Frequency-Domain Blind Source Separation employing a Non-Uniform DFT

Diego B. Haddad
CEFET-RJ/COPPE-UFRJ
Telecommunications Coord. / PADS
Email: diego@pads.ufrj.br

Mariane R. Petraglia
COPPE/UFRJ
Electrical Engineering Dept.
Email: mariane@pads.ufrj.br

Paulo Bulkool Batalheiro
UERJ
Dept. of Electronic and
Telecommunications Engineering
Email: bulkool@pads.ufrj.br

***Abstract*—Blind source separation (BSS) techniques have been extensively investigated in the last years, due to their large number of applications. There are two main approaches employed for such techniques: in the time-domain and in the frequency-domain. In this paper, we propose the use of a non-uniform DFT transform as a strategy for improving the behavior of two frequency-domain BSS algorithms.**

I. Introduction

Extensively investigated in the last years, blind source separation techniques in the convolutive context (referred to here as CBSS, for *Convolutive Blind Source Separation*) have as goal to recover the original source signals from reverberant mixtures, with no other information about the sources (originating the adjective *blind*). Music transcription, temporal series analysis, remote sensing, audio edition and speech recognition are potential applications for such techniques.

Most of the recent CBSS techniques consider convolutive configurations where the mixture system is composed of FIR (finite impulse response) filters with thousand of coefficients. It is possible to divide most of these techniques in two classes: frequency-domain techniques ([1], [2], [3], [4]) and time-domain techniques ([4], [5]). These last algorithms employ non-trivial generalizations of the cost functions used in ICA (independent component analysis [6]), tend to present source estimates with less artifacts, and imply, in general, in a high computational cost.

When implemented in the frequency-domain, almost all methods employ data windowing followed by a DFT (corresponding to a short time Fourier transform, STFT). However, it is known that the energy of most input signals is concentrated at the low frequency components and that the mixture filters have non-flat frequency responses (usually they present lowpass characteristics). Therefore, the use of uniform transforms might not be the best strategy. Usually, the use of a small number of high-frequency components of the sources turns it difficult their separation at such frequencies. On the other hand, since the mixing filters normally present larger reverberation time at low-frequencies [7], it is desirable that the separation system assigns more resources at the low part of the spectrum.

In this paper, we propose the use of non-uniform transforms for improving the performances of two recently developed frequency-domain BSS algorithms. Without loss of generality, we suppose that there are two sources and two mixtures (determined case).

II. Non-uniform DFT (NDFT)

In [8], non-uniform extensions of the standard DFT were proposed. Considering $\mathbf{F}_\mathrm{K}$ the $K \times K$ Van der Monde NDFT matrix given by

$$\mathbf{F}_\mathrm{K} = \begin{bmatrix} 1 & z_0^{-1} & \cdots & z_0^{-K+1} \\ 1 & z_1^{-1} & \cdots & z_1^{-K+1} \\ \vdots & \ddots & & \vdots \\ 1 & z_{K-1}^{-1} & \cdots & z_{K-1}^{-K+1} \end{bmatrix}, \quad (1)$$

different types of non-uniform DFTs can be obtained by choosing the locations of z_k (in particular, the uniform DFT can be obtained by setting $z_k = e^{j2\pi k/K}$).

Inspired by the Warped DFT (WDFT) proposed in [9], we set:

$$z_k = e^{j\mathrm{atan}\left[\frac{1-\alpha}{1+\alpha}\tan(\frac{2\pi k}{K})\right]}, \quad (2)$$

where α is a factor that controls the non-uniform characteristic of the NDFT transform ($\alpha = 0$ implies the standard DFT).

III. BSS-EHOD Method

Frequency-domain BSS (FD-BSS) techniques are commonly used due to their less demanding computational cost requirements. These techniques can obtain excellent results when the number of bins K is greater than the length M of the mixing filters ([10] considers as necessary the condition $K >> M$ for a good separation).

