DISCRETE AND CONTINUOUS
DYNAMICAL SYSTEMS
Volume **28**, Number **4**, December **2010**

doi:10.3934/dcds.2010.28.1753

pp. **1753–1767**

STOCHASTIC APPROXIMATION AND A NONLOCALLY WEIGHTED SOFT-CONSTRAINED RECURSIVE ALGORITHM FOR BLIND SEPARATION OF REVERBERANT SPEECH MIXTURES

Meng Yu and Jack Xin

Department of Mathematics, University of California
Irvine, CA 92697, USA

Abstract. We review statistical equations for blind source separation problems, then introduce their stochastic approximation and recursive algorithms. The recurrence resembles discretization of nonlinear systems of ordinary differential equations which may not have global solutions in general. Though scaling variables were used before to control finite time blowup, instabilities may arise from small divisor problem during silent periods of speech signals, and asymptotic balance as a necessary condition for convergence was ignored. To resolve these deficiencies, we propose a nonlocally weighted soft-constrained recursive algorithm. The nonlocal weighting of the iterations promotes stability and convergence of the algorithm. The scaling variables evolve by soft-constrained difference equations. Computations on synthetic speech mixtures based on measured binaural room impulse responses in enclosed rooms with reverberation time up to 1 second show that the new algorithm achieves consistently higher signal-to-interference ratio improvement than existing methods. The algorithm is observed to be stable and convergent, and is applied to separation of room recorded mixtures of song and music as well.

1. **Introduction.** Blind source separation (BSS) is a major area of research in signal and image processing [4, 6, 8, 14, 20] for recovering source signals from their mixtures without detailed knowledge of the mixing process (blindness). If source signals are time series and modeled by stochastic processes as is common for sound signals, one may use the independence property of source signals to estimate the mixing matrices. Such a statistical approach is called independent component analysis and has been studied extensively in the past two decades [7].

However, it remains a challenge to retrieve sound sources recorded in real-world environment such as in furnished rooms. The physical reason is that sound reflections or reverberations in enclosed rooms cause mixture signals at current time to depend on source signals and their long delays (history dependent). Mathematically, the mixing process is convolutive in time and the unknowns are high dimensional. In the case of $n \geq 2$ receivers and n sources, the discrete mixing model is:

$$x(t) = [A * s](t) = \sum_{i=0}^{q} A(i)\, s(t-i) \tag{1.1}$$

2000 *Mathematics Subject Classification.* Primary: 94A12, 65H10, 65C60; Secondary: 60G35.
Key words and phrases. Stochastic approximation, weighted constrained recurrence, reverberant speech mixtures, blind separation.
This work was partially supported by NSF grant DMS-0712881.

where t is the discrete time index, s is the digital source signal vector of duration m, $s \in R^{n\times m}$ ($m \gg n$); each $A(i)$ is $n\times n$ mixing matrix at time i; q is the index of total time delay. A typical example is $n = 2$, or two ears and two sound sources. The mixing matrix $A(i)$ is a *highly oscillatory* function in i due to multiple reflections of sound waves in the environment, and q is on the order of thousands for a typical office size room. The mathematical problem is to recover s from x without knowing $A(i)$'s.

As observed in [18], the "inversion formula" to (1.1) exists if $A(i)$ were known, and it is in the form:

$$y(l) = \sum_{p=0}^{L} W(p)\, x(l-p), \tag{1.2}$$

for some integer $L \geq q$, and $n \times n$ matrices $W(p)$ (so called demixing matrices). In case $n = 2$, $L = q$, and $W(p)$ equals the adjoint of $A(p)$ if y is recovered up to a convolution factor. Because y has independent components, one can write down infinitely many moment equations to find W. Computationally, it is the second and fourth moment equations that are often used for approximation under reasonable algebraic complexity. The second order moment (cross correlation) condition and the non-stationarity of source signals imply the statistical equations

$$E[y_j(t)\, y_k(t+\eta)] = 0,\ j \neq k,\ \ \forall \eta, \tag{1.3}$$

which are solved in [18] in an optimization framework. A more general cross correlation condition with information of speech source probability distribution is

$$E[\mathrm{sgn}(\mathrm{y_j(t)})\, \mathrm{y_k(t+\eta)}] = 0,\ \mathrm{j \neq k},\ \ \forall \eta, \tag{1.4}$$

see [19] for a derivation using maximum likelihood estimation principle. The sign function comes from logarithmic derivative of the density function of speech Laplace distribution $\exp\{-|s|^\gamma/\sigma\}$, $\gamma = 1$. If speech signals were Gaussian ($\gamma = 2$), the logarithmic derivative would be linear and standard cross correlation (1.3) resulted. Hence the generalized cross correlation condition (1.4) encodes both non-Gaussianity and non-stationarity of the source signals. As the sign function is non-smooth, statistical equations (1.4) cannot be solved by inserting (1.2) and expanding, which is the case for (1.3). Instead, a dynamic *recursive method or stochastic approximation* is expedient for finding solutions. A recursive scheme updates W in time frame by frame (a frame is an interval of time index i):

$$W_{k+1}(p) = W_k(p) - \mu\, F_k(p), \tag{1.5}$$

where k is frame index, μ is time step (learning parameter), and F_k is a nonlinear function to be designed. If computation can be done fast enough, (1.2) gives output y in real time. The separation quality normally gets better as k becomes large and more data contribute. In other words, the algorithm approaches a solution of general cross correlation equation (1.4) in time. Dynamic recurrence (1.5) resembles a discrete system of nonlinear ordinary differential equations. It is expected that F_k's have to satisfy certain properties so that solutions do not blow up.

The general theory of stochastic approximation and recursive algorithms is well-studied [15]. The scaling degree of freedom in BSS problems raise new dynamic scaling issues however. In other words, for any diagonal matrix D with nonzero elements, output y and $D\,y$ are both acceptable as solutions. This can be seen from (1.3) or (1.4) in that these equations are invariant under diagonal scaling of W. In terms of actual hearing, scaling changes only the volume of the output and does not

change the frequency contents of the signals in the output. So finding W is similar to seeking an eigenvector in the null space.

