BLIND SIGNAL SEPARATION USING A CRITERION BASED ON PRINCIPLE OF MINIMAL DISTURBANCE

Uttachai Manmontri and Patrick A. Naylor

Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ, UK
Email: {uttachai.manmontri,p.naylor}@imperial.ac.uk

ABSTRACT

The concept underlying most on-line gradient-based algorithms for blind signal separation (BSS) is that the unknown demixing matrix is adjusted with an appropriate step-size in the direction of the gradient computed at each sample instant. Associated with these algorithms is a gradient noise problem. In this paper, we develop, from the on-line processing (OP) algorithm derived using the nonstationarity and nonwhiteness properties, a normalized algorithm in which the update of the demixing matrix is based on the minimal disturbance principle. We show that the resulting updates are in the same direction as those of the original algorithm but with a scaling factor whose upper bound is unity. We evaluate the convergence speed and robustness to gradient noise of the new algorithm.

1. INTRODUCTION

Blind signal separation (BSS) has received considerable attention recently in such diverse fields as signal processing and communications. The classical instantaneous mixing and demixing processes in the blind signal separation problem have the following descriptions

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

$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n) \quad (2)$$

where $\mathbf{s}(n) = [s_1(n), s_2(n), \ldots, s_{\mathcal{N}}(n)]^T$ is the source signals vector, $\mathbf{x}(n) = [x_1(n), x_2(n), \ldots, x_{\mathcal{N}}(n)]^T$ is the observed signals vector and $\mathbf{y}(n) = [y_1(n), y_2(n), \ldots, y_{\mathcal{N}}(n)]^T$ is the estimate of the source signals called the output signals vector, $\mathbf{A}$ is an $\mathcal{N} \times \mathcal{N}$ unknown mixing matrix, $\mathbf{W}$ is a corresponding demixing matrix to be computed, the superscript T denotes transposition and n is the sample index.

The solution to this problem is feasible using second-order statistics under the following assumptions

A1: $\mathbf{A}$ is a square matrix with rank $\mathcal{N}$.

A2: Source signals are zero mean, nonstationary and nonwhite processes with

(i) $E[s_i(n)] = 0, \forall\, i = 1, 2, \ldots, \mathcal{N}$

(ii) $E\left[s_i^2(n_u)\right] \neq E\left[s_i^2(n_v)\right], \exists\, i$ and $\exists\, u \neq v$

(iii) $E[s_i(n)s_i(n-\tau)] \neq 0, \exists\, i$ and $\exists\, \tau \neq 0$.

Also, each source signal is uncorrelated with

(iv) $E[s_i(n)s_j(n-\tau)] = E[s_i(n)]\, E[s_j(n-\tau)], \forall\, 1 \leq i \neq j \leq \mathcal{N}$ and $\forall\, \tau$

where $\tau = 0, 1, \ldots, \Gamma$ denotes the time lag with Γ being the maximum nonzero time lag, $E[\cdot]$ denotes the statistical expectation operator, $\forall$ denotes *for all* and $\exists$ denotes *for some*. A1 ensures the existence of all source signals to be observed in the form of $\mathbf{x}(n)$ by the rank of $\mathbf{A}$ and makes a solution to the problem feasible. A2 is a key assumption based on the nonstationarity and nonwhiteness properties that leads to the use of joint diagonalization of matrices as a criterion in the algorithms presented thereafter.

According to (1), (2) and A2, we obtain, under the framework of second-order statistics, the following relation

$$\hat{\Lambda}_{\mathbf{s}}^{(\tau)}(n) \doteq \hat{\mathbf{R}}_{\mathbf{y}}^{(\tau)}(n) = \mathbf{W}\hat{\mathbf{R}}_{\mathbf{x}}^{(\tau)}(n)\mathbf{W}^T \quad (3)$$

where $\hat{\Lambda}_{\mathbf{s}}^{(\tau)}(n), \hat{\mathbf{R}}_{\mathbf{x}}^{(\tau)}(n)$ and $\hat{\mathbf{R}}_{\mathbf{y}}^{(\tau)}(n)$ are, respectively, *current estimates* of the correlation matrices of source signals, observed signals and output signals at time lag τ, $\doteq$ denotes an equality up to scaling factor and permutations of signals represented by $\mathbf{WA} = \mathbf{DP}$ with $\mathbf{D}$ and $\mathbf{P}$ being a *nonsingular* diagonal matrix and permutation matrix of dimension $\mathcal{N} \times \mathcal{N}$, respectively.

