ISSUES IN FREQUENCY DOMAIN BLIND SOURCE SEPARATION - A CRITICAL REVISIT

Enrique Robledo-Arnuncio & Biing-Hwang (Fred) Juang

Center for Signal and Image Processing
Georgia Institute of Technology
Atlanga, GA 30332-0250
{era,juang}@ece.gatech.edu

ABSTRACT

One of the most important problems in frequency domain blind source separation (FDBSS) is the inconsistency across frequency in the permutation of the source estimates. According to previous studies, this problem can be reduced significantly by constraining the length of the unmixing filters. This improvement has been attributed to the smoothening of the unmixing frequency response. In this paper we study the effect of modifying these length constraints taking into account the circularity of the IDFT, and we show that the smoothening of the unmixing frequency response alone can not account for the improvements in performance.

1. INTRODUCTION

The most studied modality of the Blind Source Separation (BSS) problem is the case of instantaneous linear mixtures:

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) \tag{1}$$

where $\mathbf{x}(n)$ is a random vector of measurements at time n, $\mathbf{s}(n)$ is a random vector of unknown *independent* source signals to be recovered, and the unknown matrix $\mathbf{A}$ defines the mixture.

The solution to this problem involves finding a linear system $\mathbf{W}$ which transforms the measurement vectors into vectors with statistically independent components. Such a system can be found, under certain conditions on the source statistics and on the mixture matrix, up to an arbitrary permutation and scaling of its rows.

For stationary non-gaussian sources the unmixing system can be found using higher order statistics (HOS)[1]. For non-stationary sources, it may be possible to formulate the problem using the second order statistics (SOS) of the measurement signals at different times. In both situations the required statistics are usually estimated from available instances of the measurement vectors.

A more challenging problem is the separation of convolutive mixtures. A possible approach is to move it to frequency domain:

$$\mathbf{x}(n) = \mathbf{A}(n) * \mathbf{s}(n) \implies \mathbf{x}(\omega) = \mathbf{A}(\omega)\mathbf{s}(\omega) \tag{2}$$

where $\mathbf{x}(\omega)$ and $\mathbf{s}(\omega)$ are two random vectors obtained as linear (Fourier) transformations of the original signals. It is now possible to apply either the HOS approach [2] or the SOS approach [3] to obtain the unmixing matrix at each frequency. Unfortunately the arbitrary row permutation and scaling ambiguities become now two major problems. The latter makes the resulting separated signals relate to the sources through an arbitrary linear filter, while the former produces outputs that are still mixtures of the sources in time domain, due to *permutation inconsistencies* of the outputs in frequency domain. In other words, the original additive mixing problem becomes swapping in individual sinusoids, which is further compounded by the scaling problem. It is thus a critical problem of the frequency domain approach.

The root of these inconsistencies is the fact that the separation at each frequency bin is being treated as an independent problem. One way to address this is to modify the separation algorithm to enforce some constraint that relates the solutions across frequency [3]. A different approach is to solve the independent problems and apply some post-processing at the outputs to align the solutions, using knowledge about the source signal or the mixing system.

While these problems and potential solutions have been discussed in the literature, the reports were mostly on the use of Signal-to-interference ratio (SIR) as the figure of merit and indicator of the improvement realized by the proposed solutions. Rarely discussed were the actual quality of the recovered source signals and other issues possibly rising from implementations that may not be immediately obvious in the theoretical and algorithmic development. One such example is the circularity problem that was only reported in 2003 [4], after many years of BSS research.

In this paper we revisit the problem of BSS and analyze proposed solutions, using the approach of Parra [3] as the platform, in order to gain a deeper understanding of the related issues. For example, it has been shown [5] that Parra's approach failed to remove all the permutation inconsistencies in some experiments. We explore again this problem and offer a more careful analysis of the effect of the constraint, particularly from the causality point of view that relates the solutions across frequency. Our experimental results confirm that these issues have substantial impact on the performance of a BSS algorithm that was not addressed before.

