

A MODIFIED SPATIO-TEMPORAL ORTHOGONAL ITERATION METHOD FOR MULTICHANNEL AUDIO SIGNAL REPRESENTATION

Scott C. Douglas and Malay Gupta

Department of Electrical Engineering
Southern Methodist University
Dallas, Texas 75275 USA

ABSTRACT

In this paper, we present a novel algorithm for a spatio-temporal extension of the well-known method of orthogonal iterations in linear algebra. This algorithm estimates an n -input, m -output ($m < n$) paraunitary filter bank from a multichannel data autocorrelation sequence to maximize the total output power of the filter bank when applied to an n -dimensional input signal. We then show how this procedure can be used to generate reduced rank signal representations of recordings of m audio sources in a room as collected by an n -channel microphone array. The importance of the method for determining the number of active sound sources in a room for convolutive blind source separation is also discussed.

1. INTRODUCTION

Microphone arrays are important for a number of practical applications, including speech communications, sound source localization, sound reinforcement, and sound environment monitoring [1]. Recently, much work has been focused on blind array processing using convolutive blind source separation methods for speech enhancement and noise removal [2]. As sound propagation is almost always a linear convolutive process, a typical vector signal model for an unstructured microphone array is

$$\mathbf{x}(k) = \boldsymbol{\nu}(k) + \sum_{l=0}^{\infty} \mathbf{A}_l \mathbf{s}(k-l), \quad (1)$$

where $\mathbf{s}(k) = [s_1(k) \cdots s_m(k)]^T$ contains the m sound sources of interest, $\boldsymbol{\nu}(k) = [\nu_1(k) \cdots \nu_n(k)]^T$ contains the n sensor noise signals, and the $(n \times m)$ matrices \mathbf{A}_l with elements $\{a_{ijl}\}$ represents the multichannel room impulse response from the m sources to the n sensors in the array.

In many applications of microphone arrays, the number of sources of interest m is less than the number of microphones n . Thus, if the powers of the noises $\nu_i(k)$ are small relative to the source components as heard at the microphones, it is desirable to develop a signal representation of the form

$$\hat{\mathbf{x}}(k-D) = \sum_{l=0}^L \mathbf{B}_l \mathbf{y}(k-l) \quad (2)$$

that closely approximates the input signal $\mathbf{x}(k)$ with some time delay D , where $\{\mathbf{B}_l\}$ is the $(n \times m)$ multichannel reconstruction filter impulse response and the m signals in $\mathbf{y}(k) = [y_1(k) \cdots y_m(k)]^T$ are estimated from the measured $\{x_i(k)\}$ signals using a second $(m \times n)$ multichannel linear filter. When the noises

$\{\nu_i(k)\}$ are spatially and temporally-uncorrelated with identical powers, this problem statement is the spatio-temporal extension of a well-known reduced rank signal estimation problem in array processing, for which an answer when $L = D = 0$ is well-known:

$$\hat{\mathbf{x}}(k) = \mathbf{W}_0^T \mathbf{y}(k) \quad (3)$$

$$\mathbf{y}(k) = \mathbf{W}_0 \mathbf{x}(k) \quad (4)$$

$$\mathbf{W}_0 = \mathbf{Q} \mathbf{E}, \quad (5)$$

where $\mathbf{E} = [\mathbf{e}_1 \cdots \mathbf{e}_m]^T$ contains the m principal eigenvectors of the sample autocorrelation matrix

$$\mathbf{R}_0 = \frac{1}{N} \sum_{k=1}^N \mathbf{x}(k) \mathbf{x}^T(k), \quad (6)$$

and \mathbf{Q} is any $(m \times m)$ orthogonal matrix. Such an algorithm solves the following optimization problem:

$$\text{minimize} \quad \sum_{k=1}^N \|\mathbf{x}(k) - \mathbf{W}_0^T \mathbf{W}_0 \mathbf{x}(k)\|^2 \quad (7)$$

$$\text{such that} \quad \mathbf{W}_0 \mathbf{W}_0^T = \mathbf{I}, \quad (8)$$

where \mathbf{I} is an $(m \times m)$ identity matrix.

