International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:5, No:8, 2011

A Frequency Grouping Approach for Blind Deconvolution of Fairly Motionless Sources

E. S. Gower, T. Tsalaile, E. Rakgati and M. O. J. Hawksford

***Abstract*—**A frequency grouping approach for multi-channel instantaneous blind source separation (I-BSS) of convolutive mixtures is proposed for a lower net residual inter-symbol interference (ISI) and inter-channel interference (ICI) than the conventional short-time Fourier transform (STFT) approach. Starting in the time domain, STFTs are taken with overlapping windows to convert the convolutive mixing problem into frequency domain instantaneous mixing. Mixture samples at the same frequency but from different STFT windows are grouped together forming unique frequency groups.

The individual frequency group vectors are input to the I-BSS algorithm of choice, from which the output samples are dispersed back to their respective STFT windows. After applying the inverse STFT, the resulting time domain signals are used to construct the complete source estimates via the weighted overlap-add method (WOLA). The proposed algorithm is tested for source deconvolution given two mixtures, and simulated along with the STFT approach to illustrate its superiority for fairly motionless sources.

***Keywords*—**Blind source separation, short-time Fourier transform, weighted overlap-add method

I. Introduction

A multi-channel BSS algorithm estimates the underlying source signals without any knowledge of the mixing matrix [1-4]. This degeneracy is often circumvented by making some characteristic assumption about the underlying sources, such as mutual independence, or the mixing process. For instantaneous mixing problems, the observed mixtures $\mathbf{x}(n)=[x_0(n), x_1(n),\ldots, x_{L-1}(n)]^T$ are linear mixtures of the scaled Dirac-impulse filtered sources $\mathbf{s}(n)=[s_0(n), s_1(n), \ldots, s_{L-1}(n)]^T$, where T represents vector or matrix transpose. That is

$$\mathbf{x}(n)=\mathbf{H}\cdot\mathbf{s}(n)+\boldsymbol{\varepsilon}(n),$$

where $\mathbf{H}$ is an $L\times L$ mixing matrix and $\boldsymbol{\varepsilon}(n)=[\varepsilon_0(n), \varepsilon_2(n),\ldots,\varepsilon_{L-1}(n)]^T$ is the additive noise vector. In practice, observed signals tend to be convolutive mixtures of the underlying sources as in most audio and video recordings. In this case the model is given by

$$\mathbf{x}(n)=\mathbf{H}(n)*\mathbf{s}(n)+\boldsymbol{\varepsilon}(n), \tag{1}$$

where $*$ is the convolution operator. It follows that I-BSS algorithms like those in [1-4] fail to acceptably identify the underlying sources. Consequently, estimation methods of instantaneous mixing in the frequency domain are used to facilitate the use of conventional I-BSS algorithms, particularly the STFT approach. In [5], it is said that the human voice is stationery for a period shorter than 10ms. Any longer than that the frequency components of the speech change and it is no longer stationery. Thus, given the delays and room reflection are not too long, K-point STFTs can be used to estimate the instantaneous mixing problem as proposed in [5] resulting in

$$\mathbf{x}(\omega,n)=\mathbf{H}(\omega)\mathbf{s}(\omega,n)+\boldsymbol{\varepsilon}(\omega,n), \tag{2}$$

for the spectral range $0\le\omega\le 2\pi(K-1)/K$, where $\mathbf{x}(\omega,n)$ and $\mathbf{s}(\omega,n)$ are the STFTs of the convolutive mixtures and sources respectively. An appropriate I-BSS algorithm is then applied to the spectral frame vector $\mathbf{x}(\omega,n)$ giving

$$\mathbf{y}(\omega,n)=\mathbf{W}(\omega)\mathbf{x}(\omega,n),$$

for the learned de-mixing matrix $\mathbf{W}(\omega)$. The length of the STFT is dependent on the assumed time that the source signals are stationery. Clearly this stationery assumption varies from application to application and is not truly accurate despite acceptable results in certain applications such as the cock-tail party problem.