It is seen from (3) that the BSS problem is one of finding $\mathbf{W}$ that jointly diagonalizes a set of estimated correlation matrices. In [1], the on-line processing (OP) algorithms is derived from (3) by using the current time-average correlation matrix as an estimate of the correlation matrix and by using the natural gradient method to minimize an appropriate cost function. Although it has been shown to separate nonstationary and/or nonwhite signals, the OP algorithms, like many other gradient-based algorithms, can suffer from the gradient noise problem when the computed gradient is either considerably small or large. One possibility to mitigate this problem is to choose an adaptive step-size that compensates for such small or large gradients. In this paper, a criterion based on the principle of minimal disturbance (see e.g. [2]) is used in the derivation of a normalized algorithm referred to as the normalized OP (NOP) algorithm. The NOP algorithm addresses the gradient noise problem and, as a result, exhibits fast convergence and good robustness to gradient noise.

2. OVERVIEW OF THE OP ALGORITHM

To pave the way to the derivation of NOP, we first give a brief overview of OP [1]. The OP algorithm relies on the joint diagonalization of a set of estimated correlation matrices in (3). The estimate of observed signals correlation matrix using a nonparametric recursive relation at a given time lag τ takes the form

$$\hat{\mathbf{R}}_{\mathbf{x}}^{(\tau)}(n) = \alpha\hat{\mathbf{R}}_{\mathbf{x}}^{(\tau)}(n-1) + (1-\alpha)\,\mathbf{x}(n)\mathbf{x}^T(n-\tau) \quad (4)$$

where α is a forgetting factor with $0 \leq \alpha < 1$. When the source signals are known to be stationary but nonwhite, the estimation of

observed signals correlation matrix takes the form

$$\hat{\mathbf{R}}_{\mathbf{x}}^{(\tau)}(n) = \frac{n-1}{n}\hat{\mathbf{R}}_{\mathbf{x}}^{(\tau)}(n-1) + \frac{1}{n}\mathbf{x}(n)\mathbf{x}^T(n-\tau). \quad (5)$$

A set of the above-defined matrices called the *current time-average correlation matrix* is used in the cost function J_{OP} which, after being minimized, gives a desired demixing matrix $\mathbf{W}$. The OP cost function is given by

$$J_{OP}(n) = J_{OP}(\mathbf{W}(n)) = \sum_{\tau=0}^{\Gamma} \beta^{(\tau)} J_{JD}^{(\tau)}(n) + \lambda_C J_C(n) \quad (6)$$

with $\lambda_C < 1$ being a small positive constraint constant, $J_{JD}^{(\tau)}(n)$ and $J_C(n)$ being, respectively, the joint diagonalization function at time lag τ and the constraint function, defined by

$$J_{JD}^{(\tau)}(n) = J_{JD}^{(\tau)}(\mathbf{W}(n)) = \left\| off\left(\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\right)\right\|_F^2 \quad (7)$$

$$J_C(n) = J_C(\mathbf{W}(n)) = \|diag\,(\mathbf{W}(n) - \mathbf{I})\|_F^2 \quad (8)$$

where $\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n) = \hat{\mathbf{R}}_{\mathbf{y}}^{(\tau)}(n) + \hat{\mathbf{R}}_{\mathbf{y}}^{(\tau)T}(n)$ is a symmetric matrix, called the *symmetric part* of the current time-average output signals correlation matrix $\hat{\mathbf{R}}_{\mathbf{y}}^{(\tau)}(n)$, *off* $(\cdot)$ is the matrix operator that returns a matrix with all its diagonal entries being zero, *diag* $(\cdot)$ is the matrix operator that returns a matrix with all its off-diagonal entries being zero, $\|\cdot\|_F$ denotes the Frobenius norm and $\beta^{(\tau)}$ is a positive weight satisfying $\sum_{\tau=0}^{\Gamma}\beta^{(\tau)} = 1$ and is generally set to $\frac{1}{\Gamma+1}$ giving balanced weighting for joint diagonalization. We note here that (3) still holds if all its correlation matrices are replaced by their corresponding symmetric parts. This provides a more compact form of the gradient at the same computational complexity. By using the natural gradient method [3], the demixing matrix update $\delta\mathbf{W}(n) = \mathbf{W}(n+1) - \mathbf{W}(n)$ is given by

