, K. Torkolla, and P. Smaragdis. Evaluation of blind signal separation methods. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA1999)*, Aussois, France, 1999.
25. P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22:21–34, 1998.
26. K. Torkkola. Blind separation of delayed sources based on information maximization. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'96)*, pages 3509–3512, Atlanta, Georgia, 1996.
27. S. Weiss and I. Proudler. Comparing efficient broadband beamforming architectures and their performance trade-offs. In *DSP conference*, 2002.
28. A. Westner. http://www.media.mit.edu/∼ westner.