In this paper, we propose a method for modeling an instantaneous mixing problem in the frequency domain without having to make stringent assumptions about the spectral properties of the underlying sources. However, we circumvent the BSS uncertainty problem by making an assumption on the mixing process that the sources are fairly motionless over certain period of time greater than the STFT window length. The result is more effective source separation as reflected by a measure of the net residual ISI and ICI compared to the conventional STFT approach of (2).

E. S. Gower is with the Department of Electrical Engineering, University of Botswana, Gaborone, email: ephraim.gower@mopipi.ub.bw.

T. Tsalaile is with the Department of Electrical Engineering, University of Botswana, Gaborone, email: tsalaile@mopipi.ub.bw

E. Rakgati is with the Department of Electrical Engineering, University of Botswana, Gaborone, email: rakgatie@mopipi.ub.bw.

M. O. J. Hawksford is with the School of Computer Science and Electronic Engineering, university of Essex, Colchester, United Kingdom, email: mjh@essex.ac.uk.

The rest of the paper is structured as follows: In Section II we introduce an algorithm for forming frequency collections from the STFT spectral frames, for a better approximation of the instantaneous mixing problem. Simulation results comparing this approach to the usual STFT method are presented in Section III. The discussion in Section IV summarizes the advantages and limitations of the proposed frequency grouping based algorithm.

II. The Proposed Approach

Given the mixture signal $x_l(n)$, $0 \le l \le L-1$, a windowed data frame $x_l^{(r)}(n)$, $0 \le r \le R-1$, of length N is extracted using

$$x_l^{(r)}(n) = x_l(n)\, w(n - r\Gamma), \tag{3}$$

for $r\Gamma \le n \le r\Gamma + N - 1$, where Γ is the window hop-size. This results in R vectors of the form

$$\mathbf{x}^{(r)}(n) = [x_0^{(r)}(n), x_1^{(r)}(n), \ldots, x_{L-1}^{(r)}(n)]^T. \tag{4}$$

We then take the discrete Fourier transform (DFT) of the data frames in (3) translated to time zero, that is $\tilde{x}_l^{(r)} = x_l^{(r)}(n + r\Gamma)$, to produce the spectral frames $\tilde{X}_l^{(r)}(\omega)$, for all l. As a result, from the time domain vector $\tilde{\mathbf{x}}^{(r)}(n)$ we have the spectral vector

$$\tilde{\mathbf{X}}^{(r)}(\omega) = [\tilde{X}_0^{(r)}(\omega), \tilde{X}_1^{(r)}(\omega), \ldots, \tilde{X}_{L-1}^{(r)}(\omega)]^T.$$

If using the usual STFT approach, it is at this point that an I-BSS algorithm of choice is applied to each of the R vectors. However, these are spectral frames and for source separation deconvolution filters must be used instead of an instantaneous de-mixing matrix as usually done. Fortunately if the sources have some stationery properties as mentioned earlier, an instantaneous matrix often suffices.If the sources are fairly motionless then the convolution kernels from the sources to the recording devices are fairly time-invariant. With this mixing process constraint, it is not necessary to make any assumptions about the source signals. The time-invariance means that the sources contributing to the frequency domain mixture sample say $\tilde{X}_l^{(r)}(\omega_k)$, for $0 \le k \le K-1$ (i.e. ω_k represents a unique frequency from the spectral range covered by ω), are filtered in almost the same way as for $\tilde{X}_l^{r+\lambda}(\omega_k)$, for $r + \lambda \le R-1$. In other words, we can use samples from the R STFTs of a particular mixture signal $\tilde{x}_l^{(r)}(n)$ but at the same frequency ω_k to infer the filter coefficient $H(\omega_k)$. Let the frequency group $Z_l(\omega_k)$ be a collection of samples from the R STFTs of the mixture signal $x_l(n)$ at the same frequency ω_k. That is

$$Z_l(\omega_k) = [X_l^{(0)}(\omega_k), X_l^{(1)}(\omega_k), \ldots, X_l^{(R-1)}(\omega_k)]. \tag{5}$$

