

Selective-Tap Blind Signal Processing for Speech Separation

Kostas Kokkinakis, *Member, IEEE* and Philipos C. Loizou, *Senior Member, IEEE*

Abstract—In this paper, we propose a new blind multi-channel adaptive filtering scheme, which incorporates a partial-updating mechanism in the error gradient of the update equation. The proposed blind processing algorithm operates in the time-domain by updating only a selected portion of the adaptive filters. The algorithm steers all computational resources to filter taps having the largest magnitude gradient components on the error surface. Therefore, it requires only a small number of updates at each iteration and can substantially minimize overall computational complexity. Numerical experiments carried out in realistic blind identification scenarios indicate that the performance of the proposed algorithm is comparable to the performance of its full-update counterpart, but with the added benefit of a highly reduced computational complexity.

I. INTRODUCTION

In recent years, *blind source separation* (BSS) has received much attention due to its strong potential for use in a variety of applications, such as automatic speech recognition, hearing aid devices and hands-free telephony. In scenarios where the signals of interest are mixed with other ongoing background activity and noise, BSS can be used to extract and perceptually enhance the waveform of the desired sound source(s) from a set of composite signals. By definition, BSS recovers estimates of n sources with very little to almost *no prior* knowledge about the source-to-sensor geometry or the source signal themselves. Instead, it relies only on information collected from a set of m convolutive data $\mathbf{x}(t) = [x_1(t), \dots, x_m(t)]^T \in \mathbb{R}^m$

$$\mathbf{x}(t) = \sum_{\ell=0}^{\infty} \mathbf{H}_{\ell}(t) \mathbf{s}(t - \ell), \quad t = 1, 2, \dots \quad (1)$$

where $\mathbf{H}_{\ell}(t)$ represents the unknown but linear-time invariant (LTI) multiple-input multiple-output (MIMO) mixing system at discrete time t and lag ℓ . The ‘blind’ recovery of the original sources then boils down to a multichannel inverse filtering task whereby the coefficients of an L -dimensional finite impulse response (FIR) equalizer $\mathbf{W}_{\ell}(t)$ are adjusted such that the output vector $\mathbf{u}(t) = [u_1(t), \dots, u_n(t)]^T \in \mathbb{R}^n$

$$\mathbf{u}(t) = \sum_{\ell=0}^{L-1} \mathbf{W}_{\ell}(t) \mathbf{x}(t - \ell), \quad t = 1, 2, \dots \quad (2)$$

defined for all $0 \leq \ell \leq L - 1$. Nevertheless, since the

This work was supported by Grants R03 DC 008882 and R01 DC 007527 awarded from the National Institute on Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health (NIH).

The authors are with the Center for Robust Speech Systems, Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75080, USA (e-mail: kokkinak@utdallas.edu, loizou@utdallas.edu).

computational complexity of any adaptive filtering algorithm is proportional to its tap length, time-domain multichannel BSS algorithms can become computationally prohibitive especially for applications that require a large number of filter taps (e.g., see [1], [2]).

To reduce excessive computational requirements, many authors have recently suggested to carry out the separation task in the frequency or the sub-band domain (e.g., see [3]–[4], [5]). By doing so, the overall BSS task can be elegantly reduced into several independent instantaneous problems, one for each frequency subband. Still such strategies come with perils of their own, namely *scaling* and *permutation* ambiguities, which usually have a negative effect on separation performance [6]. Another strategy for reducing the length of the inverse FIR filters is to simply increase the number of microphones present in the system. In setups where the system sensors available outnumber the active sources, the inverse FIR filters can be chosen to be significantly shorter¹ than the length of the acoustic impulse responses [7]. Nonetheless, the design requirements in modern digital signal processors (DSPs), for example such as those used in hearing aids and cochlear implant devices [8], often dictate a low power consumption in battery operated equipment and also call for a portable weight in a compact size. Naturally, such design constraints prevent the use of BSS and other similar pre-processing techniques on portable devices since they (1) prohibit large amounts of processing power and (2) limit the number of microphones that can be made available to the user.

A far more efficient alternative to lower the overall computational complexity of BSS is to resort to *selective-tap* (or *partial*) updating schemes. Such schemes operate by updating only a *subset* of the total number of filter coefficients on every iteration and hence can lead to a substantially reduced complexity. Early examples of selective-tap techniques include the sequential and periodic least-mean-squares (LMS) [9] and more recently the MMax normalized least-mean-squares (MMax-NLMS) [10], [11], whereby the updated taps are the ones associated with the largest magnitude gradient components on the error surface. In this paper, we extend the MMax tap-selection criterion and derive a *novel* selective-updating blind scheme based on the natural gradient algorithm (S-NGA). The potential of the proposed low-complexity algorithm is verified and assessed through numerical simulations in realistic acoustical scenarios.