The paper is organized as follows. In section 2, we discuss how to use scaling degree of freedom to control the recurrence (1.5) to maintain boundedness of iterates in large k. Moreover, we also require the recurrence to have an asymptotic balance to ameliorate convergence of solutions in large k. To this end, we introduce nonlocal weighting and soft constraints to (1.5). In section 3, we demonstrate numerically the convergent behavior of the proposed algorithm and separation of sound mixtures in simulated room environment. The separation quality is measured in terms of an objective quantity called signal to interference ratio improvement (SIRI). Comparisons with other methods in the literature shows better performance of our algorithm. We also show results on separating recorded song and music in a furnished office. In section 4, we use a simple dereverberation method as a preprocessing step to further improve SIRI. Concluding remarks are made in section 5.

2. **Nonlocally-weighted soft-constrained recurrence.** A natural role of $F_k(p)$ is a gradient, chosen to be proportional to the product of W_k and the generalized cross correlation function. Such a gradient is called natural gradient. More generally, $F_k(p) = \lambda_k W_k(p) + H_k(p)$, for some number λ_k, and:

$$H_k(p) \stackrel{\triangle}{=} \frac{1}{N} \sum_{l=(k-1)N+1}^{kN} \sum_{q=0}^{L} \operatorname{sgn}(y_k(l)) y_k^T(l-p+q) W_k(q), \tag{2.1}$$

where the outside sum approximates expectation, the inside sum over q is convolution of $\operatorname{sgn}(y_k(\cdot))\, y_k^T(\cdot - \eta)$ with W_k. So H_k is simply the convolutive product of generlized correlation function with W_k. The multiplication of W_k in (2.1) renders the choice of time step μ independent of the size of W_k. This way, (1.5) becomes a naturally weighted gradient descent. The convolutive gradient (2.1) was first derived in [2], and the resulting recurrence is ($\lambda_k = -1$):

$$W_{k+1}(p) = (1+\mu)\, W_k(p) - \mu\, H_k(p). \tag{2.2}$$

It was realized later [10] that recursive scheme (2.2) may diverge (blow-up) in k, and extra scaling variables are added to keep W_k bounded. A scaled recurrence is [10]:

$$W_{k+1}(p) = (1+\mu)\, d_k^{-1}\, W_k(p) - \mu\, d_k^{-2}\, H_k(p). \tag{2.3}$$

where the scaling factor d_k is homogeneous of degree one in the mixture signal x in the k-th frame. Specifically,

$$d_k = \frac{1}{n} \sum_{i,j=1}^{n} \sum_{q=0}^{2L} |g_k^{ij}(q)|, \tag{2.4}$$

where $(g_k^{ij}(q))$ are entries of an $n \times n$ matrix G for each (k,q). The matrix G is:

$$G_k(q) \stackrel{\triangle}{=} \sum_{p=0}^{L} F_k(q-p)\, W_k^{\top}(L-p), \quad q \in [0, 2L] \tag{2.5}$$

where $\top$ denotes transpose, and the matrix $F_k(p)$ is:

$$F_k(p) \stackrel{\triangle}{=} \frac{1}{N} \sum_{l=(k-1)N+1}^{kN} \operatorname{sgn}(y_k(l))\, x^{\top}(l-p). \tag{2.6}$$

In the sum of (2.5), $F_k(p)$ is set to zero if $p \notin [0, L]$.

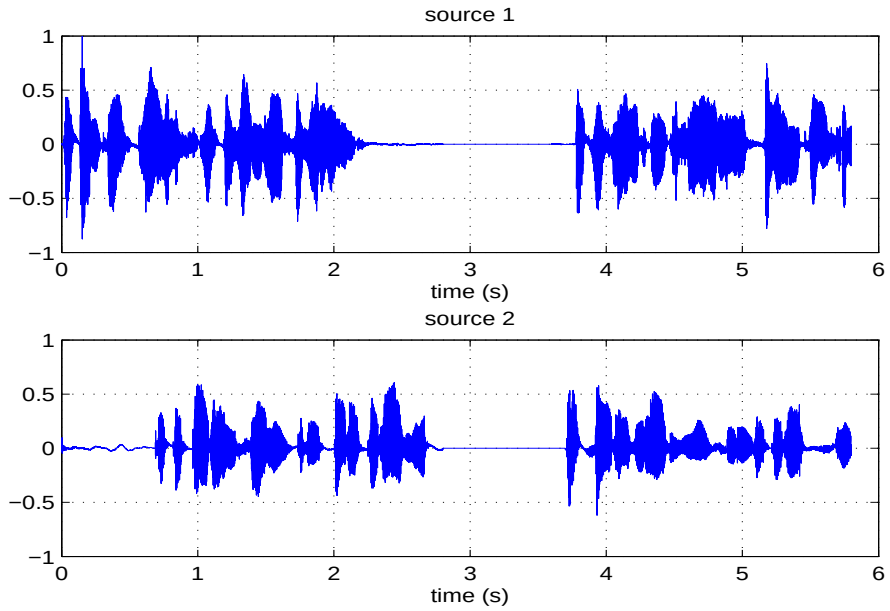

FIGURE 1. Two speech source signals with silence periods in the middle.

There are two deficiencies in (2.3). First, small divisor problem appears ($d_k \approx 0$) if the k-th frame is near or in a silence period of speech and leads to instability [19]. A pair of speech source signals with silence periods and their convolutive mixtures in anechoic condition are plotted in Figs. 1-2. The silence period lasts approximately from 2.8 second to 3.6 second in the mixtures, with a total signal duration of 5.8 second. Fig. 3 plots the evolution of scaling parameter d_k as a function of k (frame number). Once the marching frame touches the silence part, the d_k's quickly go to zero, and cause the algorithm to overflow and crash. The source estimation dies out thenceforth. In numerical experiments, we observed that as long as the duration of silence part is longer than 0.2 second, the algorithm (2.3) (time domain natural gradient algorithm) already fails.