2. THE SOURCE SEPARATION ALGORITHM

As in [5], we will restrict our discussion to the case of FIR mixing filters of order P, square mixing system of size $N \times N$ and noiseless measurements. From (2) we can compute:

$$\mathbf{R}_x(\omega) = E\{\mathbf{x}(\omega)\mathbf{x}(\omega)^*\} = \mathbf{A}(\omega)\mathbf{R}_s(\omega)\mathbf{A}^*(\omega) \tag{3}$$

where $\mathbf{R}_x(\omega)$ is a complex $N \times N$ Hermitian matrix. We want to find a $N \times N$ unmixing matrix $\mathbf{W}(\omega)$ that diagonalizes

$$\mathbf{R}_y(\omega) = E\{\mathbf{y}(\omega)\mathbf{y}(\omega)^*\} = \mathbf{W}(\omega)\mathbf{R}_x(\omega)\mathbf{W}^*(\omega) \tag{4}$$

where $\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{x}(\omega)$. $\mathbf{W}(\omega)$ has N^2 unknowns, while the diagonal condition only imposes $\frac{N(N+1)}{2}$ constraints, due to the symmetry of $\mathbf{R}_y(\omega)$. It is necessary to take advantage of the non-stationarity of the sources. We do so by analyzing separately K

different time sections of the measured signals. Defining $\mathbf{x}_L(\omega, k)$ as the Fourier transform of the kth segment of length L of the signal, then the matrices $\mathbf{R}_{x_L}(\omega, k) = E\{\mathbf{x}_L(\omega, k)\mathbf{x}_L(\omega, k)^*\}$ can be approximated as:

$$\mathbf{R}_{x_L}(\omega, k) \approx \mathbf{A}(\omega)\mathbf{R}_{s_L}(\omega, k)\mathbf{A}^*(\omega) \tag{5}$$

where $\mathbf{R}_{s_L}(\omega, k)$ is the (diagonal) correlation matrix of the Fourier coefficients for the kth segment of the source signals. The approximation requires $L \gg P$.

The correlation matrices $\mathbf{R}_{x_L}(\omega, k)$ can be estimated from a realization of the measured signal by further dividing each length-L time segment into M sub-frames of length T, computing the Fourier transform of each of these sub-frames, $\mathbf{x}_T(\omega, k, m)$, and doing time averages of the cross product of the resulting vectors:

$$\hat{\mathbf{R}}_{x_L}(\omega, k) = \frac{1}{M}\sum_{m=0}^{M-1} \mathbf{x}_T(\omega, k, m)\mathbf{x}_T^*(\omega, k, m) \tag{6}$$

Note that in order for the approximation in (5) to hold, it is now necessary that $T \gg P$. Also, this equation assumes some kind of ergodicity in the random vectors across successive length T sub-frames. This can not be a strict ergodicity, since we are assuming non-stationarity of the signal across different length L segments.

For separation of signals such as speech, this notion of ergodicity is related to the traditional one used to estimate the auto-correlation function and the power spectral density, which is also normally addressed in the context of short time spectral analysis (STSA). This STSA framework will also allow us to examine the "signal tracking" behavior of the separation process, as in many adaptive impulse response estimation and inverse problems.

The proposed solution to the source separation problem is a system of unmixing filters that diagonalize

$$\hat{\mathbf{R}}_{y_L}(\omega, k) = \mathbf{W}(\omega)\hat{\mathbf{R}}_{x_L}(\omega, k)\mathbf{W}(\omega)^* \tag{7}$$

simultaneously for several values of k so as to satisfy the assumed independence requirement. The diagonalization error matrix for one of these correlation matrices is:

$$E(\omega, k) = \text{offdiag}\left(\mathbf{W}(\omega)\hat{\mathbf{R}}_{x_L}(\omega, k)\mathbf{W}^*(\omega)\right) \tag{8}$$

A possible way to aggregate these error matrices is adding their Frobenius norms. The solution would be the filter that minimizes

$$\min_{\mathbf{W}(\omega)} \sum_{k=1}^{K} \|E(\omega, k)\|_F^2 \tag{9}$$

A simple approach to attempt this minimization is the gradient descent. The update equation for the unmixing filters is

$$\mathbf{W}^{(l+1)}(\omega) = \mathbf{W}^{(l)}(\omega) - \mu(\omega)\sum_{k=1}^{K} E(\omega, k)\mathbf{W}^{(l)}(\omega)\hat{\mathbf{R}}_{x_L}(\omega, k)$$

Note that this approach may converge to a local minimum [5].

3. OPTIMIZATION CONSTRAINTS

A possible way to deal with the row scale ambiguity in the unmixing matrices is to fix the value of one element in each row, for example [3, 5] setting the diagonal elements in the unmixing matrix to one:

$$\{\mathbf{W}(\omega)\}_{ii} = 1 \qquad \forall i = 1, \ldots, N \tag{10}$$

In time domain this makes the filters in the diagonal of the unmixing system equal to an impulse at time zero. The experiments show that with this constraint the resulting waveforms for the off-diagonal unmixing filters are concentrated around time zero too.