$$\delta\mathbf{W}(n) = -\mu\left(\sum_{\tau=0}^{\Gamma}\beta^{(\tau)}\tilde{\nabla}_{\mathbf{W}} J_{JD}^{(\tau)}(n) + \lambda_C \tilde{\nabla}_{\mathbf{W}} J_C(n)\right) \quad (9)$$

where $\tilde{\nabla}_{\mathbf{W}}$ is the natural gradient operator with respect to $\mathbf{W}(n)$ and μ is a positive step-size. The negative sign in (9) means that the algorithm moves towards the minimum in the natural gradient descent direction. The natural gradients of $J_{JD}^{(\tau)}(n)$ and $J_C(n)$ are given, respectively, by

$$\tilde{\nabla}_{\mathbf{W}} J_{JD}^{(\tau)}(n) = 4\, off\left(\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\right)\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\mathbf{W}(n) \quad (10)$$

$$\tilde{\nabla}_{\mathbf{W}} J_C(n) = 2\, diag\,(\mathbf{W}(n) - \mathbf{I})\,\mathbf{W}^T(n)\mathbf{W}(n). \quad (11)$$

By initialing $\mathbf{W}$ with all its diagonal entries being unity so as to prevent the algorithm from being trapped by possible local minima induced by J_C and by choosing appropriate μ and λ_C, the OP algorithm will converge in the mean (first moment) to a minimum corresponding to the desired demixing matrix whose existence is ensured by A1, provided that the identifiability (separability) conditions in [1] hold and variations in the nonstationarity of the source signals are sufficiently slow.

We see that the update $\delta\mathbf{W}(n)$ moves in a descent direction of the average of $\Gamma+1$ natural gradients of $J_{JD}^{(\tau)}(n)$ and a natural gradient of $J_C(n)$. It can also be seen that such natural gradients do not only give direction but also magnitude which, whether small or large, needs to be compensated for by an appropriate step-size μ. Excessively small and large gradients can result in gradient noise problems ranging from slow convergence to divergence of the algorithm.

3. DERIVATION OF THE NOP ALGORITHM

In this section, we introduce the principle of minimal disturbance to the demixing process to form a criterion for the NOP algorithm. Developing NOP is thus analogous in some ways to developing the normalized least mean square (NLMS) algorithm from the least mean square (LMS) algorithm using the principle of minimal disturbance in the field of supervised adaptive filtering [2]. A common theme is that the update of an unknown parameter in the adaptive structure should be disturbed in a minimal fashion. By applying this principle to the demixing matrix, we are able to obtain an update that perturbs the process in a minimal fashion leading to a fast convergence and robustness of the developed algorithm.

Based on $J_{OP}(n)$, the corresponding update $\delta\mathbf{W}(n)$ can be written by using a set of components comprising $J_{JD}^{(\tau)}(n)$ and $J_C(n)$ as

$$\delta\mathbf{W}(n) = \sum_{\tau=0}^{\Gamma}\beta^{(\tau)}\delta\mathbf{W}_{JD}^{(\tau)}(n) + \lambda_C\delta\mathbf{W}_C(n) \quad (12)$$

where $\delta\mathbf{W}_{JD}^{(\tau)}(n)$ is an update based on $J_{JD}^{(\tau)}(n)$ and $\delta\mathbf{W}_C(n)$ is an update based on $J_C(n)$.

In light of the principle of minimal disturbance, every component of $\delta\mathbf{W}(n)$ in (12) should disturb the separation system in a minimal fashion. Let us consider an update $\delta\mathbf{W}_{JD}^{(\tau)}(n)$ based on a particular given $J_{JD}^{(\tau)}(n)$ and form the following constrained minimal disturbance problem

$$\text{minimize } \frac{1}{2}\left\|\delta\mathbf{W}_{JD}^{(\tau)}(n)\right\|_F^2 \text{ subject to } J_{JD}^{(\tau)}(n+1) = 0 \quad (13)$$