From all the L mixture signals, we have the vectors

$$\mathbf{Z}(\omega_k) = [Z_0(\omega_k), Z_1(\omega_k), \ldots, Z_{L-1}(\omega_k)]^T \tag{6}$$

for $0 \le k \le K-1$. This frequency grouping is illustrated in Fig. 1. After forming the vectors of (6), an I-BSS algorithm of choice is applied to each of them. Since these samples are at one frequency, it is mathematically appropriate to use an instantaneous de-mixing matrix regardless of the nature of the underlying sources. If $\mathbf{W}(\omega_k)$ is the learned de-mixing matrix from the frequency group vector $\mathbf{Z}(\omega_k)$, then the separated source signal frequency group vector at ω_k is given by

$$\mathbf{Q}(\omega_k) = \mathbf{W}(\omega_k)\mathbf{Z}(\omega_k), \quad \text{for} \quad 0 \le k \le K-1. \tag{7}$$

If the I-BSS output to the frequency sample $\tilde{X}_l^{(r)}(\omega_k)$ is

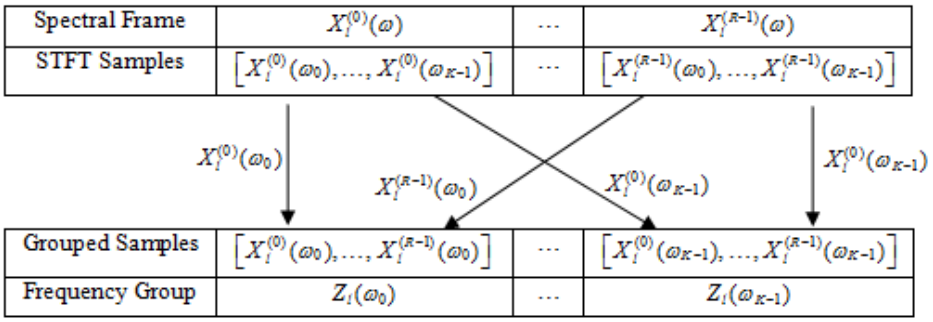

Fig. 1. A frequency group $Z_l(\omega_k)$ is formed by collecting samples at the same frequency from the R spectral frames. Since there are K frequencies from the DFT, there are K frequency groups from R spectral frames.

$\tilde{Y}_l^{(r)}(\omega_k)$, then the output frequency group to $Z_l(\omega_k)$ is

$$Q_l(\omega_k) = [\tilde{Y}_l^{(0)}(\omega_k), \tilde{Y}_l^{(1)}(\omega_k), \ldots, \tilde{Y}_l^{(R-1)}(\omega_k)], \quad (7)$$

in accordance with (5). Hence the output vector is of the form

$$\mathbf{Q}(\omega_k) = [Q_0(\omega_k), Q_1(\omega_k), \ldots, Q_{L-1}(\omega_k)]^T, \quad (8)$$

for $0 \le k \le K-1$. Inherent to most BSS algorithms is the permutation of the outputs. As there are K frequency grouped vectors for I-BSS, it is possible that a particular source is produced at different values of l, $0 \le l \le L-1$. An attempt to construct the sources directly from the vectors of (8) might lead to spectral errors. Typical of most frequency domain BSS algorithms utilizing the STFT, there is a need for a permutation solving algorithm for accurate source estimates construction [5-7]. In [8, 9] the fourth order cross-cumulant (kurtosis) is used as a statistical dependency measure between the different I-BSS spectral outputs. Here we use it as

$$\begin{aligned}\kappa\left[Q_l(\omega_k), Q_m(\omega_f)\right] = {} & \mathrm{E}\left[\left|Q_l(\omega_k)\right|^2 \left|Q_m(\omega_f)\right|^2\right] \\ & - \mathrm{E}\left[\left|Q_l(\omega_k)\right|^2\right]\mathrm{E}\left[\left|Q_m(\omega_f)\right|^2\right] \\ & - \left|\mathrm{E}\left[Q_l(\omega_k) Q_m^H(\omega_f)\right]\right|^2 \\ & - \left|\mathrm{E}\left[Q_m(\omega_f) Q_l^H(\omega_k)\right]\right|^2,\end{aligned} \quad (9)$$

where E[.] is mathematical expectation, H denotes conjugate transpose, $0 \le l, m \le L-1$ and $0 \le f, k \le K-1$. Assuming the original sources are statistically independent and adequate source separation at each one of the frequency grouped vectors, a relatively high value of (9) suggests a high probability that the outputs at frequencies ω_k and ω_f are for the same source signal.