Most of the FD-BSS algorithms exploit the property that convolutions in the time domain convert to products in the frequency domain. Supposing that each mixture bin is an instantaneous mixture of the corresponding sources bins, instantaneous BSS techniques can then be employed in each bin. This simple but powerful idea presents three inconveniences: *i)* if the STFT is employed, the product in the frequency domain corresponds, in the time domain, to a circular convolution and not to the desired linear convolution; *ii)* the permutation problem, inherent to instantaneous BSS techniques, becomes non-trivial when we need to distinguish, in a consistent way along the bins, the estimates that belong to each source and *iii)* the scaling ambiguity, also inherent to BSS techniques, can generate an intolerable source filtering, randomly emphasizing

(or attenuating) some bins.

For easing *(ii)*, several approaches were proposed such as constraining the maximum lengths of the separation filters in the time-domain ([10]), using the envelope correlation ([11], [12]) or estimates of the directions of arrival ([13], [1]), or even employing a hybrid approach ([2]). We chose the EHOD (Exploration of High-Order Dependencies), proposed in [3], which employs the multidimensional *score* function that, as it preserves the statistics dependencies of higher orders[1] among the bins during the iterations, ensures coherent estimates at the end of the optimization process.

The problem *(i)* requires a large K, such as to obtain a good approximation for the linear convolution from the circular convolution. We verify the possibility of improving the algorithm performance when K is reduced.

If $s_i(n)$ is the n-th sample of the i-th mixture and $h_{j,k}$ is the filter impulse response corresponding to the multiple paths from the k-th source to the j-th sensor, we have that $x_p(n)$ (corresponding to the n-th sample of the p-th mixture) can be written as:

$$x_p(n) = \sum_{m=1}^{2} \sum_{k=0}^{M-1} h_{p,m}(k) s_m(n-k), \tag{3}$$

where M, as defined before, is the length of the mixture filters.

The EHOD method initially applies a STFT to the mixtures (usually employing a Hanning window of length K and sliding of $K/4$ samples among frames), supposing instantaneous mixtures ($M = 1$) in each frequency bin. This allows the application of a 2×2 separation matrix to the mixture samples in each bin in order to estimate the bins values of each source, considering the TITO (two-input two-output) case. With $\mathbf{W}^{(k)}$ the 2×2 separation matrix (constant along the frames, supposing that the mixture system is stationary) of the k-th bin, it is possible, from an initial guess[2], to recur to an iterative procedure for estimating such matrix. Denoting $x_i^{(k)}(m)$ the k-th bin of the i-th mixture in the m-th frame, we can find the estimates $y_j^{(k)}(m)$ from:

$$\begin{bmatrix} y_1^{(k)}(m) \\ y_2^{(k)}(m) \end{bmatrix} = \mathbf{W}^{(k)} \begin{bmatrix} x_1^{(k)}(m) \\ x_2^{(k)}(m) \end{bmatrix} \tag{4}$$

At each iteration, we apply the following recursion:

$$\mathbf{W}_{\text{new}}^{(k)} = \mathbf{W}_{\text{old}}^{(k)} + \mu \Delta \mathbf{W}^{(k)}, \tag{5}$$

where μ is the learning factor (typically smaller or equal to 0.1). The elements of $\Delta\mathbf{W}^{(k)}$ can be calculated from the expression:

$$\Delta w_{ij}^{(k)} = \sum_{l=1}^{2} \left\{ \delta(i-l) - \mathrm{E}\left[\chi^{(k)}\right] \right\} w_{lj}^{(k)}, \tag{6}$$

where $\mathrm{E}[\cdot]$ is the statistic mean operator (which is applied along the frames), with $\chi^{(k)}$ and $\phi^{(k)}$ (a multidimensional *score* function) defined as:

$$\chi^{(k)} = \phi^{(k)}\left(y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(k)}\right) y_l^{*(k)} \tag{7}$$

$$\phi^{(k)}\left(y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(k)}\right) = \frac{y_i^{(k)}}{\sqrt{\sum_{k=1}^{K} \left|y_i^{(k)}\right|^2}} \tag{8}$$