¹For example in a two-source and three-sensor configuration convolved with 1,024 sample point acoustic impulse responses, the lower bound for the FIR filter length L yielding a perfect system inverse can decrease to around 512 taps, if we assume that the number of microphones doubles.

II. SELECTIVE-TAP BLIND SOURCE SEPARATION

A. Natural Gradient Algorithm

Spatial independence is a key assumption for BSS. In short, it implies that the output joint probability density function (PDF) $p_u(\mathbf{u}(t))$ is equal to the product of the output PDFs

$$p_u(\mathbf{u}(t)) = \prod_{i=1}^n p_{u_i}(u_i(t)). \quad (3)$$

A simple way to check for independence is to measure the distance between the two sides of (3) with an appropriate distance measure such as the Kullback-Leibler (K-L) divergence. Alternatively, we can resort to the *maximum likelihood* (ML) principle and choose to optimize the negative log-likelihood (objective) function with respect to the unmixing matrix \mathbf{W}_ℓ such that

$$J_i(u_i) = - \sum_{i=1}^n E[\log p_{u_i}(u_i)] - \log |\det(\mathbf{W}_\ell)| \quad (4)$$

Due to the Riemannian nature of the optimization parameter space, a less computationally burdensome choice to minimize the cost function in (4) with respect to the unmixing matrix \mathbf{W}_ℓ is the so-called *natural gradient* [12], which is an optimal re-scaling of the standard (stochastic) entropy gradient (e.g., see [13]). For multipath conditions where $L > 1$, the *natural gradient algorithm* (NGA) is given by [1]–[2], [12]

$$\mathbf{W}_\ell(t+1) = \mathbf{W}_\ell(t) + \Delta \mathbf{W}_\ell(t) \quad (5)$$

$$\Delta \mathbf{W}_\ell(t) = \mu [\mathbf{W}_\ell(t) - \varphi(\mathbf{u}(t)) \tilde{\mathbf{u}}^H(t-\ell)] \quad (6)$$

$$\tilde{\mathbf{u}}(t) = \sum_{k=0}^{L-1} \mathbf{W}_{L-1-k}^H(t) \mathbf{u}(t-k) \quad (7)$$

where $0 < \mu < 1$ is a positive learning parameter controlling the rate of convergence and rate of adaptation, $(\cdot)^H$ is the Hermitian operator and symbol $\varphi(\cdot)$ represents the nonlinear monotonic activation (or score) function operating elementwise on the output signal vector, such that²

$$\varphi(\mathbf{u}(t)) \triangleq [\varphi_1(u_1(t)), \dots, \varphi_1(u_1(t-L+1)), \dots, \varphi_n(u_n(t)), \dots, \varphi_n(u_n(t-L+1))]^T \quad (8)$$

²Instead of using the definition in (8), we could re-write the update in (6) as $\Delta \mathbf{W}_\ell(t) = \mu [\mathbf{W}_\ell(t) - \varphi(\mathbf{u}(t-L+1)) \tilde{\mathbf{u}}^H(t-\ell)]$ whereby we are introducing the $(L-1)$ -sample delay in term $\mathbf{u}(t)$ to accommodate for non-causal parts of the equalizer filters [12].

where

$$\varphi_i(u_i) = - \frac{\partial \log p_{u_i}(u_i)}{\partial u_i}, \quad i = 1, 2, \dots, n. \quad (9)$$

with $p_{u_i}(u_i)$ denoting the PDF of each source estimate u_i . Note also that vector $\tilde{\mathbf{u}}(t)$ represents the reverse-filtered output computed by using the latest $(L-1)$ -samples backwards from the current sample t for all lags $\ell = 0, 1, \dots, L-1$ as shown in (7).