Secondly, there is "inconsistency problem". Suppose that y_k converges up to a limiting signal s with independent components at suitably large values of N, and that $W_k(q)$ converges to $W_\infty(q)$, then

$$H_\infty(p) \stackrel{\triangle}{=} \lim_{k\to\infty, N\to\infty} H_k(p) = \sum_{q=0}^{L} E[\mathrm{sgn}(s(\cdot))\, s^T(\cdot - (p-q))] W_\infty(q) \tag{2.7}$$

is a *convolutive (non-local) product* that *cannot be balanced by the local terms* (constant multiple of W_k) in (2.3)! The generalized correlation matrix $E[\mathrm{sgn}(s)(\cdot)\, s^T(\cdot - \eta)]$ is diagonal with diagonal entries being functions of η with finite support. Only in the special case that s is independent from time to time (white in time), the matrix $E[\mathrm{sgn}(s)(\cdot)\, s^T(\cdot - \eta)]$ is zero if $\eta \neq 0$, and the sum in (2.7) reduces to a local multiplication. In other words, convergence cannot happen in (2.3) for mixtures of colored signals such as speech.

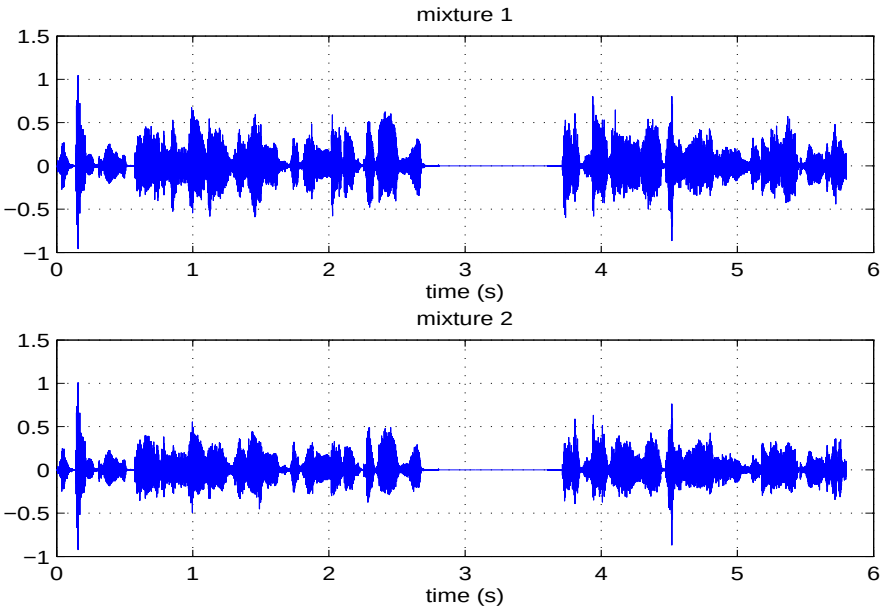

FIGURE 2. Two convolutive speech mixtures with silence periods in an anechoic room.

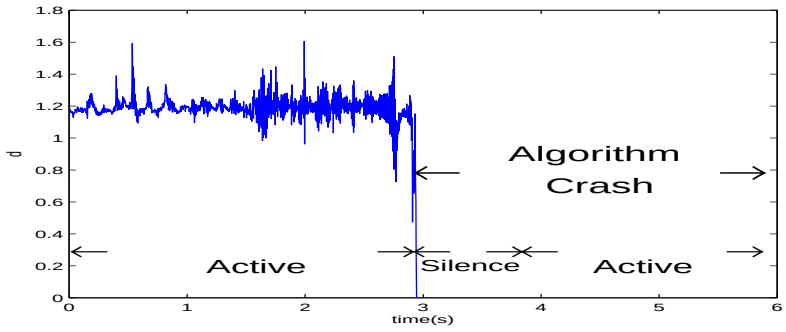

FIGURE 3. Illustration of small divisor problem: scaling variable d_k as a function of iteration (frame) number k shows oscillatory behavior before rapid fall-off to zero at the beginning of silence period in mixture signals in Fig. 2. The scaled natural gradient algorithm (2.3) crashes and is unable to function thereafter.

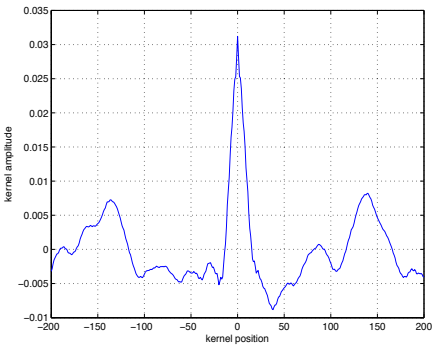

FIGURE 4. Measured (sliding window) generalized correlation function from a clean speech signal

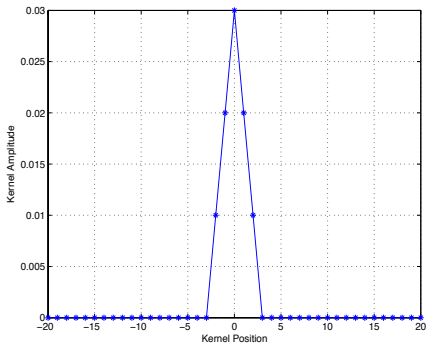

FIGURE 5. Piecewise linear kernel function in the computation, with base width $L_M = 5$ and peak height 0.03.

We introduce a new recursive method to resolve both the inconsistency and the small divisor problem for convolutive source separation by correcting (2.3) with two important ingredients: *nonlocal weighting and soft constraints*. The nonlocally weighted demixing matrix iteration is:

$$W_{k+1}(p) = W_k(p) + \sigma_{1,k}\,[M * W_k](p) - \sigma_{2,k}\,H_k(p) \tag{2.8}$$

where H is by (2.1), $*$ is linear convolution, M is a nonlinear kernel function. The nonlocal weighting by M of W_k or $M*W$ approximates the limiting functional form in (2.7) so that with proper scaling coefficicents $(\sigma_{1,k}, \sigma_{2,k})$ the last two terms of (2.8) balance out. We shall denote by L_M is the length of the support of the kernel function. The kernel function $M = M(i)$ used in our computation is plotted in Fig. 5. It is constructed from numerical data of (sliding window) generalized correlation functions of clean speech signals, see Fig. 4 for a typical example. We simplify the "measured kernels" into a piecewise linear function (hat function) with $L_M = 5$ and amplitude equal to 0.03, shown in Fig. 5.