Regarding the permutation inconsistency, Parra [3] proposed constraining the waveform of the unmixing filters to a length Q, with $Q < T$, at each step of the algorithm. This is equivalent to performing a convolution with a linear-phase circular sinc function in frequency domain, and it thus enforces a smoothing of the frequency response of the unmixing filters, which is hopefully not compatible with permutation inconsistencies. But Ikram [5] observed that some degree of permutation inconsistency remains. Also, since the smoothing constraint occurs as part of the gradient search, it perturbs the intermediate solutions and it will spread search errors uncontrollably and non-uniformly across frequency.

A related problem of the truncation approach then is that it is difficult to define a consistent stopping condition which achieves a reasonable result for all frequencies.

In [3] and [5] the filter length constraint was implemented by keeping only the first Q samples of the time domain frames:

$$\{\mathbf{W}\}_{ij} \leftarrow FZF^{-1}\{\mathbf{W}\}_{ij} \tag{11}$$

where $\{\mathbf{W}\}_{ij}$ is a frequency frame, F is the DFT matrix of size T, and Z is the truncation matrix:

$$Z_{ij} = \begin{cases} 1 & \text{if } i = j \text{ and } i \leq Q \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

It is interesting to note the relation between the constraints (10) and (12). The former concentrates the waveform of the filters around the time origin, as will be shown later, while the latter limits their waveforms to a segment starting at the origin. The result is a causal waveform with a decaying tail, as shown in figure 1(a). There is no obvious need to enforce this shape in the waveform.

4. EFFECTS OF THE CIRCULARITY OF THE DFT

By *circularity* of the DFT/IDFT we refer to the implicit periodicity assumed in the time domain signals subject to these transformations. It can be seen as an artifact due to sampling the continuous frequency representation of a finite length signal. It is a basic theoretical fact, and its consequences are pervasive in frequency domain algorithms.

One of the consequences, always explicitly taken into account in previous FDBSS work, is the circular convolution property. This is what makes equation (5) an approximation.

There are two other important consequences found when applying the IDFT. The first one is *time aliasing*, which happens when the frequency domain representation corresponds to a time domain signal longer than the DFT frame, i.e, when the frequency representation is under-sampled. This problem is well known in the context of frequency domain block-convolution. It has also been treated in the context of frequency domain adaptive filters, where it becomes more difficult to avoid. The adaptive algorithms often work independently at each frequency bin, converging to samples of the frequency representation of an optimum solution which can be long in time.

We have only found very recent work [4] dealing with time aliasing in the context of FDBSS. It is a problem hard to analyze or predict: the unmixing filter is being computed directly in frequency domain, leading to the same issues found in frequency domain adaptive filters. And this can be complicated by the linear

filtering ambiguity in the solutions, which can lead to filter waveforms of different lengths with the same separation performance.

The second and more obvious consequence of the circularity is the possibility of obtaining a waveform that wraps around the edges of the time frame. This happens when the frequency domain representation has a significant linear phase component, or when it corresponds to a non-causal signal, as we will see below. These conditions can lead to incorrect waveforms if care is not taken to avoid them in the frequency domain or to undo the circular wrapping in the result. This may not be trivial if the frequency domain representation is obtained through some optimization algorithm and no previous knowledge is available about the characteristics of the solution.

The constraint (11) used to smooth the frequency responses has the advantage of avoiding these circularity issues by defining the portion of the frame where the waveform is to be placed. But as we will see, they arise when the constraint is removed or modified.

5. EXPERIMENTAL SETUP

We constructed 72 different pairs of speech signals. 24 of them female/female, 24 male/male and 24 female/male. Each speech signal consisted of 57.6 seconds of concatenated recordings from a specific speaker taken from the clean test section of the TIDIGIT database. The signals were down-sampled to 8kHz.

For the mixing system we used room response recordings from the Microphone Array section of the *RWCP Sound Scene Database in Real Acoustic Environment*. The selected recordings correspond to the Echo room A (E2A), with 30 ms of reverberation time. The two microphones are 30 cm apart, and the two sources are located at 50 and 110 degrees with respect to the reference plane.

We used frame sizes (T) of 2400 and 3600 samples with 50% overlap. The signals were divided in $K = 4$ non-overlapping segments. The unmixing filter lengths were $Q = T$ (no constraint), $Q = \frac{T}{2}$ and $Q = \frac{T}{4}$. The two latter lengths were enforced as both causal (12) and modified constraints, described below (13).