B. Selective-Natural Gradient Algorithm

When approaching convergence, $\Delta \mathbf{W}_\ell(t) \simeq 0$, assuming a sufficiently small step-size. The stationary points of (6) can guarantee both temporal and spatial statistical independence under the following two conditions

$$E[\varphi_i(u_i(t)) u_j^*(t-\ell)] = \begin{cases} \delta_\ell, & \forall \ell \neq 0 \text{ and } i = j \\ 0, & \forall t, \ell \text{ and } i \neq j \end{cases} \quad (10)$$

where $E[\cdot]$ is the statistical expectation, $(\cdot)^*$ represents the complex-conjugate operator and δ_ℓ is the Kronecker delta, which is equal to 1 for $\ell = 0$ and equal to 0 otherwise. As it can be seen from (10), the convergence behavior of the NGA depends solely upon the magnitude of the so-called *estimating function* at each iteration, which is equal to

$$\mathbf{R}_\ell(t) = E[\varphi(\mathbf{u}(t)) \mathbf{u}^H(t-\ell)] \quad (11)$$

for lags $\ell = 0, 1, \dots, L-1$. Since the above estimating function is not equally sensitive to variations from all the updated filter coefficients, a tap-selection criterion can be constructed by employing only M out of L coefficients with the largest values of $|\mathbf{R}_\ell(t)|$ for all lags $\ell = 0, 1, \dots, L-1$, at each iteration. The subset of the filter coefficients to be partially updated at any particular time t is specified in the $(nL \times mL)$ matrix $\mathbf{Q}(t)$, which is coined the tap selection matrix, and is given by

$$\mathbf{Q}(t) \triangleq \begin{pmatrix} \text{diag}[\mathbf{q}_{11}(t)] & \dots & \text{diag}[\mathbf{q}_{1m}(t)] \\ \vdots & \ddots & \vdots \\ \text{diag}[\mathbf{q}_{n1}(t)] & \dots & \text{diag}[\mathbf{q}_{nm}(t)] \end{pmatrix} \quad (12)$$

where each element of the tap-selection matrix is given by

$$\mathbf{q}_{ij}(t) \triangleq [q_{ij}(t), q_{ij}(t-1), \dots, q_{ij}(t-L+1)]^T \quad (13)$$

such that, after dropping the time index t for convenience,

$$q_{ij}(\tau) = \begin{cases} 1, & \text{if } |r_\ell(\tau)| \in [M \text{ maxima } |\mathbf{R}_\ell(t)|] \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

where $|\cdot|$ denotes absolute value and $r_\ell(\tau)$ are the elements of (11) at the (sorted) indices $\tau = 0, 1, \dots, L-1$. Every element $q_{ij}(\tau)$ is either equal to one or zero, depending on

whether the condition in (14) is satisfied or not. In order to calculate the filter coefficients that are to be updated at different time instants, a fast sorting algorithm needs to be executed at every iteration [14]. After the sorting is performed, each block of the tap selection matrix $\mathbf{Q}(t)$ contains $M < L$ coefficients equal to one in the positions (or indices) calculated from (14) and zeros elsewhere, such that

$$M = \text{trace}[\mathbf{Q}(t)] \quad (15)$$

where operator $\text{trace}[\cdot]$ denotes the sum of the diagonal elements of the input matrix argument. Accordingly, to update only M taps in the equalizer $\mathbf{W}_\ell(t)$, we can write the *selective-natural gradient algorithm* (S-NGA), as follows

$$\widetilde{\mathbf{W}}_\ell(t+1) = \widetilde{\mathbf{W}}_\ell(t) + \Delta\widetilde{\mathbf{W}}_\ell(t) \quad (16)$$

$$\Delta\widetilde{\mathbf{W}}_\ell(t) = \lambda \left[\widetilde{\mathbf{W}}_\ell(t) - \varphi(\tilde{\mathbf{u}}(t))\tilde{\mathbf{y}}^H(t-\ell) \right] \quad (17)$$

$$\tilde{\mathbf{y}}(t) = \sum_{k=0}^{M-1} \widetilde{\mathbf{W}}_{M-1-k}^H(t) \tilde{\mathbf{u}}(t-k) \quad (18)$$

$$\tilde{\mathbf{u}}(t) = \sum_{k=0}^{M-1} \mathbf{Q}(t) \mathbf{u}(t-k) \quad (19)$$

where parameter λ represents the new learning rate. Note that for $M = L$, the S-NGA algorithm in (16)–(19) reduces to the *full-update* NGA algorithm given in (5)–(7). In general, the separation performance of the S-NGA depends on the degree of coefficient reduction achieved and therefore on the overall number of the sorted filter taps in $\widetilde{\mathbf{W}}_\ell(t)$.

