A NONLOCALLY WEIGHTED SOFT-CONSTRAINED NATURAL GRADIENT ALGORITHM FOR BLIND SEPARATION OF REVERBERANT SPEECH

Jack Xin, Meng Yu, Yingyong Qi, Hsin-I Yang, Fan-Gang Zeng

Mathematics and Biomedical Engineering
University of California at Irvine
Irvine, CA 92697, USA.
{jack.xin, m.yu}@uci.edu.

ABSTRACT

A nonlocally weighted soft-constrained natural gradient iterative method is introduced for robust blind separation in reverberant environment. The nonlocal weighting of the iterations promotes stability and convergence of the algorithm for long demixing filters. The scaling degree of freedom is controled by soft-constraints built into the auxillary difference equations. The small divisor problem of iterations in silence durations of speech is resolved. Computations on synthetic speech mixtures based on measured binaural room impulse responses show that the algorithm achieves higher signal-to-inteference ratio improvement than existing method (natural gradient time domain algorithm) in an office size room with reverberation time over 0.5 second.

Index Terms— Adaptive estimation, weighted and controled iteration, reverberant speech, blind separation.

1. INTRODUCTION

A variety of methods rooted in the independent component analysis have been generalized to convolutive mixtures in the past decade, [2, 3, 4, 5, 6, 7, 8, 9, 10, 12] among others. However room reverberation remains one of the main barriers to applications of blind source separation methods (BSS) in realistic listening conditions. A typical office size room has reverberation time (T_{60}, [1]) near 0.5 second. At 10 kHz sampling frequency, separating two channel mixtures of two sources already requires finding 5000×4 or 20k demixing filter taps. It is incredibly challenging to find an accurate solution to this high dimensional problem, considering also that the associated objective function is non-convex. On the other hand, a practically useful BSS method should perform well in a reverberant environment ($T_{60} \geq 0.5s$), with at least 5 dB improvement after processing. The 5 dB is a typical gain of beamforming directional microphones.

In this paper, we introduce a new time domain convolutive natural gradient algorithm for robust BSS of reverberant speech signals. Computational results shall demonstrate that it outperforms the recent method [4], and improves signal-to interference ratio (SIR) of reverberant speech ($T_{60} = 0.56$ second) by over 10 dB.

To motivate the main ingredients of our algorithm design, let us recall the existing convolutive natural gradient method [2, 4]. Let the mixing model be

$$x(t) = [A * s](t) \tag{1.1}$$

where $[A * s](t) = \sum_{i=0}^{q} A(i)s(t-i)$ is the linear convolution of source vector $s \in R^n$ with $A(i)$, a group of $n \times n$ matrices. Let L be the demixing filter length, and N be the number of data points in each block (frame). The demixing (inversion) formula is:

$$y_k(l) = \sum_{p=0}^{L} W_k(p)\, x(l-p). \tag{1.2}$$

The scaled natural gradient method [4] is:

$$W_{k+1}(p) = (1+\mu)\, d_k^{-1}\, W_k(p) - \mu\, d_k^{-2}\, H_k(p), \tag{1.3}$$

where μ is the learning parameter, the scaling factor d_k is linear in the mixture signal x in the current block, and:

$$H_k(p) \triangleq \frac{1}{N} \sum_{l=(k-1)N+1}^{kN} \sum_{q=0}^{L} f(y_k(l)) y_k^T(l-p+q) W_k(q), \tag{1.4}$$

where y is related x as in (1.2), and f is the sign function. The scaling variable $d(k)$ is intended to cure divergence that arises if d_k were a positive constant independent of k as originally proposed [2]. However, during silence (or near silence) period of speech, d_k may be very small to cause a drastic growth (instability) in the solution, or a small divisor problem. Specifically,

$$d_k = \frac{1}{n} \sum_{i,j=1}^{n} \sum_{q=0}^{2L} |g_k^{ij}(q)|, \tag{1.5}$$

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

$$G_k(q) \triangleq \sum_{p=0}^{L} F_k(q-p)\, W_k^{\top}(L-p), \quad q \in [0, 2L] \tag{1.6}$$

where the matrix F is:

$$F_k(p) \triangleq \frac{1}{N} \sum_{l=(k-1)N+1}^{kN} f(y_k(l))\, x^{\top}(l-p). \tag{1.7}$$

In the sum of (1.6), $F_k(p)$ is set to zero if $p \notin [0, L]$. It follows from a direct calculation that

$$H_k(p) = \sum_{q=0}^{L} G_k(L+p-q)\, W_k(q), \tag{1.8}$$

Partially supported by NSF Grant DMS-0712881, NIH Grant 2R44DC006734, UCI grant CORCLR-MI-2006-07-6.