In this paper, we present an algorithm for finding the spatio-temporal equivalent of (5) that solves the following optimization problem, where L is even-valued:

$$\text{minimize} \quad \sum_{k=1}^N \|\mathbf{x}(k-L) - \mathbf{u}(k)\|^2 \quad (9)$$

$$\text{such that} \quad \sum_{q=0}^L \mathbf{W}_q \mathbf{W}_{q+l}^T = \mathbf{I} \delta_l, \quad -\frac{L}{2} \leq l \leq \frac{L}{2}, \quad (10)$$

where $\|\cdot\|$ is the Euclidean norm and

$$\mathbf{u}(k) = \sum_{q=0}^L \mathbf{W}_{L-q}^T \mathbf{y}(k-q) \quad (11)$$

$$\mathbf{y}(k) = \sum_{l=0}^L \mathbf{W}_l \mathbf{x}(k-l). \quad (12)$$

The iterative procedure developed in this paper can be viewed as a spatio-temporal equivalent of a modified orthogonal iterations method [3] that uses a novel embedded iterative procedure for enforcing the constraints in (10) at each iteration. As is obvious from

the problem formulation, such a procedure is useful for determining reduced rank approximations to multichannel audio recordings of m spatially-distinct sources in a room as collected by an n -channel microphone array. Numerical experiments on two different data sets illustrate the efficacy of the method. The importance of the method for determining the number of active sound sources in a room for convolutive blind source separation is also discussed.

2. THE ALGORITHM

To describe the algorithm, consider the cost function in (9) under the additional assumption that the sequence $\{\mathbf{x}(k)\}$ is prepended by $(L-1)$ zero vectors. Then, using a change of time variables, it is straightforward to show that an equivalent formulation to (9)–(10) is

$$\text{maximize} \quad \mathcal{J}(\{\mathbf{W}_q\}) = \frac{1}{2} \sum_{k=1}^N \mathbf{y}^T(k) \mathbf{y}(k) \quad (13)$$

$$\text{such that} \quad \sum_{q=0}^L \mathbf{W}_q \mathbf{W}_{q+l}^T = \mathbf{I} \delta_l, \quad -\frac{L}{2} \leq l \leq \frac{L}{2}. \quad (14)$$

Our method for solving (13)–(14) employs a gradient ascent procedure in which each matrix tap \mathbf{W}_l is replaced by the derivative of $\mathcal{J}(\{\mathbf{W}_q\})$ with respect to \mathbf{W}_l , after which the updated coefficient sequence is adjusted to maintain the paraunitary constraints in (14). It can be shown that

$$\frac{\partial \mathcal{J}(\{\mathbf{W}_q\})}{\partial \mathbf{W}_l} = \sum_{q=0}^L \mathbf{W}_q \mathbf{R}_{\frac{L}{2}-q+l}, \quad (15)$$

where the multichannel autocorrelation sequence \mathbf{R}_q is given by

$$\mathbf{R}_q = \frac{1}{N} \sum_{k=1}^N \mathbf{x}(k) \mathbf{x}^T(k-q). \quad (16)$$

Thus, the first step of our procedure at each iteration sets

$$\mathbf{W}_l^{(0)} = \sum_{q=0}^L \mathbf{W}_l \mathbf{R}_{\frac{L}{2}-q+l}, \quad 0 \leq l \leq L. \quad (17)$$

At this point, the coefficient sequence $\{\mathbf{W}_l^{(0)}\}$ needs to be modified to enforce the paraunitary constraints in (14) while maintaining the row span of the multichannel impulse responses. Recently, a technique for enforcing paraunitary constraints on an $(m \times n)$ multichannel FIR filter has been described [4]. This algorithm can be described in matrix form as follows: While $\{\mathbf{W}_l^{(p)}\}$ is not paraunitary, do

$$\mathbf{C}_l^{(p)} = \begin{cases} \sum_{q=0}^{L-l} \mathbf{W}_q^{(p)} \mathbf{W}_{q+l}^{(p)T}, & 0 \leq l \leq \frac{L}{2} \\ \mathbf{C}_l^{(p)T} & -\frac{L}{2} \leq l < 0 \\ \mathbf{0} & \text{otherwise.} \end{cases} \quad (18)$$

$$\mathbf{W}_l^{(p+1)} = \frac{3}{2} \mathbf{W}_l^{(p)} - \frac{1}{2} \sum_{q=0}^L \mathbf{C}_{\frac{L}{2}-q}^{(p)} \mathbf{W}_q^{(p)} \quad (19)$$

In this procedure, p is the iteration index of the adaptive paraunitary constraint procedure. The value of p is incremented, and

(18)–(19) are repeated until the matrix sequence $\mathbf{C}_l^{(p)}$ is close to $\mathbf{I} \delta_l$ for $|l| \leq L/2$. Once such a condition is achieved, the sequence \mathbf{W}_l is replaced by $\mathbf{W}_l^{(p)}$, and the entire process is repeated until an appropriate convergence condition on $\{\mathbf{W}_l\}$ is met. This condition could be determined by the change in the value of $\mathcal{J}(\{\mathbf{W}_q\})$ over time, or by the overall value of the error criterion in (9).