The measured kernel as in Fig. 4 suggests the peak height value of 0.03. The width of the kernel function $L_M = 5$ is determined from separation quality of anechoic convolutive mixtures of speech samples which are synthesized by measured anechoic binaural room impulse responses (BRIRs)[23]. For simplicity, we have chosen the same kernel function M for all entries of $W_k(p)$ matrix. The locally scaled version (2.3) is a special case when the support of piecewise linear M shrinks to a single point.

The scaling variables $(\sigma_{1,k}, \sigma_{2,k})$ are updated by the soft-constrained equations:

$$\begin{aligned} \sigma_{1,k+1} &= \sigma_{1,k} \exp\{-\nu F_{1,k}(\sigma_{1,k}, \sigma_{2,k})\} \\ \sigma_{2,k+1} &= \sigma_{2,k} \exp\{-\nu F_{2,k}(\sigma_{1,k}, \sigma_{2,k})\} \end{aligned} \tag{2.9}$$

where ν is a positive constant, and:

$$F_{1,k} \triangleq \sigma_{1,k} \sum_{p=0}^{L} \sum_{j=1}^{n} |W_k^{1,j}(p)| + \sigma_{2,k} \sum_{p=0}^{L} \sum_{j=1}^{n} |H_k^{1,j}(p)| - c_1$$
$$F_{2,k} \triangleq \sigma_{1,k} \sum_{p=0}^{L} \sum_{j=1}^{n} |W_k^{2,j}(p)| + \sigma_{2,k} \sum_{p=0}^{L} \sum_{j=1}^{n} |H_k^{2,j}(p)| - c_2 + \sigma_{2,k} E \tag{2.10}$$

where c_1, c_2 and E are positive constants.

Equations (2.9)-(2.10) say that if $|W_k|$ were too large, F_1 and F_2 would be positive and reduce (σ_1, σ_2) so that $|W_k|$ would not continue to grow in k and become unbounded. Similarly, if $|W_k|$ were too small, the nonlinear term H would be smaller, and so F_1 and F_2 would turn negative so that (σ_1, σ_2) would grow in k and not become trivial. The growth of σ_1 in turn helps the growth of $|W|$ which dominates that of H in equation (2.8) when $|W|$ is small enough. It follows that the dynamics by (2.8)-(2.10) admit an invariant region where the iterates are bounded and nontrivial, which implies weak convergence of the iteration and the separation condition that the off-diagonal elements of $< f(y_k(t) y_k^T(t-\eta) >$ tend to zero as $k \gg 1$ for a range of η, here $< \cdot >$ is a temporal average in t, [19].

3. **Numerical results and comparisons.** The proposed nonlocally weighted soft-constrained natural gradient (NLW-SCNG) recursive algorithm is tested on convolutive and reverberant speech mixtures. Two clean speech signals are convolved with a set of measured binaural room impulse responses (BRIRs), [23]. Two groups of tests are conducted. One group contains source signals that are sentences of two male speakers, 5 seconds in duration, recorded at a sampling rate of 10 kHz (kilo-Hertz). The other group consists of two source signals from two female speakers of same duration as in the first group. The average separation quality is considered. The BRIRs are measured in a $5 \times 9 \times 3.5$ m ordinary classroom using the Knowles Electronic Manikin for Auditory Research (KEMAR), positioned at 1.5 m above the floor and at ear level [23]. By convolving the speech signals with the pre-measured room impulse responses, one source is virtually placed directly at the front of the listener and the other at an angle of 60° in the azimuth to the right, while both are located at 1 m away from the KEMAR. We then calculate the reverberation time to quantify the degree of difficulty of a separation task [22]. The reverberation time, denoted by T_{60}, is the time it takes for the energy of an impulse sound (such as a startgun shot) to decay by 60 decibel (dB) [1]. To assess the separation ability of the algorithm, we calculate the signal-to-interference-ratio improvement (SIRI, [13]) by measuring the overall amount of crosstalk reduction achieved by the algorithm *before* (SIR_i) and *after* (SIR_o) the demixing stage. Following [13], the SIRI in decibel (dB) is:

$$\text{SIRI} \triangleq \text{SIR}_o - \text{SIR}_i \triangleq$$
$$10 \log_{10} \left(\sum_{i=1}^{m} \frac{\sum_{l=1}^{L_s} |\hat{s}_{ii}|^2}{\sum_{j=1, j\neq i}^{n} \sum_{l=0}^{L_s} |\hat{s}_{ij}|^2} \right) - 10 \log_{10} \left(\sum_{i=1}^{m} \frac{\sum_{l=1}^{L_s} |x_{ii}|^2}{\sum_{j=1, j\neq i}^{n} \sum_{l=0}^{L_s} |x_{ij}|^2} \right) \tag{3.1}$$

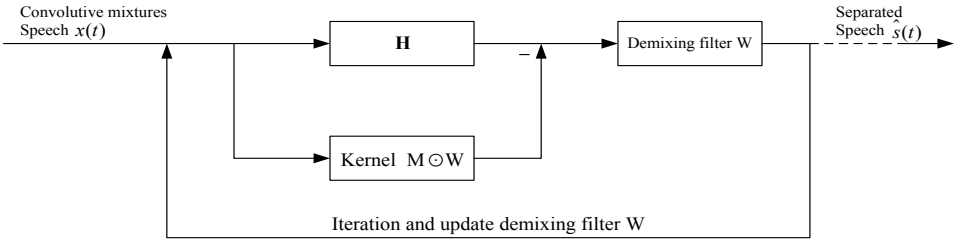

FIGURE 6. Schematic diagram of NLW-SCNG. The input mixed reverberant speech signals are iterated frame by frame to update demixing filter through nonlocally weighted and soft-constrained natural gradient iterations. The notation $\odot$ stands for convolution $(*)$.

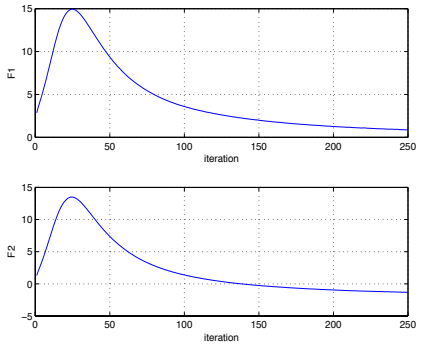

FIGURE 7. Convergence of control variables F_1 and F_2 in terms of iteration numbers.