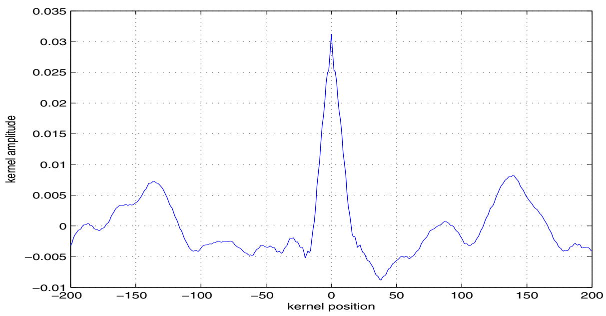

Figure 1: Measured (sliding window) generalized correlation function from a clean speech signal

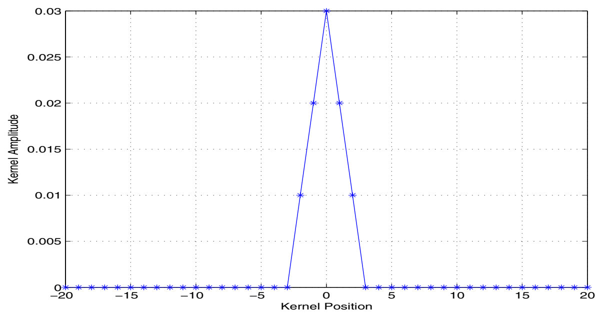

Figure 2: Piecewise linear kernel function used in the computation (right), with base width $L_M = 5$ and peak height 0.03.

which is another way to express $H_k(p)$ in (1.4). One observes that if $x(i) = 0$ for i in a block, then $F_k(p) = 0$ by (1.7), implying that $G_k(p) = 0$ for $p \in [0, L]$, and $d_k = 0$. Likewise, if $x(i) \approx 0$ for i in a block, $d_k \approx 0$ causing small divisor problem in (1.3). The expressions (1.6)-(1.8) are much less transparent than (1.4), but are helpful for implementation.

There is also an inconsistency problem in (1.2)-(1.4). Suppose that y_k converges up to a limiting signal s with independent components at large values of N, and that $W_k(q)$ converges to $W_\infty(q)$, then

$$H_\infty(p) \equiv \lim_{k\to\infty, N\to\infty} H_k(p) = \sum_{q=0}^{L} E[f(s)\, s^T(q-p)] W_\infty(q), \tag{1.9}$$

which is a *convolutive (non-local) product* that *cannot be balanced by the local term* (constant multiple of W_k) in (1.3)! The generalized correlation matrix $E[f(s)(\cdot)\, s^T(\cdot - \eta)]$ 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, the matrix $E[f(s)(\cdot)\, s^T(\cdot - \eta)]$ is zero if $\eta \neq 0$, and the sum in (1.9) reduces to a local multiplication. In other words, with *local term* $(1+\mu)\, d_k^{-1}\, W_k(p)$ in (1.3), convergence cannot happen for mixtures of colored stochastic signals such as speech for which f is the sign function as a consequence of Laplace distribution of speech amplitudes. The solution is to adopt a *nonlocal term* which is equal to convolution of a kernel function with $W_k(p)$, see (2.2) below.

The present contribution resolves both the inconsistency and the small divisor scaling problem for convolutive source separation by equiping the natural gradient iteration with two important ingredients: *nonlocal weighting and soft constraints*. In section 2, the new algorithm and its properties are outlined. In section 3, experimental results on reverberant speech show robust and excellent performance. Conclusions are in section 4.

2. NONLOCALLY WEIGHTED SOFT-CONSTRAINED ITERATIVE SCHEME

Let us introduce the nonlocally weighted demixing matrix iteration as:

$$W_{k+1}(p) = W_k(p) + \sigma_{1,k}\, [M \odot W_k](p) - \sigma_{2,k}\, H_k(p), \tag{2.1}$$

where H is by (1.4) and $\odot$ is a nonlocal product with a nonlinear kernel function M:

$$[M \odot W_k](p) \triangleq \sum_{i=1}^{L_M} M(i) W_k\left(p - \frac{L_M - 1}{2} + i - 1\right) \tag{2.2}$$

and 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. 2. It is constructed from numerical data of (sliding window) generalized correlation functions of clean speech signals, see the Fig. 1 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 the right panel. The shape of kernel function is obtained by optimal searching. The value of W_k on the right hand side of (2.2) is set to zero if $p - \frac{L_M-1}{2} + i - 1 \notin [0, L]$.