Having solved the permutation problem, we decompose the samples of the source estimates frequency group $Q_l(\omega_k)$, $0 \le l \le L-1$, to their respective STFT windows, yielding

$$\tilde{\mathbf{Y}}^{(r)}(\omega) = [\tilde{Y}_0^{(r)}(\omega), \tilde{Y}_1^{(r)}(\omega), \ldots, \tilde{Y}_{L-1}^{(r)}(\omega)]^T, \quad (10)$$

for $0 \le r \le R-1$. This re-grouping is illustrated in Fig. 2. Applying the inverse STFT to the spectral frames gives the time zeroed outputs

$$\tilde{\mathbf{y}}^{(r)}(n) = \left[\tilde{y}^{(r)}(n), \tilde{y}^{(r)}(n), \ldots, \tilde{y}_{L-1}^{(r)}(n)\right]^T.$$

It is possible that some spectral errors due to blocking effects and the I-BSS signal processing be pronounced in each of the time domain output frames. To minimize these, a synthesis/output window $f(n)$ is applied to the output $\tilde{y}_l^{(r)}(n)$ giving a weighted output $\tilde{y}_l^{(r)}(n) f(n)$. After windowing, the rth output frame is translated back to time $r\Gamma$ yielding $y_l^{(r)}(n) = \tilde{y}_l^{(r)}(n - r\Gamma) f(n - r\Gamma)$. The output vector frame is now

$$\mathbf{y}^{(r)}(n) = \left[y^{(r)}(n), y^{(r)}(n), \ldots, y_{L-1}^{(r)}(n)\right]^T. \quad (11)$$

The use of the input window $w(n)$ and the output window $f(n)$ in an overlapped fashion is termed as the weighted overlap-add method (WOLA), and is presented in [10]. For ideal signal reconstruction, it is necessary that

$$\sum_{r=-\infty}^{\infty} w(n - r\Gamma) f(n - r\Gamma) = 1.$$

So, given say a Hann window, $w(n)$ and $f(n)$ can be derived as the square root function. The complete time domain source estimates are obtained via

$$y_l(n) = \sum_{r=-\infty}^{\infty} y_l^{(r)}(n - r\Gamma), \quad (12)$$

for $0 \le l \le L-1$. The proposed algorithm steps are:

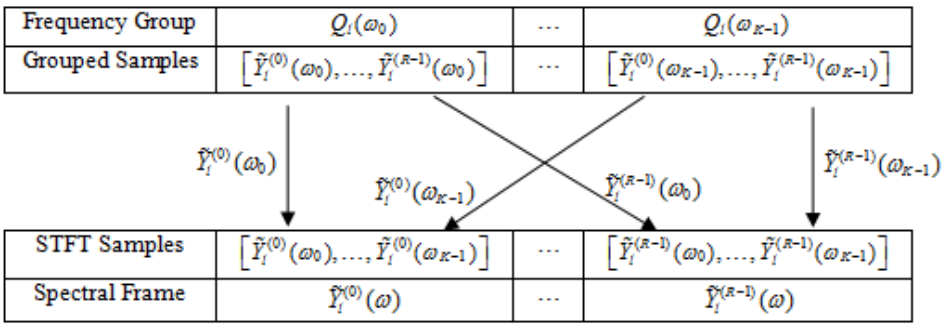

Fig. 2. An output frequency group $Q_l(\omega_k)$ is decomposed by returning frequency samples $\tilde{Y}_l^{(r)}(\omega_k)$ to their respective STFTs, forming the spectral frames $\tilde{Y}_l^{(r)}(\omega)$ after solving the permutation problem.