III. EXPERIMENTAL METHODOLOGY

A. Material

To assess the performance of the S-NGA, five male speakers are corrupted by interfering speech. The signals are chosen from the IEEE speech corpus, which consists of phonetically balanced sentences, with each sentence being composed of approximately 7 to 12 words (e.g., see [15]). Every sentence produced by a male talker is designated as the *target* speech. To simulate the speech interferer or competing voice in this experiment, a female talker uttering the sentence “Tea served from the brown jag is tasty” is chosen as the *interferer* (or *masker*) source. The source signals are approximately 4 s in duration and are recorded at a sampling rate of 8 kHz. A set of five convolutive speech mixtures are produced by convolving the clean signals with the two *binaural* room impulse responses (BRIRs) depicted in Fig. 1 (a) and (b) (e.g., see [16]). The length of the acoustic impulse responses is 2,048 sample points, corresponding to a delay of around 256 ms at a sampling rate of 8 kHz.

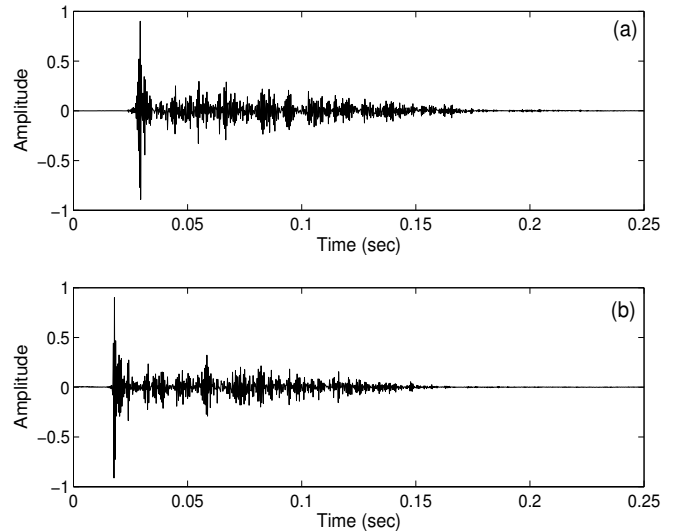


Fig. 1. Impulse responses of the cross-channel acoustic paths measured in a small classroom. (a) Impulse response corresponding to the acoustic path from the target source to microphone 2. (b) Impulse response corresponding to the acoustic path from the female masker to microphone 1. The sampling rate is 8 kHz.

In order to simplify the estimation problem, the channel distortions from the target source to microphone 1 and from the female masker to microphone 2 are assumed negligible. The BRIRs are measured in a $5 \times 9 \times 3.5$ m ordinary classroom using a KEMAR positioned at 1.5 m above the floor and at ear level [16]. The broadband reverberation time³ of this enclosure is equal to $T_R = 200$ ms, which is a typical value for a moderately reverberant environment. Both speech signals have the same onset and are normalized so that their maximum amplitude is unity. By convolving the speech signals with these pre-measured impulse responses, the target male is positioned directly at the front of the listener at a 0° azimuth, whereas the female interferer is placed at an angle of 60° to the right.

B. Performance Evaluation

The S-NGA is executed with $L = 2,048$, whereas M is set to 2,048, 1,024, 512 and 256 taps. The learning rates are explicitly tuned to yield the maximum possible steady-state performance. In all cases, the equalizer is initialized using a center-tap scheme, such that $\mathbf{W}_\ell(0) = \delta_{\ell-M/2} \mathbf{I}$. The algorithm operates with the hyperbolic tangent score function $\varphi_i(u_i) = \tanh(u_i)$ and converges after approximately 20 passes through the convolutive speech data. To assess separation performance we resort to the signal-to-interference-ratio improvement (SIRI). SIRI is defined as the *overall* amount of crosstalk reduction achieved by the algorithm *before* (SIR_i) and *after* (SIR_o) the unmixing stage and is described in [2].

³The parameter T_R defines the interval in which the reverberating sound energy, due to decaying reflections, reaches one millionth of its initial value. In other words, it is the time it takes for the reverberation level to drop by 60 dB below the original sound energy present in the room at a given instant.