Notice that the nonlocal product in (2.2) captures the limiting form of $H_k(p)$ in (1.9). For simplicity, we have chosen the same kernel function M for all entries of $W_k(p)$ matrix. The locally scaled version (1.3) is a special case, for example 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 equations:

$$\begin{aligned} \sigma_{1,k+1} &= \sigma_{1,k} \exp\{-\nu F_{1,k}(\sigma_{1,k}, \sigma_{2,k})\} \\ \sigma_{2,k+1} &= \sigma_{2,k} \exp\{-\nu F_{2,k}(\sigma_{1,k}, \sigma_{2,k})\} \end{aligned} \tag{2.3}$$

where ν is a positive constant, and:

$$F_{1,k} \triangleq \sigma_{1,k} \sum_{p=0}^{L} \sum_{j=1}^{n} |W_k^{1,j}(p)| + \sigma_{2,k} \sum_{p=0}^{L} \sum_{j=1}^{n} |H_k^{1,j}(p)| - c_1, \tag{2.4}$$

$$\begin{aligned} F_{2,k} \triangleq \sigma_{1,k} \sum_{p=0}^{L} \sum_{j=1}^{n} |W_k^{2,j}(p)| + \sigma_{2,k} \sum_{p=0}^{L} \sum_{j=1}^{n} |H_k^{2,j}(p)| \\ - c_2 + \sigma_{2,k} E \end{aligned} \tag{2.5}$$

where c_1, c_2 and E are positive constants.

Equations (2.3)-(2.5) 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 (2.1) when $|W|$ is small enough. It follows that the dynamics by (2.1)-(2.5) 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, $< \cdot >$ being a temporal average in t, [11].

3. EXPERIMENTAL RESULTS

The proposed nonlocally weighted soft-constrained natural gradient (NLW-SCNG) 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), [14]. The source signals are sentences of two male speakers, 5 seconds in duration, recorded at a sampling rate of 10 kHz. 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 [14]. The BRIR data are more faithful to the setting of a human wearing hearing aids, and are also used in [4, 8]. 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 [13]. To assess the separation ability of the algorithm, we calculate the signal-to-interference-ratio improvement (SIRI, [8]) by measuring the overall amount of crosstalk reduction achieved by the algorithm *before* (SIR_i) and *after* (SIR_o) the demixing stage. Following [8], the SIRI in dB is:

$$\begin{aligned} \mathrm{SIRI} &\triangleq \mathrm{SIR}_o - \mathrm{SIR}_i \\ &\triangleq 10\log_{10}\left(\sum_{i=1}^{m} \frac{\sum_{l=1}^{Length} |\hat{s}_{ii}|^2}{\sum_{j=1,j\neq i}^{n}\sum_{l=0}^{Length} |\hat{s}_{ij}|^2}\right) \\ &- 10\log_{10}\left(\sum_{i=1}^{m} \frac{\sum_{l=1}^{Length} |x_{ii}|^2}{\sum_{j=1,j\neq i}^{n}\sum_{l=0}^{Length} |x_{ij}|^2}\right) \end{aligned} \tag{3.6}$$

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 $Length$ is for the length of speech signal. We shall focus on SIRI in this work, and report on subjective speech quality measure in a future study.

Let us compare the novel algorithm NLW-SCNG with the recent natural gradient time domain algorithm (NGTD). It is well known that beamforming initialization and data prewhitening improve separation. In order to evaluate the algorithms themselves, neither NGTD nor NLW-SCNG has beamforming initialization and prewhitening preprocessing.

First consider the BRIR mixtures with reverberation time $T_{60} =$ 150 ms. We use the piecewise linear kernel in Fig. 2 with base width 5 and peak height 0.03. Two schemes are tested. Scheme I (in Fig. 3) refers to applying NLW-SCNG once through the mixtures, and scheme II (in Fig. 4) refers to processing by reinitializing and rerunning NLW-SCNG on the output of scheme I. The parameters are: $\nu = 0.00125$, $c_1 = 1$, $c_2 = 3$, $E = 0.04$, frame length is 10000, frame step size is 100 point, and demixing filter

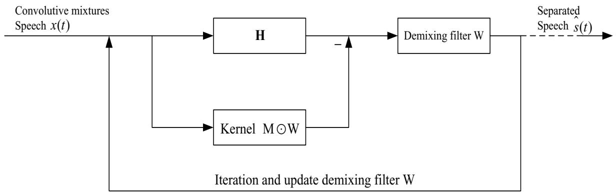

Figure 3: Schematic diagram of Scheme I (NLW-SCNG). The input mixed reverberant speeches are iterated frame by frame to update demixing filter through nonlocally weighted and soft-constrained natural gradient iterations.