1. Extract the rth windowed data frames of length N using (3) to form the vector frames $\mathbf{x}^{(r)}(n) = [x_0^{(r)}(n), x_1^{(r)}(n), \ldots, x_{L-1}^{(r)}(n)]^T$, for $0 \le r \le R-1$, using the input window $w(n)$.
2. Take the K-point DFTs of the frames translated to time zero, resulting in the spectral frames $\tilde{\mathbf{X}}^{(r)}(\omega) = [\tilde{X}_0^{(r)}(\omega), \tilde{X}_1^{(r)}(\omega), \ldots, \tilde{X}_{L-1}^{(r)}(\omega)]^T$, for $0 \le r \le R-1$.
3. Form the frequency groups $Z_l(\omega_k)$, for all $0 \le k \le K-1$, ending up with the vectors (6) as illustrated in Fig. 1.
4. Apply the I-BSS algorithm of choice such as JADE [1] or RADICAL [2] to each of the vectors of (6), resulting in the output vectors given by (8).
5. Solve the possible permutation of the outputs using an algorithm such as that given by (9).
6. Translate the output samples from the frequency groups $Q_l(\omega_k)$, for $0 \le l \le L-1$ and $0 \le k \le K-1$, to their respective spectral frames as illustrated in Fig. 2, resulting in the vectors of (10).
7. Apply the inverse DFT to each of the output spectral frames and translate each output to time $r\Gamma$ after applying the synthesis window $f(n)$ to give the output vectors of (11).
8. Construct the complete time domain source estimates using (12).

III. Simulation Results

If $\mathbf{W}(\omega_k)$ is the optimal de-mixing matrix to $\mathbf{H}(\omega_k)$, then the overall transformation matrix $\mathbf{G}(\omega_k) = \mathbf{W}(\omega_k)\mathbf{H}(\omega_k)$ can be represented as

$$\mathbf{G}(\omega_k) = \mathbf{\Delta}(\omega_k)\mathbf{P}_k,$$

for the diagonal matrix $\mathbf{\Delta}(\omega_k) = \mathrm{diag}(\alpha_{11}e^{-\tau_1}, \alpha_{22}e^{-\tau_2}, \ldots, \alpha_{LL}e^{-\tau_N})$ and the permutation $\mathbf{P}_k$. In $\mathbf{\Delta}(\omega_k)$, α_{ij} for $0 \le i, j \le L-1$, is the resulting scaling ambiguity of the source estimates and $e^{-\tau_l}$ is the resulting delay after processing. If $\mathbf{G}(\omega_k)$ is as defined above, then the sources are extracted without any ISI and ICI. Practically, such a result is rare due to a number of reasons, like an insufficient number of deconvolution coefficients, and as such there is always some amount of ISI and ICI in the extracted sources. Therefore, we can quantify the performance of a source deconvolution algorithm by a direct measure of the net residual ISI and ICI as in [8, 9]. That is

$$\mathrm{P}_M\left[\mathbf{G}(\omega_k)\right] = \mathrm{ISI}\left[\mathbf{G}(\omega_k)\right] + \mathrm{ICI}\left[\mathbf{G}(\omega_k)\right].$$

If $\mathrm{g}_{ij}(\omega_k)$ is the row-I and column-j coefficient at frequency ω_k, the net residual interference at ω_k is given by

$$\mathrm{P}_M\left[\mathbf{G}(\omega_k)\right] = \sum_{i=0}^{L-1}\left(\sum_{j=0}^{L-1}\frac{\left|\mathrm{g}_{ij}(\omega_k)\right|^2}{\max_j\left|\mathrm{g}_{ij}(\omega_k)\right|^2} - 1\right) + \sum_{j=0}^{L-1}\left(\sum_{i=0}^{L-1}\frac{\left|\mathrm{g}_{ij}(\omega_k)\right|^2}{\max_i\left|\mathrm{g}_{ij}(\omega_k)\right|^2} - 1\right). \quad (13)$$