At the end of all iterations, the minimum distortion principle [14] is applied by means of the following modification in all matrices:

$$\mathbf{W}^{(k)} \leftarrow \text{diag}\left[\left(\mathbf{W}^{(k)}\right)^{-1}\right] \mathbf{W}^{(k)}, \tag{9}$$

where the operator $\text{diag}(\cdot)$ zeroes the elements which are not in the main diagonal. Supposing that the separation is reasonable and (without loss of generality) that there is no permutation, we can approximate the matrix $\mathbf{W}^{(k)}$ as:

$$\mathbf{W}^{(k)} \approx \mathbf{\Delta}^{(k)} \left[\mathbf{H}^{(k)}\right]^{-1}, \tag{10}$$

where $\mathbf{\Delta}^{(k)}$ is a diagonal matrix which corresponds to scaling and $\mathbf{H}^{(k)}$ is the mixture matrix, that supposes instantaneous mixtures in each bin. Therefore, we have:

$$\text{diag}\left[\mathbf{W}^{(k)}\right]^{-1} \mathbf{W}^{(k)} \approx \text{diag}\left[\mathbf{H}^{(k)}\right] \left(\mathbf{H}^{(k)}\right)^{-1}, \tag{11}$$

from which we observe that the minimum distortion principle results in a reasonable scaling (although not ideal) instead of arbitrary.

IV. GABSOS Method

Using separation filters of length K, the GABSOS (Generalization of Blind Source Separation Algorithms Based on Second-Order Statistics) [4] obtains such filters ($w_{i,j}$) so that the sources estimates are given by:

$$y_p(n) = \sum_{q=1}^{2} \sum_{k=0}^{K-1} w_{p,q}(k) x_q(n-k). \tag{12}$$

A short description of this algorithm is presented as follows. Let

$$\underline{\mathbf{X}}_p(m) = \text{diag}\left\{\mathbf{F}_{4\mathrm{K}}[x_p(mK-3K) \ \ldots \ x_p(mK+K-1)]^T\right\}, \tag{13}$$

where $(\cdot)^T$ is the transpose operator. To work in the frequency domain, it is convenient to define the following input matrix:

$$\underline{\mathbf{X}}(m) = [\underline{\mathbf{X}}_1(m) \underline{\mathbf{X}}_2(m)]. \tag{14}$$

and the coefficient matrix:

$$\underline{\mathbf{W}}_{pq}(m) = \text{diag}\left\{\mathbf{F}_{4K}[w_{pq,0}, \ldots, w_{pq,K-1}, 0, \ldots, 0]^T\right\} \tag{15}$$

We should emphasize that, after each iteration of the algorithm (based on the natural gradient), we should convert the coefficients to the time-domain so as to zero the last coefficients, in order to limit the order of the filter and guarantee the implementation of a linear convolution. We also define:

$$\underline{\mathbf{W}} = \begin{bmatrix} \underline{\mathbf{W}}_{11} & \underline{\mathbf{W}}_{12} \\ \underline{\mathbf{W}}_{21} & \underline{\mathbf{W}}_{22} \end{bmatrix}, \tag{16}$$

$$\mathbf{S}_{\text{xx}} = \underline{\mathbf{X}}^H \mathbf{F}_{4K} \left[\mathbf{0}_{3K\times3K} \quad \mathbf{0}_{3K\times3K} \mathbf{F}_{4K}^{-1} \underline{\mathbf{X}},\right], \tag{17}$$

[1] And not only the correlation, which is not enough to preserve the consistency among the frequency bins of the estimates.

[2] Usually a whitening matrix is employed.