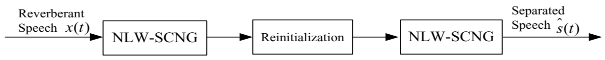

Figure 4: Schematic diagram of Scheme II (NLW-SCNG with reinitialization). The first-step output "separated" speech signals are sent into NLW-SCNG again as input "mixed" speech signals.

length L is 1000. The initial conditions are: $W_0(0) = 0.2\, I_2$, I_2 the 2x2 identity matrix; $W_0(p) = 0$, $p \geq 1$; $\sigma_1 = 1.2$ and $\sigma_2 = 0.24$. We found that scheme II improves on scheme I under 1 dB in slightly reverberant conditions ($T_{60} \leq 150$ ms). The improved SIR is 17.37 dB.

Shown in Fig. 5 and Fig. 6 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 reverberant environment converges in approximately 250 iterations as illustrated in the Figures. If only *soft-constrained*, i.e. without *nonlocal weighting*, the control and scaling sequences appear as damped oscillations which may take as many as 3000 iterations to converge. With additional *nonlocal weighting*, the control variables (F_1, F_2) and scaling variables (σ_1, σ_2) in the new algorithm NLW-SCNG converge significantly faster.

For the NGTD algorithm, the initial condition of W is given by $W_0(1) = 0.001I$, $W_0(p) = 0$ if $p > 1$; step size $\mu = 0.2$ [4], demixing filter length $L = 1000$; frame length 10000, and frame step size is equal to 100 sample points. Under the acoustic environment $T_{60} = 150$ ms, the improved SIR is 12.75 dB.

Now a typical reverberant room environment with $T_{60} = 560$

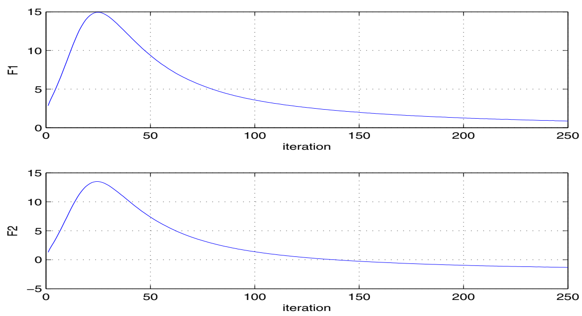

Figure 5: Convergence of control variables F_1 and F_2 in terms of iteration numbers.

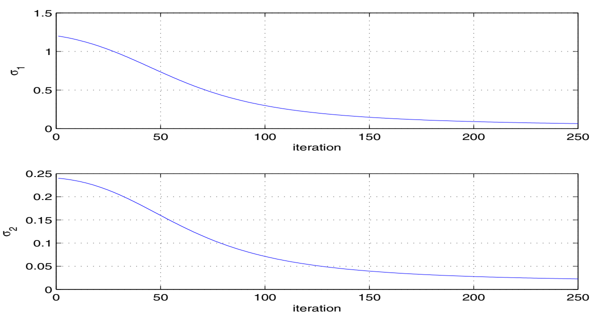

Figure 6: Convergence of scaling variables σ_1 and σ_2 in terms of iteration numbers.

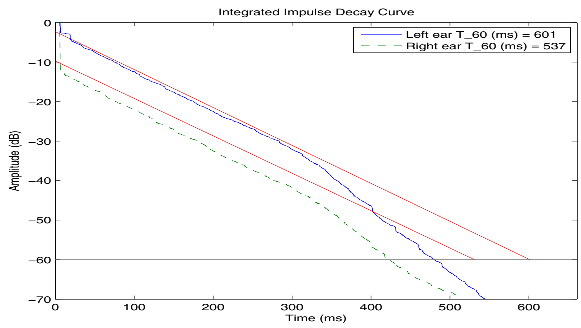

Figure 7: Energy decay curves of the left ear and right ear impulse responses. The straight lines are standard for calculating the intersection with -60 dB line to find reverberation time T_{60}[13].

ms is used to test NLW-SCNG and NGTD. The energy decay curves [13] of BRIRs are in Fig. 7. The parameters and initial conditions of the algorithms are kept the same as those for $T_{60} = 150$ ms. Table 1 shows that for NLW-SCNG, the scheme II with reinitialization improves on scheme I by nearly 1 dB, and both schemes exceed 9 dB SIRI, with NGTD lagging behind by 2.2 dB.