Two speech signals ($L = 2$) were used to give two mixtures ($M = 2$) A and B shown in Fig. 3 using a system of finite impulse response (FIR) filters of order $K = 6$. The FIR filters are obtained by truncating to six samples the impulse response of the transfer matrix

$$\mathbf{H}(z) = \begin{bmatrix} \dfrac{0.1+0.4z^{-1}}{0.3+0.6z^{-1}} & \dfrac{0.2+0.8z^{-1}}{0.4+0.7z^{-1}} \\ \dfrac{0.6+0.3z^{-1}}{0.9+0.4z^{-1}} & \dfrac{0.9+0.7z^{-1}}{0.7+0.8z^{-1}} \end{bmatrix}. \quad (14)$$

The coefficients are time-invariant since the matrix $\mathbf{H}(z)$ is fixed, depicting a mixing process of motionless sources. Table I shows the performance results of the proposed approach against the usual STFT approach in decibel (dB). The I-BSS algorithm of choice is RADICAL [2] due its simplicity and robustness to outliers. The results suggest that the frequency frame method allows the I-BSS algorithm to separate mixtures better than the STFT or spectral frame approach, usually with a net residual interference difference of about 9dB according to (13).

TABLE I
Net Residual Interference Measures in Decibel

Frequency	STFT	Frequency Frame
ω_0	-23	-32
ω_1	-19	-28
ω_2	-21	-29
ω_3	-21	-30
ω_4	-25	-31
ω_5	-23	-33

Fig. 3 shows the time domain plots of the original sources as well as their estimates. The estimates' temporal structures closely resemble those of the originals, albeit scaling ambiguities inherent to I-BSS algorithms. However, there is an observable amount of noise/discrepancies which are due to an insufficient number of deconvolution filters, i.e. $K = 6$. This means that the I-BSS algorithm is applied to only six frequency bands.

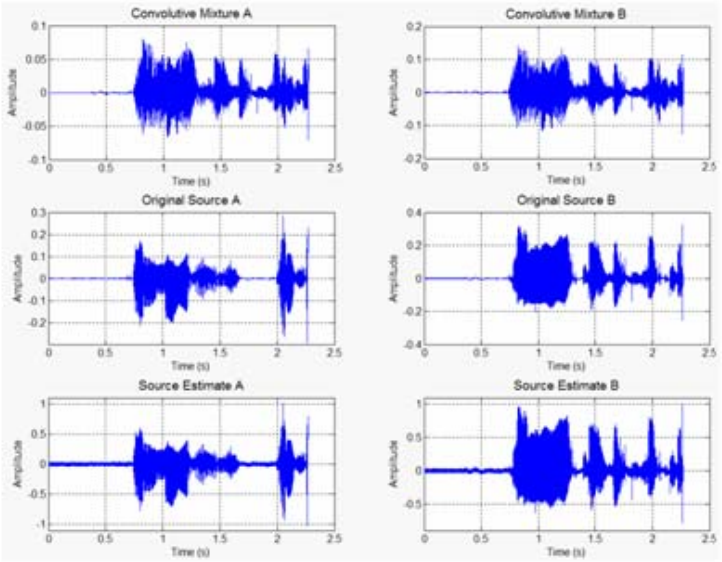

Fig. 3. The extracted source estimates via the frequency grouping approach closely resemble the original sources albeit some observable discrepancies which are due to insufficient deconvolution due to a limited number of deconvolution filters, $K = 6$.

Key to the STFT approach is that the window length should be less than or equal to the time when the frequency components are stationery. We illustrate the relative performance of the frequency grouping approach against the STFT approach for varying window lengths in Fig. 4.