$$\mathbf{G}_{4K} = \mathbf{F}_{4K} \begin{bmatrix} \mathbf{I}_{2K\times 2K} & \mathbf{0}_{2K\times 2K} \\ \mathbf{0}_{2K\times 2K} & \mathbf{0}_{2K\times 2K} \end{bmatrix} \mathbf{F}_{4K}^{-1}, \tag{18}$$

$$\mathbf{G}_{8K} = \begin{bmatrix} \mathbf{G}_{4K} & \mathbf{0}_{4K\times 4K} \\ \mathbf{0}_{4K\times 4K} & \mathbf{G}_{4K} \end{bmatrix}, \tag{19}$$

$$\mathbf{S}_{\text{yy}} = \underline{\mathbf{W}}^H \mathbf{G}_{8K} \mathbf{S}_{\text{xx}} \mathbf{G}_{8K} \underline{\mathbf{W}}, \tag{20}$$

$$\mathbf{L} = \begin{bmatrix} \mathbf{F}_{4K} \begin{bmatrix} \mathbf{I}_{D\times D} \\ \mathbf{0}_{4K-D\times D} \end{bmatrix} & \mathbf{0}_{4K\times D} \\ \mathbf{0}_{4K\times D} & \mathbf{F}_{4K} \begin{bmatrix} \mathbf{I}_{D\times D} \\ \mathbf{0}_{4L-D\times D} \end{bmatrix} \end{bmatrix}, \tag{21}$$

$$\mathbf{L}_{\mathbf{I}} = \begin{bmatrix} \mathbf{1}_{4L\times 1} & \mathbf{0}_{4K\times 1} \\ \mathbf{0}_{4L\times 1} & \mathbf{1}_{4K\times 1} \end{bmatrix} \tag{22}$$

where D is an arbitrary parameter (which should satisfy the condition $1 \leq D \leq K$), $(\cdot)^H$ is the hermitian operator and $\mathbf{0}_{i\times j}$, $\mathbf{I}_{i\times j}$, $\mathbf{1}_{i\times j}$ are, respectively, the null matrix, the identity matrix and the matrix with all elements equal to 1, all of dimension $i \times j$. The natural gradient of the adopted cost function is given by:

$$\nabla_{\underline{\mathbf{W}}}^{\text{NG}} \Im = \sum_{i=1}^{b} \frac{2}{b} \mathbf{G}_{8K} \underline{\mathbf{W}} \mathbf{L} \mathbf{L}^H \left\{ \mathbf{S}_{\text{yy}} - \text{bdiag}(\mathbf{S}_{\text{yy}}) \right\} \mathbf{L}$$

$$. \left\{ \text{bdiag}(\mathbf{L}^H \mathbf{S}_{\text{yy}} \mathbf{L}) \right\}^{-1} \mathbf{L}^H, \tag{23}$$

where the bdiag $\{\cdot\}$ operator on a partitioned block matrix consisting of several submatrices sets all submatrices on the off-diagonals to zero.

The matrix $\underline{\mathbf{W}}$ has redundancies, which are removed after each iteration through the transformation $\underline{\mathbf{W}}' = \underline{\mathbf{W}} \mathbf{K}_{\mathbf{I}}$, with the matrix $\underline{\mathbf{W}}'$ having two columns, each one with $2K$ rows. The natural gradient of $\underline{\mathbf{W}}'$ is:

$$\nabla \underline{\mathbf{W}}'^{\text{NG}} = \mathbf{G}'_{8K} \nabla_{\underline{\mathbf{W}}}^{\text{NG}} \Im \mathbf{L}_{\mathbf{I}}, \tag{24}$$

with:

$$\mathbf{G}'_{4K} = \mathbf{F}_{4K} \begin{bmatrix} \mathbf{I}_{K\times K} & \mathbf{0}_{K\times 3K} \\ \mathbf{0}_{3K\times K} & \mathbf{0}_{3K\times 3K} \end{bmatrix} \mathbf{F}_{4K}^{-1}, \tag{25}$$

and

$$\mathbf{G}'_{8K} = \begin{bmatrix} \mathbf{G}'_{4K} & \mathbf{0}_{4K\times 4K} \\ \mathbf{0}_{4K\times 4K} & \mathbf{G}'_{4K} \end{bmatrix}. \tag{26}$$

V. Performance Measures

In [15], three performance measures for source separation (in noise-free case): SIR (signal-interference ratio, the most important of the three), SAR (signal-artifact ratio) e SDR (signal-distortion ratio) were presented. These measures are vastly employed in the recent literature, and, consequently, used in this work. It should be emphasized that the ISI (intersymbol interference) is not a useful measure in the source separation for convolutive mixtures (only in the source separation for instantaneous mixtures or in blind deconvolution), since it is important to preserve the temporal structure of the sources (filtered versions of the sources are allowed).