which defines a search for $\delta\mathbf{W}_{JD}^{(\tau)}$ at the sample index n that disturbs the separation process in a minimal manner while forcing $J_{JD}^{(\tau)}$ at one step ahead to be zero. To allow (13) to be differentiable with respect to $\delta\mathbf{W}_{JD}^{(\tau)}(n)$, we exploit the Taylor series expansion and neglect its high-order to estimate $J_{JD}^{(\tau)}(n+1)$ giving

$$J_{JD}^{(\tau)}(n+1) = J_{JD}^{(\tau)}(n) + \left\langle \tilde{\nabla}_{\mathbf{W}} J_{JD}^{(\tau)}(n) \,\middle|\, \delta\mathbf{W}_{JD}^{(\tau)}(n)\right\rangle \quad (14)$$

where $\langle \mathbf{X} \mid \mathbf{Y}\rangle = Trace\left(\mathbf{X}^T\mathbf{Y}\right)$. We note that it is more accurate to use the natural gradient in the Taylor series expansion when estimating $J_{JD}^{(\tau)}(n+1)$ from $J_{JD}^{(\tau)}(n)$ in a matrix environment.

Following the method of Lagrange multipliers, we first replace $J_{JD}^{(\tau)}(n+1)$ with (14) and then convert (13) to the following unconstrained problem

$$J(n) = J(\delta\mathbf{W}_{JD}^{(\tau)}(n)) = \frac{1}{2}\left\|\delta\mathbf{W}_{JD}^{(\tau)}(n)\right\|_F^2 + \lambda_L\left(J_{JD}^{(\tau)}(n) + \left\langle\tilde{\nabla}_{\mathbf{W}} J_{JD}^{(\tau)}(n)\,\middle|\,\delta\mathbf{W}_{JD}^{(\tau)}(n)\right\rangle\right) \quad (15)$$

where λ_L is the Lagrange multiplier.

Given $\underline{\hat{\mathbf{R}}}_{\mathbf{x}}^{(\tau)}(n)$ and $\mathbf{W}(n)$, we obtain the first-order conditions of (15) as

$$\tilde{\nabla}_{\delta\mathbf{W}_{JD}^{(\tau)}} J(n) = \delta\mathbf{W}_{JD}^{(\tau)}(n) + \lambda_L\tilde{\nabla}_{\mathbf{W}} J_{JD}^{(\tau)}(n) = \mathbf{0} \quad (16)$$

$$\frac{\partial J(n)}{\partial\lambda_L} = J_{JD}^{(\tau)}(n) + \left\langle\tilde{\nabla}_{\mathbf{W}} J_{JD}^{(\tau)}(n)\,\middle|\,\delta\mathbf{W}_{JD}^{(\tau)}(n)\right\rangle = 0 \quad (17)$$

and solve for λ_L by substituting $\delta\mathbf{W}_{JD}^{(\tau)}(n)$ from (16) into (17) giving

$$\lambda_L = \frac{J_{JD}^{(\tau)}(n)}{\left\|\tilde{\nabla}_{\mathbf{W}} J_{JD}^{(\tau)}(n)\right\|_F^2}. \tag{18}$$

Replacing λ_L in (16) with (18), we obtain the following optimal component

$$\delta\mathbf{W}_{JD}^{(\tau)}(n) = -J_{JD}^{(\tau)}(n)\frac{\tilde{\nabla}_{\mathbf{W}} J_{JD}^{(\tau)}(n)}{\left\|\tilde{\nabla}_{\mathbf{W}} J_{JD}^{(\tau)}(n)\right\|_F^2}. \tag{19}$$

A key feature of (19) is that the update $\delta\mathbf{W}_{JD}^{(\tau)}(n)$ obtained from the use of minimal disturbance principle mitigates the gradient noise by normalizing the natural gradient of $J_{JD}^{(\tau)}(n)$ with its squared Frobenius norm. In particular, we can interpret (19) as an update $\delta\mathbf{W}_{JD}^{(\tau)}(n)$ that moves towards the minimum of $J_{JD}^{(\tau)}(n)$ in the natural gradient descent (negative sign) direction with distance proportional to $J_{JD}^{(\tau)}(n)$ and then vanishes at the minimum. It is also seen that (19) is *adaptive* and *self-adjustable* in the sense that the update $\delta\mathbf{W}_{JD}^{(\tau)}(n)$ moves with a large step-size when the difference between the current point and the minimum point, or simply $J_{JD}^{(\tau)}(n)$, is large and it moves with a smaller step-size as the point becomes close to the minimum.