SIR values were computed by doing time domain convolutions with the resulting filters, both before and after applying a post-alignment of the permutations. For the alignment we followed the procedure described in [5]. These SIR values were averaged across the different speaker pairs and frame sizes.

6. RESULTS WITH NO LENGTH CONSTRAINT

The experiments with no time domain truncation (11) in the gradient loop (i.e, with $T = Q$) were aimed at providing a baseline that would clarify the effect of the length constraint. As we expected, the results present a high degree of permutation inconsistency.

Figure 1(b) shows a typical resulting waveform for one of the filters outside of the diagonal of the unmixing system. This shape suggests that a negative-time tail is being wrapped around the beginning of the frame. Under such hypothesis, Figure 1(c) shows the result of unwrapping the waveform. We computed the separation performance with both sets of unmixing filters 1(b) and 1(c) trying to decide which of them is the correct one.

The separation performance results are shown in Table 1. The table shows that taking the wrapped filters in Figure 1(b) with aligned permutations leads to worse SIR results (3.1 dB) than when using the unwrapped waveform in Figure 1(c) (7.9 dB). This confirms that the algorithm without the filter length constraint is converging to non-causal filters in its off-diagonal components, with

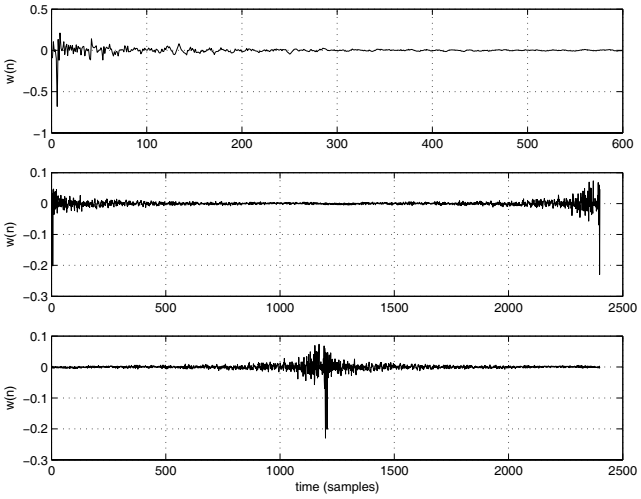

Fig. 1. Waveforms for an off-diagonal component of the unmixing system, with (a) window (12), (b) $T = Q$, (c) $T = Q$ unwrapped

an approximately symmetric envelope centered around the time origin, and suffering permutation inconsistencies. We have observed this behavior in all our experiments.

	Orig. frames	unwrapped frames
original permut.	1.8 (0.3)	0.4 (0.5)
aligned permut.	3.1 (0.9)	7.9 (1.4)

Table 1. Mean and standard deviation SIR gain in dB, after optimization with no length constraint.

7. RESULTS WITH MODIFIED CONSTRAINTS

The windowing performed in (12) corresponds to a convolution with a linear phase sinc function in frequency. If a zero phase sinc is used instead, the corresponding window would be:

$$Z_{ij} = \begin{cases} 1 & \text{if } i = j \text{ and } \left(i \leq \frac{Q}{2} \text{ or } i > T - \frac{Q}{2}\right) \\ 0 & \text{otherwise} \end{cases} \qquad (13)$$

In this case the window is centered at the origin, allowing the algorithm to converge to a wrapped non-causal filter, as in the previous section. The smoothing effect should be similar when using this centered window.

The results are very different, though. Table 2 shows the gain in separation performance using each filter. The SIR improvement when using causal window is 7.0dB. With centered window it is close to 0dB, but permutation post-alignment seems to fix the problem. Surprisingly, the smoothing provided by this new window is not helping to fix the permutation inconsistencies.

8. SIR VARIATIONS ACROSS FREQUENCY

It is possible to analyze the separation performance at each frequency bin. The representation of SIR versus frequency allows to visualize the permutation inconsistencies, since they make the sign of the SIR negative while correct permutations make it positive.

	Causal window	0-phase window
original permut.	7.0 (0.8)	0.2 (0.7)
aligned permut.	6.7 (0.8)	7.3 (1.0)

Table 2. Mean and standard deviation SIR gain in dB, using constraint (11) (left column) and constraint (13) (right column).

Figure 2(a) shows the SIR achieved at each frequency bin in a typical experiment when using the causal window (12) as proposed by Parra. The performance is not constant across frequency, possibly due to stochastic errors in the source correlation matrices and to convergence problems. There are some remaining permutation inconsistencies, but some frequency bands present a quite homogeneous SIR with values above 10 dB.