Decomposing the i-th estimate y_i of the sources in three components (for more details about this decomposition, see [15]):

$$y_i = s_{\text{desired}} + e_{\text{interf}} + e_{\text{artif}}, \tag{27}$$

SIR, SAR and SDR are defined as [3]:

$$\text{SIR} = 10\log_{10} \frac{||s_{\text{desired}}||^2}{||e_{\text{interf}}||^2}, \tag{28}$$

[3]We made a simplification when we assumed that the sensors do not introduce noise in the mixtures.

TABLE I
SIR, SAR and SDR measurements (in dB) for different values of α.

α	SIR	SAR	SDR
0	9.3297	19.9848	8.7531
0.0025	9.3432	20.0037	8.7673
0.005	9.4946	20.2316	8.9398
0.0075	9.7855	20.58	9.256
0.01	10.2763	21.1436	9.7834
0.0125	10.7723	21.7602	10.3183
0.015	11.0727	22.138	10.6423
0.0175	11.2103	22.2531	10.7851
0.02	11.2832	22.4297	10.8737
0.0225	11.2652	22.6684	10.8875
0.025	11.1309	22.7502	10.7765

TABLE II
SIR measurements (in dB) for different values of α and μ.

α	$\mu = 10^{-6}$	$\mu = 10^{-5}$	$\mu = 10^{-4}$
0	2.0743	9.3785	1.4799
0.001	7.942	7.7514	6.7462
0.01	5.2732	9.6256	3.5903
0.05	2.3612	8.3299	9.0171
0.1	1.6491	8.9129	8.9041

$$\text{SAR} = 10\log_{10} \frac{||s_{\text{desired}} + e_{\text{interf}}||^2}{||e_{\text{artif}}||^2}, \tag{29}$$

$$\text{SDR} = 10\log_{10} \frac{||s_{\text{desired}}||^2}{||e_{\text{interf}} + e_{\text{artif}}||^2}. \tag{30}$$

VI. Simulation Results

In both experiments presented in this section, two speech signals were employed: a male and a female. The length of the mixing filters was $M = 8$.

A. Experiment 1

In this experiment, the number of bins was $K = 8$. The EHOD algorithm was implemented in its off-line form, with 2500 iterations and $\mu = 0.1$. The results obtained in this experiment are presented in Table 1. This table shows the improvement in the SIR, SAR and SDR is approximately 2 dB for $\alpha = 0.02$.

B. Experiment 2

The GABSOS method in the frequency-domain does not introduce artifacts and distortions in the estimates, as opposed to the EHOD method. Therefore, the SAR and SDR measurements are very high and carry little information. For this reason, we present only the SIR values for this experiment. The GABSOS algorithm was implemented in its on-line version and the value of the step-size μ influences the final result. Table 2 contains the final SIR obtained for different values of α and μ (the initial SIR was 1.67 dB).

In the average, non-zero values of α generate a larger SIR than when using the conventional DFT. It can also be observed that a smaller variation in the results (with respect to the values of μ) is obtained when the NDFT is employed, which might be another reason for using non-uniform transforms. Figure 1 shows the increase of SIR in the same configuration (but using another μ).

VII. Conclusions

In this paper, a non-uniform DFT (NDFT) was used as an alternative to the employment of uniform transforms for the purpose of source separation, both in the on-line form and off-line form. We verified that a 2 dB improvement is obtained in the SIR, SAR and SDR with the WDFT when compared to the DFT results. In the on-line configurations, the use of the WDFT yielded more stable