To simplify (19), we use (7) and (10) and expand (19) to

$$\delta\mathbf{W}_{JD}^{(\tau)}(n) = -\frac{\left\|off\left(\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\right)\right\|_F^2 \tilde{\nabla}_{\mathbf{W}} J_{JD}^{(\tau)}(n)}{16\left\|off\left(\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\right)\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\mathbf{W}(n)\right\|_F^2}. \tag{20}$$

By applying the inequality property of the squared Frobenius norm to the squared Frobenius norm terms in (20), we obtain $\frac{\left\|off\left(\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\right)\right\|_F^2}{\left\|off\left(\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\right)\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\mathbf{W}(n)\right\|_F^2} \geq \frac{\left\|off\left(\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\right)\right\|_F^2}{\left\|off\left(\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\right)\right\|_F^2\left\|\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\mathbf{W}(n)\right\|_F^2}$. Using this inequality and expanding the remaining $\tilde{\nabla}_{\mathbf{W}} J_{JD}^{(\tau)}(n)$, (20) is approximately simplified to

$$\delta\mathbf{W}_{JD}^{(\tau)}(n) \approx -\frac{off\left(\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\right)\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\mathbf{W}(n)}{4\left\|\underline{\hat{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}(n)\mathbf{W}(n)\right\|_F^2}. \tag{21}$$

We note that the squared Frobenius norm of (21) is always less than that of (20) due to the squared Frobenius norm inequality. Therefore, $\delta\mathbf{W}_{JD}^{(\tau)}(n)$ in (21) still obeys the principle of minimal disturbance. Following the above methodology, we similarly obtain

$$\delta\mathbf{W}_C(n) = -J_C(n)\frac{\tilde{\nabla}_{\mathbf{W}} J_C(n)}{\left\|\tilde{\nabla}_{\mathbf{W}} J_C(n)\right\|_F^2} \tag{22}$$

$$\approx -\frac{diag\left(\mathbf{W}(n)-\mathbf{I}\right)\mathbf{W}^T(n)\mathbf{W}(n)}{2\left\|\mathbf{W}^T(n)\mathbf{W}(n)\right\|_F^2}. \tag{23}$$

In order to control the rate of convergence, we introduce a scaling factor $\bar{\mu}$ to (12) and write

$$\delta\mathbf{W}(n) = \bar{\mu}\left(\sum_{\tau=0}^{\Gamma}\beta^{(\tau)}\delta\mathbf{W}_{JD}^{(\tau)}(n) + \lambda_C\delta\mathbf{W}_C(n)\right) \tag{24}$$

where $0 < \bar{\mu} \leq 1$ with 1 being an upper bound that still keeps the squared Frobenius norm of $\delta\mathbf{W}(n)$ in accordance with the minimal disturbance principle, $\delta\mathbf{W}_{JD}^{(\tau)}(n)$ and $\delta\mathbf{W}_C(n)$ can be obtained from their simplified forms in (21) and (23), respectively.

Comparing (24) with (9), we see that the NOP algorithm is clearly the normalized version of the OP algorithm. Both algorithms move in the same natural gradient descent direction but NOP employs additional normalized terms achieved at the additional cost of $2(\Gamma+2)\mathcal{N}^2$ multiplications.

4. NUMERICAL EXPERIMENTS

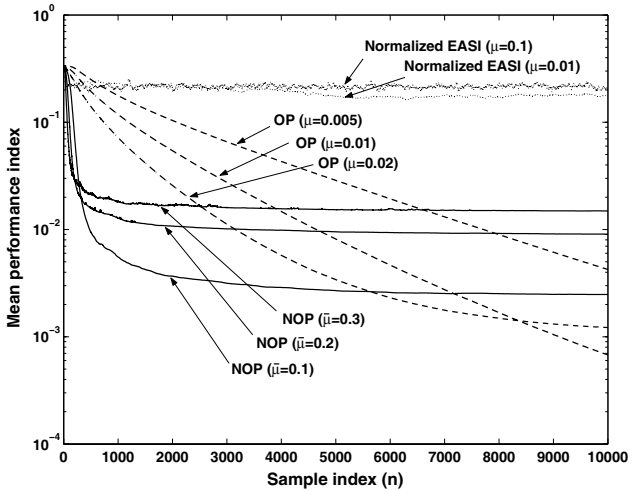

Fig. 1. Mean performance indices of the OP, NOP and normalized EASI algorithms obtained from the mixtures of two stationary but nonwhite source signals generated by AR(2) and AR(4) models.