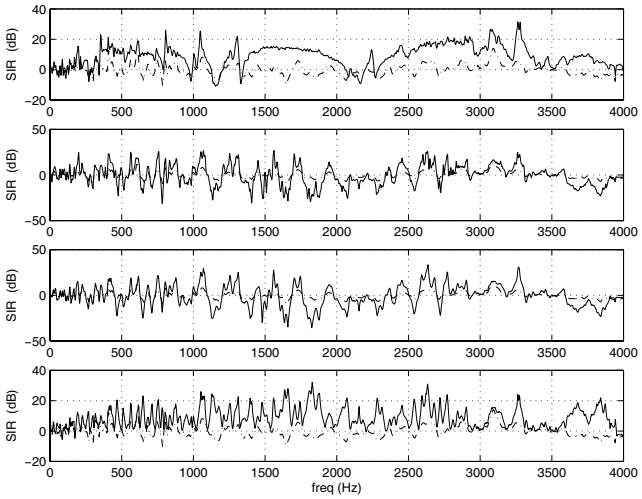

Fig. 2. SIR for each frequency bin at the inputs (broken line) and at the outputs (continuous line) with (a) $Q = \frac{T}{4}$ and causal window. (b) $Q = T$, (c) $Q = \frac{T}{4}$ and centered window, (d) previous one with permutation post-alignment

Figure 2(b) shows typical results with no filter length constraint. SIR oscillates between positive and negative values but, interestingly, these oscillations are not simple changes in sign. We have consistently observed that the magnitude of the achieved SIR presents gradual but significant changes. These might be related to convergence problems: We have observed different convergence rates across frequency, even when using the frequency dependent normalization of the gradient proposed in [3]. This seems to be related to wide variations across frequency in the condition number of the mixing matrices we used. The gradient approach used for the optimization seems to be too sensitive to these variations.

Figure 2(c) shows the results when the time domain window is centered at the origin, keeping the non-causal tail. The evolution of SIR is almost identical to the one with no windowing. Some smoothening of the sharpest variations can be observed, but the broad variations remain. The last plot shows the results of the previous experiment after performing a post-alignment of the permutations. The sign of the SIR is now consistent, but the variations in its magnitude are still present, giving a very different result to the one obtained with the causal window in Figure 2(a).

9. CONCLUSIONS

The results in section 6 reveal that the gradient optimization of eq. (9) converges to non-causal filters with two tails and symmetric envelope when no additional constraints are imposed on the length of the filters. This shows that the filter length constraints used in [3] have an unexpected effect: they enforce a causal waveform on the unmixing filters in addition to the previously reported smoothening of their frequency responses.

This observation raises the question of how much each of these two effects, smoothening and enforcement of causality, is contributing to the improvement in separation performance. The results in section 7 show that in all of our experiment conditions the improvement is mostly due to the enforcement of causality, since an alternative constraint with similar smoothing effect but which preserves the non-causal tail does not improve the results.

The improvement in performance achieved with the original causal constraints has so far [3, 5] been attributed to the alignment of permutation inconsistencies. The reason for this is that a post-alignment of the permutations when no constraints are used provides an equivalent improvement in performance. But the results in section 8 show that the effect of those constraints is more complex than a simple alignment of permutations. Figures 2(a) and (d) show that the causal window used in the first one is actually making the algorithm converge to a different solution. It seems that enforcing causality in the unmixing filters modifies the constrained solution space in a way that benefits both the convergence behavior of the algorithm and the consistency of the permutations.

10. ACKNOWLEDGEMENTS

We thank Prof. Tomoko Matsui for valuable discussion and help setting up the experiments. This work was partially funded by the Fulbright Commission Spain through a grant to one of the authors.

11. REFERENCES

[1] Jean-Francois Cardoso, "Blind signal separation statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[2] Marcel Joho and Philip Schniter, "Frequency domain realization of a multichannel blind deconvolution algorithm based on the natural gradient," in *Proc. 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, Apr. 2003, pp. 543–548.

[3] Lucas C. Parra and Clay Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[4] H. Sawada, R. Mukai, S. de la Kethulle de Ryhove, S. Araki, and S. Makino, "Spectral smoothing for frequency-domain blind source separation," in *Proc. 8th International Workshop on Acoustic Echo and Noise Control (IWAENC 2003)*, Kyoto, Japan, Sept. 2003, pp. 311–314.

[5] Muhammad Z. Ikram and Dennis R. Morgan, "Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Istambul, Turkey, June 2000, vol. 2, pp. 1041–1044.