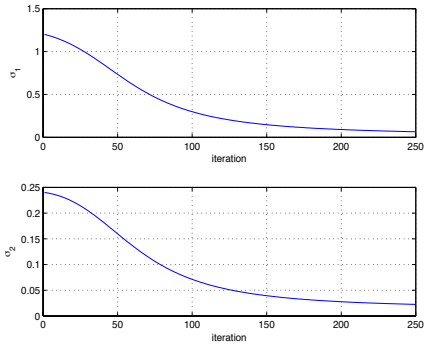

FIGURE 8. Convergence of scaling variables σ_1 and σ_2 in terms of iteration numbers.

where the $\hat{s}$'s are the output signals, and x's are the input signals; m denotes the number of microphones, n denotes the number of sources, and L_s is for the length of speech signal.

Next, let us compare the performance of a couple of convolutive BSS algorithms in reverberant room conditions based on SIRI. As noted in [10, 12], time domain natural gradient algorithm (NGTD) and spatio-temporal FastICA algorithm (STFICA) are significantly better than other methods, such as Parra's decorrelation-based method [21], and bin-wise natural gradient frequency-domian (NGFD) without beamforming initialization. Therefore we compared the following algorithms:

- STFICA algorithm, with sequential and symmetric orthogonalization and one stage least squares prewhitening at the length of $M = 400$ taps per filter [12]. The separation system used $L = 500$ taps per input-output filter channel.
- Natural gradient time-domain (NGTD) method without prewhitening or any special initialization. The initial condition of W is given by $W_0(1) = 0.001I$, $W_0(p) = 0$ if $p > 1$; step size $\mu = 0.2$ [10], demixing filter length $L = 1000$; frame length 10^4, and frame step size equal to 100 sample points.

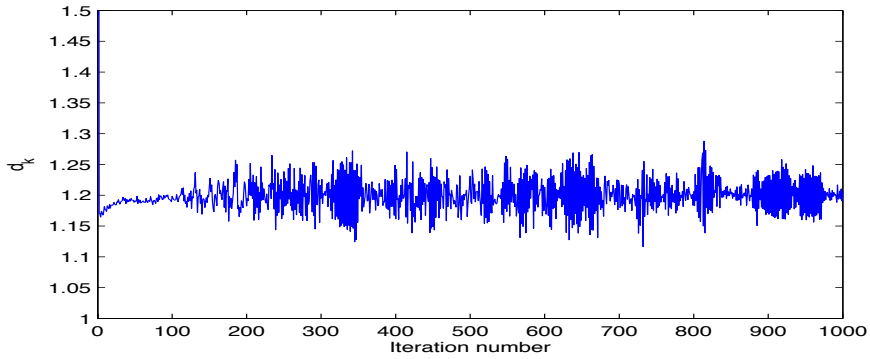

FIGURE 9. Scaling variable d_k as a function of iteration number k shows oscillatory behavior without nonlocal weighting and soft constraint.

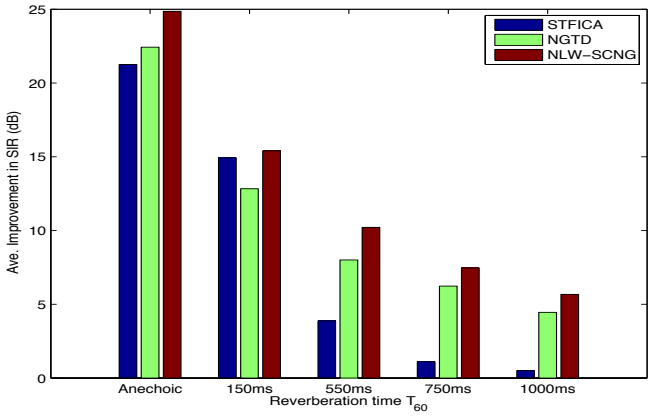

FIGURE 10. Comparison of SIR improvement of three separation methods (STFICA, NGTD, NLW-SCNG) in different reverberation conditions (anechoic, 150 ms, 550 ms, 750 ms and 1 s).

- Nonlocally weighted soft-constrained natural gradient (NLW-SCNG) method as shown in Fig. 6; kernel function as in Fig. 5. The initial condition of W is same as NGTD; and initial $\sigma_1 = 1.2$, $\sigma_2 = 0.24$. Parameter $\nu = 0.00125$, $c_1 = 1$, $c_2 = 3$, $E = 0.04$, frame length 10^4, frame step size equal to 100 sample points, and demixing filter length $L = 1000$.

Shown in Fig. 7 and Fig. 8 are the evolution of control variables (F_1, F_2) and the scaling variables (σ_1, σ_2) in terms of iteration numbers. The algorithm in moderately reverberant environment (reverberation time $T_{60} = 150$ ms) converges in 250 iterations as illustrated in the Figures. Without *nonlocal weighting* and *soft-constraint*, the control and scaling sequences appear as persistent oscillations when mixture signals contain no silence durations (as in Fig. 11), see Fig. 9 on scaling variable d_k of recurrence (2.3). In contrast, the control variables (F_1, F_2) and scaling variables (σ_1, σ_2) in the new algorithm NW-SCNG converge after an initial transient period.

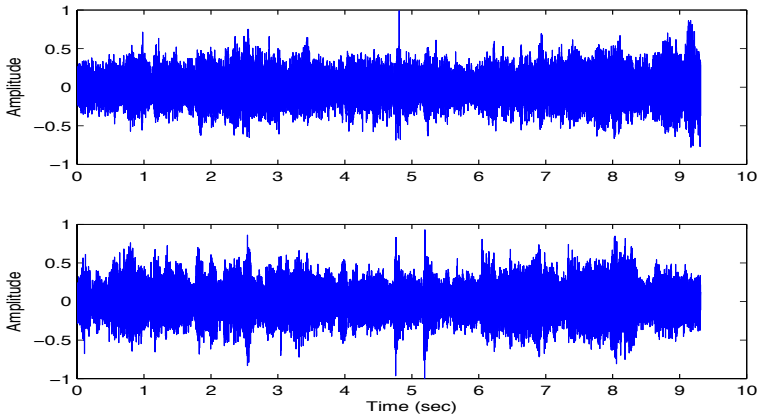

FIGURE 11. Recorded mixture of a song and a violin piece.