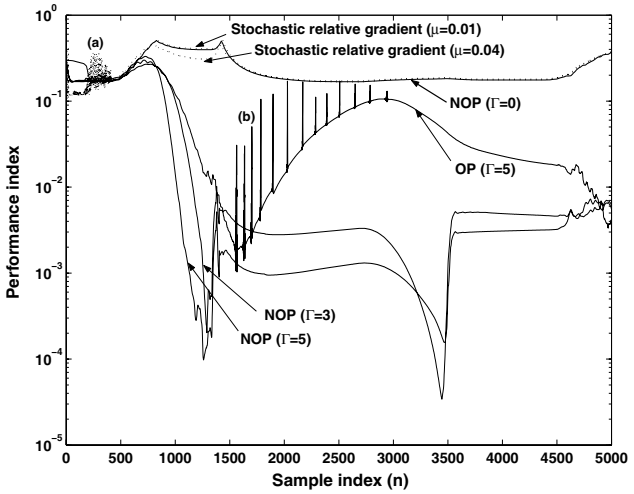

Fig. 2. Performance index of the OP, NOP and stochastic relative gradient algorithm [4] obtained from the mixtures of two speech signals. Examples of effect of gradient noise from (a) stochastic relative gradient ($\mu = 0.04$) and (b) OP ($\Gamma = 5$).

To evaluate the performance of the BSS algorithms, the closeness of $\mathbf{C} = [c_{ij}] = \mathbf{WA}$ to $\mathbf{DP}$ is measured.

We employ the performance index [5] which is defined by PI= $\frac{1}{2\mathcal{N}(\mathcal{N}-1)}\left(\sum_{i=1}^{\mathcal{N}}\left(\sum_{j=1}^{\mathcal{N}}\frac{c_{ij}^2}{\max_j\left(c_{ij}^2\right)}-1\right)+\sum_{j=1}^{\mathcal{N}}\left(\sum_{i=1}^{\mathcal{N}}\frac{c_{ij}^2}{\max_i\left(c_{ij}^2\right)}-1\right)\right)$ where $max_{i,j}\left(\cdot\right)$ is the maximum value for $1 \leq i, j \leq \mathcal{N}$. Accordingly, smaller values of PI indicate better performance.

Fig. 1 shows the PI averaged over 500 independent trials for two stationary but nonwhite source signals generated by autoregressive (AR) models of order two and four. For these signals, algorithms that employ the nonstationarity property only such as [6], [4], [7] when used with an estimate of the correlation matrix are not applicable. We therefore compare the OP and NOP algorithms, which are based on second-order statistics, with the normalized equivariant adaptive source separation via independence (EASI) algorithm [8], which is based on higher-order statistics. All entries of the mixing matrix are drawn from a normally distributed random process. We set Γ to four and λ_C to 0.01 for both OP and NOP. From Fig. 1, it is seen that both OP and NOP outperform the normalized EASI algorithm with the faster convergence from NOP. This is due to the fact that the distinctive difference between AR source signals is their nonwhiteness property on which OP and NOP rely rather than the probability density function of source signals on which normalized EASI relies. For AR source signals, their probability density functions can be slightly different. After the algorithms converge, an effect of large μ and $\bar{\mu}$ on an increase in misadjustment of NOP and normalized EASI can be seen in the figure. This suggests the need of a smaller step-size in order to obtain satisfactory misadjustment.

In the next simulation, the OP, NOP algorithms, which employ both the nonstationarity and nonwhiteness properties, and the stochastic relative gradient algorithm based on the maximum likelihood criterion [4], which employs the nonstationarity property only, are compared. The source signals are speech signals.[1] We set α to 0.999 for all algorithms, λ_C to 0.01 for OP and NOP, μ to 0.5 for OP, $\bar{\mu}$ to 0.2 for NOP and μ of the stochastic relative gradient algorithm is shown in the figure. In Fig. 2, it is seen that, at $\Gamma = 5$, NOP outperforms OP because of the normalized terms. All results from NOP except at $\Gamma = 0$ are better than the stochastic relative gradient algorithm because of the use of the nonwhiteness property in addition to the nonstationarity property. As Γ increases, the performance of NOP improves. The stochastic relative gradient and NOP algorithms give similar PI curve at $\Gamma = 0$ due to the fact that only the nonstationarity property of speech signals is used. We also see that the OP and the stochastic relative gradient algorithms exhibit large dynamic PI due to the effect of large computed gradient induced by the nonstationarity of speech signals and a large step-size. This effect is mitigated in the NOP algorithm.