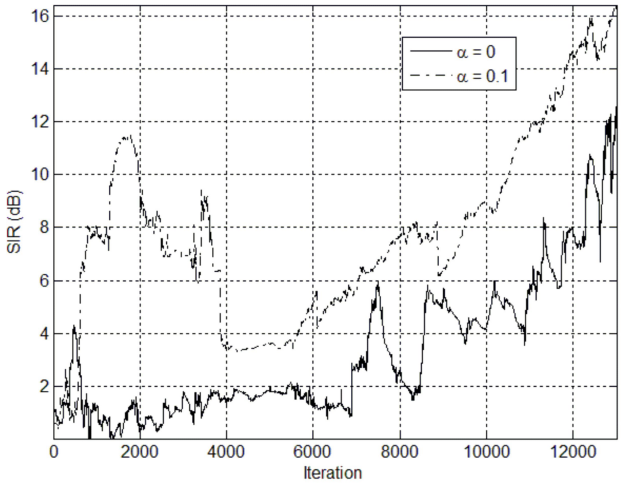

Fig. 1. SIR (dB) evolution of GABSOS method with $\mu = 2.10^{-6}$.

performances of the algorithms. The optimal use of the WDFT requires the selection of the parameter α, which controls the non-uniformity of the NDFT. Such problem will be explored in our future research.

REFERENCES

[1] M. Z. Ikram and D. R. Morgan, *A Beamforming Approach to Permutation Alignment for Multichannel Frequency-Domain Blind Speech*, Proc. ICASSP, pp. 881-884, 2002.

[2] H. Sawada, R. Mukai, S. Araki and S. Makino, *A Robust and Precise Method for Solving the Permutation Problem of Frequency-domain Blind Source Separation*, IEEE Transactions on Speech and Audio Processing, vol. 12, pp. 530-538, 2004.

[3] T. Kim, H. T. Attias, S.-Y. Lee, T.-W. Lee, *Blind Source Separation Exploiting Higher-Order Frequency Dependencies*, IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, pp. 70-79, 2007.

[4] H. Buchner, R. Aichner and W. Kellerman, *A Generalization of Blind Source Separation Algorithms for Convolutive Mixtures Based on Second-Orders Statistics*, IEEE Transactions on Speech and Audio Processing,vol. 13, pp. 120-134, 2005.

[5] J. Thomas, Y. Deville and S. Hosseini, *Time-domain Fast Fixed-Point Algorithms for Convolutive ICA*, IEEE Signal Processing Letters, vol. 13, pp. 228-231, 2006.

[6] A. Hyvarinen, J. Karhunen, E. Oja, *Independent Component Analysis, John Wiley and Sons*, 2001.

[7] S. Araki, S. Makino, R. Aichner, T. Nishikawa and H. Saruwatari, *Subband-Based Blind Separation for Convolutive Mixtures of Speech*, IEICE Trans. Fundamentals, vol. E88-A, no. 12, pp. 3593-3603, 2005.

[8] S. Bagchi and S. K. Mitra, *Nonuniform Discrete Fourier Transform and its Signal Processing Applications*, Norwell, MA: Kluwer, 1999.

[9] A. KMakur and S. K. Mitra, *Warped Discrete-Fourier Transform: Theory and Applications*, IEEE Transactions on Circuits and Systems, vol. 48, no. 9, pp. 1086-1093, 2001.

[10] L. Parra and C. Spence, *Convolutive Blind Separation of Non-Stationary Sources*, IEEE Transactions on Speech and Audio Processing, vol. 8, pp. 320-327, 2000.

[11] J. Anemller and B. Kollmeir, *Amplitude Modulation Decorrelation for Convolutive Blind Source Separation*, Proc. ICA, pp. 215-220, 2000.

[12] N. Murata, S. Ikeda and A. Ziehe, *An Approach to Blind Source Separation Based on Temporal Structure of Speech Signals*, Neurocomputing, vol. 41, pp. 1-24, 2001.

[13] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda and F. Itakura, *Evaluation of Blind Signal Separation Method using Directivity Pattern Under Reverberant Conditions*, Proc. ICASSP, pp. 3140-3143, 2002.

[14] K. Matsuoka, *Minimal Distortion Principle for Blind Source Separation*, SICE, pp. 2138-2143, 2002.

[15] E. Vincent, R. Gribonval and C. Fvotte, *Performance Measurement in Blind Audio Source Separation*, IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 4, pp. 1462-1469, 2006.