IV. RESULTS AND DISCUSSION

We test the S-NGA on a set of convolutive speech mixtures generated using the pre-measured impulse responses in Fig. 1. Table I contrasts performance and reduction in complexity relative to the complexity of the full-update algorithm. As expected, the full-update NGA yields the best SIRI performance. In addition, even when the S-NGA operates with a reduced number of filter coefficients at every iteration, it exhibits only a slightly lower separation performance. SIRI values indicate that the overall degree of separation remains unchanged for $M = 1,024$. In fact, even when setting $M = 512$, which accounts for a 75% reduction in the total equalizer length (with a processing delay of just 64 ms at 8 kHz) and around 40% reduction in computational complexity, the algorithm manages to retain almost 80% of its full-update counterpart performance. The overall computational complexity of the S-NGA in terms of floating point operations per second (FLOPS) is around 50% less when only 3/16 of filter coefficients (amounting to a 81.25% reduction) are adapted at every iteration. In all cases, to ensure that the complexity due to coefficient selection is kept low, we use a fast sorting routine (e.g., see SORTLINE [14]), which only requires an additional $\log_2 L + 2$ tap comparison operations per sample.

V. CONCLUSIONS

We have developed a selective-updating BSS scheme (S-NGA), which can learn multiple filters at a substantially reduced computational overhead, while retaining a satisfactory separation performance. Experiments in reverberant settings, show that the overall tradeoff between filter length reduction and performance loss is acceptable. The S-NGA has great potential for use in portable devices, e.g., hearing aids and cochlear implants, since it can operate with considerably shorter filters and still equalize long acoustic echo paths with sufficient accuracy.

VI. REFERENCES

- [1] S. C. Douglas, H. Sawada and S. Makino "Natural gradient multi-channel blind deconvolution and speech separation using causal FIR filters," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 92–104, 2005.
- [2] K. Kokkinakis and A. K. Nandi, "Multichannel blind deconvolution for source separation in convolutive mixtures of speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 200–212, 2006.
- [3] L. Parra and C. Spence, "Convolutive blind separation of nonstationary sources," *IEEE Trans. Speech and Audio Process.*, vol. 8, pp. 320–327, 2000.
- [4] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomp.*, vol. 22, pp. 21–34, 1998.
- [5] K. Kokkinakis and P. C. Loizou, "Subband-based blind signal processing for source separation in convolutive mixtures of speech," In *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process.*, pp. 917–920, 2007.
- [6] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: Investigation and solutions," *IEEE Trans. Speech and Audio Process.*, vol. 13, pp. 1–13, 2005.

TABLE I

COMPLEXITY REDUCTION VS. SEPARATION PERFORMANCE FOR FIVE MALE TARGET SPEAKERS WHEN PROCESSED WITH THE S-NGA

$L M$	Complexity Reduction	Speaker	SIRI (dB)
2,048 2,048	–	A	17.97
		B	17.83
		C	19.62
		D	18.01
		E	18.24
2,048 1,024	25.00%	A	17.24
		B	17.01
		C	18.79
		D	17.57
		E	17.92
2,048 512	39.50%	A	16.18
		B	15.67
		C	17.52
		D	16.31
		E	16.98
2,048 256	65.75%	A	9.07
		B	8.72
		C	10.95
		D	9.16
		E	9.94

- [7] M. Hofbauer, "On the FIR inversion of an acoustical convolutive mixing system: Properties and limitations" In *Proc. Fifth Int. Conf. on ICA and BSS*, pp. 643–651, 2004.
- [8] K. Kokkinakis and P. C. Loizou, "Using blind source separation techniques to improve speech recognition in bilateral cochlear implant patients," *J. Acoust. Soc. Am.*, vol. 123, pp. 2379–2390, 2008.
- [9] S. C. Douglas, "Adaptive filters employing partial updates," *IEEE Trans. Circuits Syst. II*, vol. 44, pp. 209–216, 1997.
- [10] T. Aboulnasr and K. Mayyas, "Complexity reduction of the NLMS algorithm via selective coefficient update," *IEEE Trans. Signal Process.*, vol. 47, pp. 1421–1424, 1999.
- [11] A. W. H. Khong and P. A. Naylor, "Selective-tap adaptive filtering with performance analysis for identification of time-varying systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 1681–1695, 2007.
- [12] S.-I. Amari, S. C. Douglas, A. Cichocki and H. H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," In *Proc. IEEE Workshop on Signal Process. Adv. in Wireless Comms.*, pp. 101–104, 1997.
- [13] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computat.*, vol. 7, pp. 1129–1159, 1995.
- [14] I. Pitas, "Fast algorithms for running ordering and max/min calculation," *IEEE Trans. Circuits Syst.*, vol. 36, pp. 795–804, 1989.
- [15] IEEE Subcommittee, "IEEE recommended practice speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.
- [16] B. G. Shinn-Cunningham, N. Kopco and T. J. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *J. Acoust. Soc. Am.*, vol. 117, pp. 3100–3115, 2005.