Scheme	SIR_i (dB)	SIR_o (dB)	SIRI (dB)
NLW-SCNG I	-1.53	7.85	9.38
NLW-SCNG II	-1.53	8.68	10.21
NGTD	-1.53	6.48	8.01

Table 1: Performances of NLW-SCNG scheme I, NLW-SCNG scheme II and NGTD in a reverberant room with $T_{60} = 560$ ms.

4. CONCLUSIONS

In this paper, we introduced a nonlocally weighted soft-constrained natural gradient time domain iterative algorithm to fix the inconsistency and small divisor problems in previous convolutive natural gradient methods. The new algorithm is stable and rapidly convergent, and it offers over 10 dB signal-to-interference ratio improvement in reverberant room with reverberation time $T_{60} \geq 500$ ms, outperforming recent convolutive BSS method NGTD algorithm. Future work will incorporate preprocessing [15] and evaluate the resulting algorithm in strongly reverberant environments ($T_{60} \approx 1$ s). We also plan to evaluate our NLW-SCNG algorithm on patients with hearing loss, and collect subjective scores on performance.

5. REFERENCES

[1] J. Allen,"Effects of small room reveberation on subjective preference", *J. Acoust. Soc. Amer*, 71(1982), p. S5.

[2] S. Amari, A. Cichocki, H-H. Yang, "A new learning algorithm for blind signal separation", *Adv. Neural Infom. Proc. System*, 8, pp. 757-763, MIT press, Cambridge, MA, 1996.

[3] S. Choi, A. Cichocki, H. Park, S. Lee,"Blind source separation and independent component analysis: A review", *Neural Information Processing Letters and Reviews*, Vol. 6, No. 1, 2005, pp. 1-57.

[4] S. Douglas, M. Gupta, "Scaled natural gradient algorithm for instantaneous and convolutive blind source separation", *Proc. IEEE ICASSP*, II 637- II 640, 2007.

[5] S. Douglas, M. Gupta, "Convolutive Blind Source Separation for Audio Signals", pp 3-45, in *Blind Speech Separation*, S. Makino, T-W. Lee, H. Sawada (Eds.), Signal and Communication Technology, Springer, 2007.

[6] S. Douglas, H. Sawada, and S. Makino, "A spatio-temporal FastICAalgorithm for separating convolutive mixtures", *Proc. IEEE ICASSP*, Vol. 5, pp. 165-168, Mar. 2005.

[7] N. Murata, S.Ikeda, A. Ziehe, "An approach to blind separation based on temporal structure of speech signals", *Neurocomputing*, 41(2001), pp. 1-24.

[8] K. Kokkinakis, P. Loizou, "Subband-based blind signal processing for source separation in convolutive mixtures of speech", *Proc. IEEE ICASSP*, IV, pp. 917-920, 2007.

[9] J. Liu, J. Xin and Y. Qi, "A dynamic algorithm for blind separation of convolutive sound mixtures", *Neurocomputing*, 72(2008), pp 521-532.

[10] J. Liu, J. Xin, Y. Qi, 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*, Vol. 7, No.1, pp 109-128, 2009.

[11] J. Liu, J. Xin, Y. Qi, "A Soft-Constrained Dynamic Iterative Method of Blind Source Separation", *Soc. Industrial Applied Math J. on Multi-scale Modeling and Simulations*, Vol.7, No.4, 2009 (on-line, DOI: 10.1137/080736168).

[12] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources", *IEEE Trans. Speech Audio Processing*, Vol.8, pp. 320-327, 2000.

[13] M. R. Schroeder, "New method of measuring reverberation time", *Journal of the Acoustical Society of America*, 37, pp. 409-412, 1965.

[14] B. Shinn-Cunningham, N. Kopco, T. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses", *J. Acoust. Soc. Amer*, 117(5), pp. 3100-3115, 2005.

[15] M. Wu, D-L. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement", *IEEE Transaction on Audio, Speech, and Language Processing*, Vol. 14, No. 3, pp. 776-778, 2006.