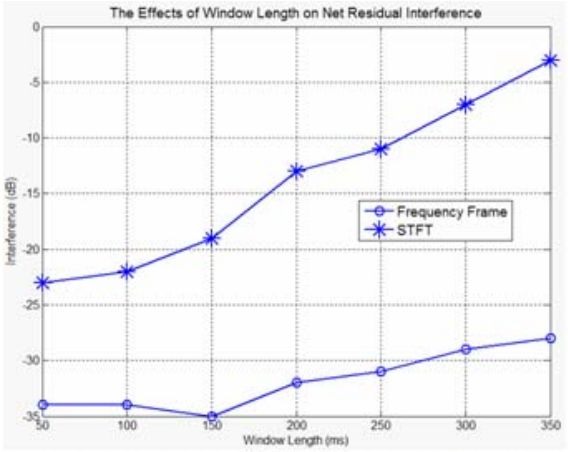

Fig. 4. The frequency grouping approach has a lower net residual interference at all window lengths than the STFT method. For time periods greater than 50ms, the performance of the STFT method degrades much more significantly than that of frequency grouping, illustrating its robustness to the characteristics of the underlying source signals.

For the used speech signals, when the STFT window length is greater than ~50ms, the STFT approach results in a significantly higher increase of the net residual interference as measured by (13) relative to the proposed method. The frequency grouping method does seem to eventually degrade in performance as the window length is increased, but this is mainly due to smaller values of R, or frequency samples per group $Z_l(\omega_k)$, $0 \le l \le L-1$. This compromises the data mining convergence for identifying the optimal basis vectors of the de-mixing matrix $\mathbf{W}(\omega_k)$.

IV. Discussion

Blind source separation of convolutive mixtures is usually tackled in the frequency domain to avoid the more complex time domain convolutive model. The usual approach is to apply several STFTs and by assuming that the frequency components are fairly stationery over the window length, the mixing process is approximately instantaneous over the spectral frames of the STFT. This approach means that it is necessary to know the spectral properties prior to source separation so as to use an optimal window length.

Instead of making assumptions about the underlying sources, the BSS problem can be divided into two types of physically stationery sources (considered here) and moving sources. Given fairly motionless sources, the convolution kernels are time-invariant and this allows us to form frequency gatherings from the STFT spectral frames, forming a truly instantaneous problem. Simulation results illustrate that this is a better approach and is more consistent over varying window lengths of the STFT.

References

[1] J. F. Cardosso and A. Souloumiac, "Blind Beamforming for Non-Gaussian Signals, " *IEE Proceedings Part F*, VOl. 140, No. 6, pp 362-370, 1993.

[2] E. G. L. Miller and J. W. Fisher III, "ICA using Spacing Estimates of Entropy, "*Journal of Machine Learning Research*, vol. 4, pp. 1271-1295, 2003.
[3] A. Hyvarinen, "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis, " *IEEE Transactions on Neural Networks*, vol. 10, no. 2, pp. 626-634, 1999.
[4] S. Amari, A. Cichocki and H. H. Yang, "A new Learning Algorithm for Blind Signal Separation, "*Advances in Neural Information Processing Systems*, vol. 8, pp. 752-763, 1996.
[5] S. Ikeda and N. Murata, "A Method of ICA in Time-Frequency Domain, " *In Proc. ICA*, pp. 365-371, 1999.
[6] K. Rahbar and J. P. Reilly, "A frequency Domain Method for Blind Source Separation f Convolutive Audio Mixtures, " *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, September 2005.
[7] S. Sanei, W. Wenwu and J. A. Chambers, "A Coupled HMM for Solving the Permutation Problem in Frequency Domain BSS, " *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 565-568, May 2004.
[8] C. Mejuto, A. Dapena and L. Castedo, "Frequency Domain Informax for Blind Separation of Convolutive Mixtures, " *Proceedings of ICA*, pp. 315-320, Hensinki, Finland, June 2000.
[9] A. Dapena and C. Serviere, " A Simplified Frequency-Domain Approach for Blind Separation of Convolutive Mixtures, " *Proceedings of ICA*, San Diego, USA, pp. 569-574, 2001.
[10] R. Crochiere, "A Weighted Overlapp-add Method for Short Time Fourier Transform Analysis/Synthesis, " *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 1, pp. 99-102, January 2003.