To compare the separation performance of convolutive BSS algorithms, we mixed two-source speech samples using BRIRs of increasing reverberation times, from anechoic to strongly reverberant $T_{60} = 1$ s, see Fig. 10. It is well known that beamforming initialization and data prewhitening improve separation. In order to assess the algorithms themselves, NGTD and NLW-SCNG are without these preprocessing steps, while STFICA is with prewhitening since it is not claimed to work without such processing. In an anechoic room, NLW-SCNG improves SIR by 24.9 dB while STFICA and NGTD increases SIR by 21.3 dB and 22.4 dB respectively. In anechoic room conditions, the difference is minor. If reverberation time $T_{60} = 150$ ms, NLW-SCNG processing gains 15.4 dB in SIR, STFICA is slightly lower yet exceeds NGTD by 2.1 dB gain in SIR. As reverberation time goes up to $T_{60} = 550$ ms, the SIR gains by STFICA drop sharply to less than 5 dB. NLW-SCNG achieves the highest gain of 10.2 dB in SIR improvement, with NGTD lagging behind by 2.2 dB. At 750 ms reverberation time, the STFICA gains less than 1 dB in SIR. At $T_{60} = 1$ second, NGTD falls under 5 dB in SIRI. In contrast, performance of NLW-SCNG goes down most slowly as reverberation becomes stronger. It performs the best among the three algorithms and gains more than 5 dB even when reverberation time reaches 1 second.

We applied NLW-SCNG algorithm to mixtures of violin and song recorded in our lab room with reverberation time about 0.5 second. The sound sources and microphones form a unit (1 m) square. The recorded mixture signals are plotted in Fig. 11. The two outputs by NLW-SCNG and NGTD are in Fig. 12 and Fig. 13 respectively. The extraction of the song in the lower panel of Fig. 12 is seen to contain less music.

4. **NLW-SCNG with dereverberation preprocessing.** In this section, we study a preprocessing method to reduce late reverberations, and combine it with NLW-SCNG to improve separation performance on strongly reverberant speech ($T_{60} = 1$ s). Room reverberation causes degradation of speech signals [5] and performance deterioration of BSS algorithms [24]. Room reverberations are composed of early reflections and late reflections [5]. The distinction is marked by 50 ms [16]. Early reflections depend on details of reflecting surfaces in a room, its relatively

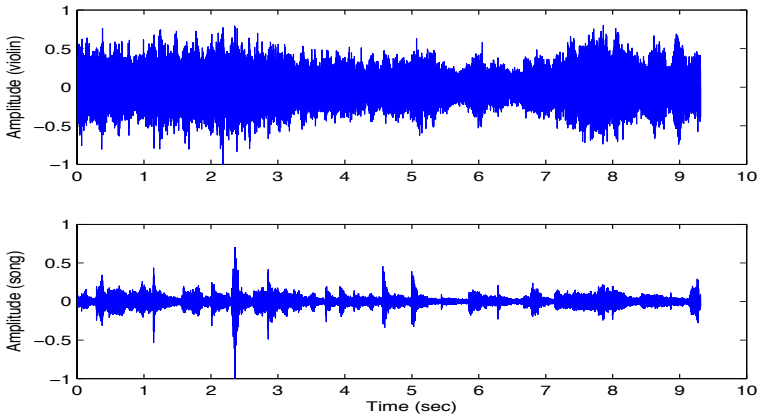

FIGURE 12. Output signals from NLW-SCNG algorithm.

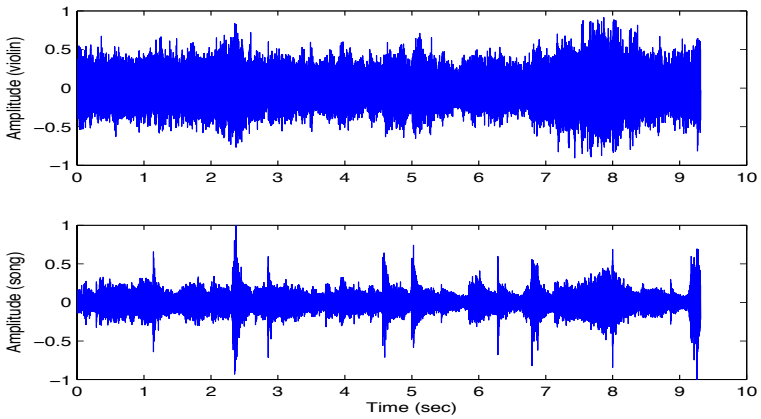

FIGURE 13. Output signals from NGTD algorithm.

short duration amounts to 400 filter taps at 8 kHz sampling rate. The late reflections smear the speech spectra and reduce the intelligibility and quality of speech signals. It is the late reflections of impulse responses that are noisy and expensive to resolve computationally. The idea then is to estimate and remove the contributions from late impulse responses, and resolve the early impulse responses which are in lower dimensions.

We shall integrate a simple and efficient one-microphone spectral subtraction method [25] into NLW-SCNG as a preprocessing step. The smearing effects of late impulses lead to the smoothing of the signal spectrum [25]. The power spectrum of late-impulse components is a smoothed and shifted version of the power spectrum of the reverberant speech [25]:

$$|S_l(k;i)|^2 = \gamma\, \omega(i-\rho) * |S(k;i)|^2, \tag{4.1}$$

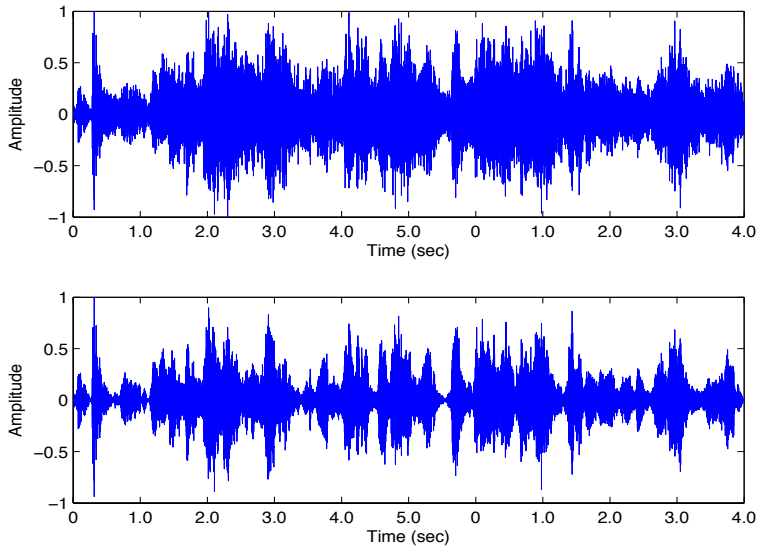

FIGURE 14. Time domain plots of reverberant speech enhancement. Upper panel: two-source mixture of 5 second duration at 10kHz sampling frequency in a room with reverberation time $T_{60} = 1$ second. Lower panel: the dereverberated speech processed by spectral subtraction.

where $|S(k;i)|^2$ and $|S_l(k;i)|^2$ are the short-term (framewise) power spectra of the reverberant speech and the late-impulse components respectively; γ is a positive parameter depending on the degree of reverberations. Indices k and i refer to frequency bin and time frame. The symbol $*$ denotes convolution in the time domain and $\omega(i)$ is a smoothing function in the form of the Rayleigh distribution:

$$\omega(i) = \begin{cases} \frac{i+a}{a^2} \exp(\frac{-(i+a)^2}{2a^2}), & \text{if } i > -a \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

The power spectrum of the early-impulse components can be estimated by subtracting the power spectrum of the late-impulse components from that of the reverberant speech. Specifically, spectral subtraction [9] is employed to estimate the power spectrum of original speech $|S_o(k;i)|^2$

$$|S_o(k;i)|^2 = |S(k;i)|^2 \max\Big[\frac{|S(k;i)|^2 - |S_l(k;i)|^2}{|S(k;i)|^2}, \epsilon\Big] \tag{4.3}$$

where the parameters $\epsilon = 0.001$, $a = 5$, and $\rho = 7$. Finally, the phase spectrum of the enhanced speech is set to that of the input speech. The processed speech is reconstructed from the magnitude and phase spectrum.

Effects of reverberant speech enhancement are shown in the time domain (Fig. 14) and the frequency domain (Fig. 15) where one second long reverberation blurs a source speech signal. After processing by spectral subtraction method first to reduce long-term reverberations, the reverberation is reduced and speech intelligibility is improved. Our simulations showed that room reverberation time $T_{60} = 0.55\,s$ can

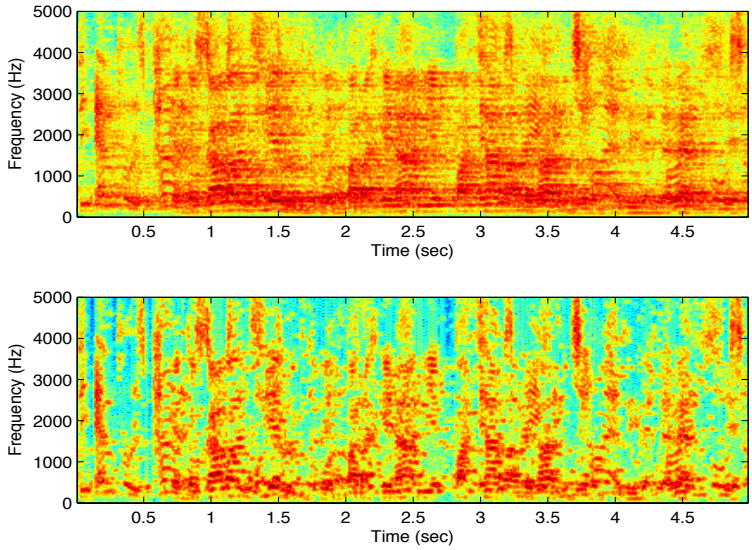

FIGURE 15. Plots of reverberant speech enhancement. Upper panel: spectrogram of the reverberant speech in in Fig. 14. Lower panel: spectrogram of the dereverberated speech by spectral subtraction.

be approximately reduced to the range $[0.15, 0.20]\,s$; $T_{60} = 0.75\,s$ to $[0.20, 0.25]\,s$; and $T_{60} = 1\,s$ to $[0.25, 0.30]\,s$.

After preprocessing on the mixture data in Section 3, we then apply the NLW-SCNG algorithm. Blind separation in strongly reverberant environment ($T_{60} = 0.55$ s, 0.75 s and 1.0 s) is studied again. The scaling factor γ in (4.1) takes the value of 1.2, 1.0 and 0.8 respectively. Fig. 16 illustrates the SIR gains after the combined two-stage processing. With the help of preprocessing, additional SIR improvement is observed, and is particularly valuable when the reverberation time is long, namely 3.4 dB more SIR at $T_{60} = 1\,s$, 2.0 dB more SIR at $T_{60} = 0.55\,s$. As a result, the overal SIR improvement is 9 dB at $T_{60} = 1\,s$.

5. **Discussion and conclusion.** We introduced a nonlocally weighted soft constrained natural gradient time domain recursive algorithm to resolve the inconsistency and small divisor problems in previous convolutive natural gradient methods. The new algorithm is stable and rapidly convergent in numerical experiments. It offers over 10 dB signal-to-interference ratio improvement in reverberant room with reverberation time T_{60} =550 ms, and over 5 dB at $T_{60} = 1$ s, outperforming several recent methods in the literature. When integrated with a recent spectral subtraction dereverberation method, as much as 9 dB gain in SIR is achieved in strongly reverberant condition ($T_{60} = 1$ s).

We applied a spectral subtraction method [25] as a preprocessing step to reduce the reverberation effect and achieve better separation. However, preprocessing may also introduce signal distortions. In future work, we shall study how to improve the recursive algorithm by spectral widening technique [26] and filter bank based bandwise processing. We plan to perform mathematical analysis on uniform boundedness and convergence of the proposed recursive algorithm.