5. DISCUSSION AND CONCLUSIONS

We have presented a normalized version of the on-line processing (OP) algorithm referred to as the normalized OP (NOP) algorithm for blind signal separation. The similarity between the OP/NOP and the well-known LMS/NLMS algorithms [2] is that the normalized algorithms can be derived from the principle of minimal disturbance and that the normalized terms are the terms that do not involve the criterion of the original algorithms i.e. the estimation error $e(n)$ for LMS [2], *off* $\left(\hat{\underline{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}\right)$ and *diag* $(\mathbf{W}-\mathbf{I})$ for OP, which, in fact, cannot be used as a normalized term since they approach zero when the unknown moves close to the minimum. Consider *off* $\left(\hat{\underline{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}\right)$ used as a criterion of OP. It turns out that the normalized term of this criterion, i.e. $\left\|\hat{\underline{\mathbf{R}}}_{\mathbf{y}}^{(\tau)}\mathbf{W}\right\|_F^2$, is dependent on $\mathbf{W}$ whereas the normalized term of NLMS is independent of the weight vector $\mathbf{w}$. This can be explained by the fact that the criterion used in NLMS is of the elements of $\mathbf{w}$ whereas, in NOP, the criterion is of the *quadratic* elements of $\mathbf{W}$ – hence the normalized term of NOP takes a form that is still dependent on $\mathbf{W}$ after the *first-order* gradient differentiation. The use of squared Frobenius norm of correlation matrix as a normalized term is given without derivation in [6], [9], [7]. This normalized term when used in NOP, although similar in some degree to NLMS in the sense that it is independent of the unknown, i.e. $\mathbf{W}$, does not mitigate the gradient noise problem (see (19)) and destroys the desirable property of having the unity upper bound on the scaling factor.

In conclusion, although derived from a different perspective, we can also view NOP as an algorithm that employs the same natural gradient descent direction as OP but with adaptive step-size. From this viewpoint, it can be said that NOP not only utilizes the nonstationarity and the nonwhiteness properties but also exhibits fast convergence and robustness by overcoming the gradient noise problem. It is shown by numerical experiments that an improved performance, when compared to its original version, is achieved by the proposed normalized algorithm.

[1] Available: http://www.bsp.brain.riken.jp/ICALAB/ICALABSignalProc/benchmarks/Speech4.mat

6. REFERENCES

[1] U. Manmontri and P. A. Naylor, "Blind identification using second-order statistics: a nonstationarity and nonwhiteness approach," in *Proc. IEEE ICASSP'05*, vol. 5, PA, USA, Mar. 2005, pp. 305–308.

[2] S. Haykin, *Adaptive Filter Theory*. Upper Saddle River, NJ: Prentice-Hall, 2002, ch. 6.

[3] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, pp. 251–276, Feb. 1998.

[4] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. Signal Processing*, vol. 49, pp. 1837–1848, Sept. 2001.

[5] E. Moreau, "A generalization of joint-diagonalization criteria for source separation," *IEEE Trans. Signal Processing*, vol. 49, pp. 530–541, Mar. 2001.

[6] L. Parra and C. Spence, "Convolutive blind separation of nonstationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 320–327, May 2000.

[7] W. Wang, S. Sanei, and J. A. Chambers, "Penalty function-based joint diagonalization approach for convolutive blind separation of nonstationary sources," *IEEE Trans. Signal Processing*, vol. 53, pp. 1654–1669, May 2005.

[8] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 44, pp. 3017–3030, Dec. 1996.

[9] M. Joho and H. Mathis, "Joint diagonalization of correlation matrices by using gradient methods with application to blind signal separation," in *Proc. IEEE SAM'02*, VA, USA, Aug. 2002, pp. 273–277.