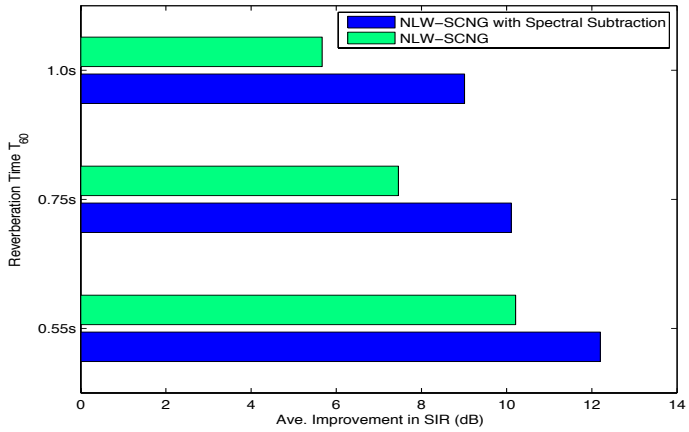

FIGURE 16. Comparison of SIR improvement of NLW-SCNG with preprocessing by spectral subtraction.

Acknowledgments. The authors would like to thank Deliang Wang, Hsin-I Yang and Qiang Liu for helpful conversations during the preparation of this work.

REFERENCES

[1] J. Allen, *Effects of small room reverberation on subjective preference*, Journal of Acoustical Society of America, **71** (1982), S5.

[2] S. Amari, A. Cichocki and H-H. Yang, *A new learning algorithm for blind signal separation*, Adv. Neural Infomation Processing System, **8** (1996), 757–763.

[3] S. Araki, R. Mukai, S. Makino, T. Nishikara and H. Saruwatari, *The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech*, IEEE Trans. Speech Audio Processing, **11** (2003), 109–116.

[4] A. Bijaoui and D. Nuzillard, "Blind Source Separation of Multi-Spectral Astronomical Images," in Proc. MPA/ESO/MPE Joint Astronomy Conference on Mining the Sky, Germany, Aug. 2000.

[5] A. Boothroyd, *Room acoustics and speech perception*, Seminars in Hearing, **25** (2004), 155–166.

[6] G. Brown and D-L. Wang, *Separation of speech by computational auditory scene analysis*, in "Speech Enhancement" (eds. J. Benesty, S. Makino and J. Chen), Springer, New York, (2005), 371–402.

[7] S. Choi, A. Cichocki, H. Park and S. Lee, *Blind source separation and independent component analysis: A review*, Neural Information Processing Letters and Reviews, **6** (2005), 1–57.

[8] A. Cichocki and S. Amari, "Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications," John Wiley and Sons, 2005.

[9] J. R. Deller, J. G. Proakis and J. H. L. Hansen, "Discrete-Time Processing of Speech Signals," Upper Saddle River, NJ: Prentice-Hall, 1987.

[10] S. Douglas and M. Gupta and H. Saruwatari, *Scaled natural gradient algorithm for instantaneous and convolutive blind source separation*, in Proc. IEEE ICASSP, **II** (2007), 637–640.

[11] S. Douglas and M. Gupta,, *Convolutive blind source separation for audio signals*, in "Blind Speech Separation" (eds. S. Makino, T-W. Lee and H. Sawada), Signal and Communication Technology, Springer, (2007), 3–45.

[12] S. Douglas, M. Gupta, H. Sawada and S. Makino, *Spatio-temporal FastICA algorithms for the blind separation of convolutive mixtures*, IEEE Trans. on Audio, Speech, and Language Processing, **15** (2007), 1511–1520.

[13] K. Kokkinakis and P. Loizou, *Subband-based blind signal processing for source separation in convolutive mixtures of speech*, in Proc. IEEE ICASSP, **IV** (2007), 917–920.

[14] J. Kolba and I. Jouny, *Blind source separation in tumor detection in mammograms*, in Proc. of the IEEE 32nd Annual Northeast Bioengineering Conference, (2006), 65–66.
[15] H. J. Kushner and G. Yin, "Stochastic Approximation and Recursive Algorithms and Applications," 2^{nd} edition, Springer-Verlag, New York, 2003.
[16] H. Kuttruff, "Room Acoustics," Taylor & Francis, 2000.
[17] J. Liu, J. Xin and Y. Qi, *A dynamic algorithm for blind separation of convolutive sound mixtures*, Neurocomputing, **72** (2008), 521–532.
[18] J. Liu, J. Xin, Y. Qi and F-G Zeng, *A time domain algorithm for blind separation of convolutive sound mixtures and l_1 constrained minimization of cross correlations*, Comm Math Sciences, **7** (2009), 109–128.
[19] J. Liu, J. Xin and Y. Qi, *A soft-constrained dynamic iterative method of blind source separation*, SIAM Interdisciplinary Journal on Multiscale Modeling and Simulations, **7** (2009), 1795–1810.
[20] M. Naceur, M. Loghmari and M. Boussema, *The contribution of the sources separation method in the decomposition of mixed pixels*, IEEE Trans. on Geoscience and Remote Sensing, **42** (2004), 2642–2653.
[21] L. Parra and C. Spence, *Convolutive blind separation of non-stationary sources*, IEEE Trans. Speech Audio Processing, **8** (2000), 320–327.
[22] M. Schroeder, *New method of measuring reverberation time*, Journal of the Acoustical Society of America, **37** (1965), 409–412.
[23] B. Shinn-Cunningham, N. Kopco and T. Martin, *Localizing nearby sound sources in a classroom: Binaural room impulse responses*, Journal of the Acoustical Society of America, **117** (2005), 3100–3115.
[24] D-L. Wang and G. Brown, "Computational Auditory Scene Analysis: Principles, Algorithms, and Applications," John Wiley and Sons, 2006.
[25] M. Wu and D-L. Wang, *A two-stage algorithm for one-microphone reverberant speech enhancement*, IEEE Trans. on Audio, Speech, and Language Processing, **14** (2006), 776–778.
[26] H. Yasukawa, *A simple method of broad band speech recovery from narrow band speech for quality enhancement*, in Proc. of the IEEE Digital Signal Processing Workshop, (1996), 173–175.

Received October 2009; revised February 2010.

E-mail address: myu3@uci.edu
E-mail address: jxin@math.uci.edu