Several important points about this procedure can be made:

1. When $L = 0$, the procedure is mathematically equivalent to the following iterative method:

$$\mathbf{W}_0 \leftarrow (\mathbf{W}_0 \mathbf{R}_0^2 \mathbf{W}_0^T)^{-1/2} \mathbf{W}_0 \mathbf{R}_0, \quad (20)$$

where $(\mathbf{W}_0 \mathbf{R}_0^2 \mathbf{W}_0^T)^{-1/2}$ is the inverse of the symmetric square root of the matrix $\mathbf{W}_0 \mathbf{R}_0^2 \mathbf{W}_0^T$ and \leftarrow denotes an assignment procedure. The dynamics of the procedure in (20) are well-understood [3]. In particular, so long as the initial value of \mathbf{W}_0 satisfies $\mathbf{W}_0 \mathbf{e}_i \neq 0$ for $1 \leq i \leq m$ and the values of the m th and $(m+1)$ st ordered eigenvalues of \mathbf{R}_0 satisfy $\lambda_m > \lambda_{m+1}$, \mathbf{W}_0 is guaranteed to converge exponentially to the m -dimensional principal subspace of \mathbf{R}_0 with an asymptotic rate of at least λ_{m+1}/λ_m .

2. The procedure is related to recently-developed eigenfilter methods for multichannel processes [5]. In fact, the multichannel filter impulse responses produced by our procedure have been found through numerical experiments to closely approximate the subspaces generated from select eigenvectors of the $(nL \times nL)$ -dimensional sample autocorrelation matrix of the multichannel data. The critical difference between our procedure and these existing methods is that the block Toeplitz structure of the sample autocorrelation matrix is exploited in our approach, such that a single filter can represent an entire nL -dimensional signal subspace by multichannel shifts of the corresponding filter impulse responses.

3. The convergence of the iterative paraunitary procedure in (18)–(19) can be sensitive to filter edge effects caused by the finite value of L and to the overall scaling of $\{\mathbf{W}_l^{(0)}\}$. To address these concerns, we recommend that

- each filter impulse response $\{w_{ij0}^{(0)}, w_{ij1}^{(0)}, \dots, w_{ijL}^{(0)}\}$ be windowed using a tapered window function prior to applying the iterative procedure, *e.g.* using a Hamming window, and
- the initial value of each $\mathbf{W}_l^{(0)}$ be scaled according to the relation

$$\mathbf{W}_l^{(0)} \leftarrow \sqrt{3} \frac{\mathbf{W}_l^{(0)}}{\left(\sum_{q=0}^L \sum_{i=1}^m \sum_{j=1}^n |w_{ijq}^{(0)}| \right)^{1/2}} \quad (21)$$

The scaling used in (21) has been chosen based on the analysis in [4] and guarantees that the iterative paraunitary procedure is convergent. Typically, the windowing procedure is applied before the scaling constraint.

4. The complexity of the method is dominated by the computation of the spatio-temporal autocorrelation sequence \mathbf{R}_q , $|q| \leq L$, which is of $\mathcal{O}(n^2LN)$. Typically, the data block size N is much larger than the filter length L , so the coefficient updates require fewer computations to complete. In addition, the filtering operations can be implemented using FFT-based fast convolution procedures, such that the complexity of the multichannel convolutions in (17), (18), and (19) are of $\mathcal{O}(mn^2L \log_2 L)$, $\mathcal{O}(mn^2L \log_2 L)$,



Figure 1: Laboratory measurement environment used for numerical evaluations.

and $\mathcal{O}(m^2 n L \log_2 L)$, respectively. Note that the adaptive paraunitary constraint procedure in (18)–(19) typically converges in less than 20 iterations due to its quadratic convergence rate. Moreover, only two to five iterations of (17) were needed for each chosen filter length in the multichannel audio examples in the next section to obtain adequate convergence with no significant change in performance.

3. MULTICHANNEL AUDIO SIGNAL REPRESENTATION

The goal of multichannel audio signal representation is to calculate a set of m signals of length N and an $(n \times m)$ -channel FIR filter that can be used to accurately represent a set of n signals of length N , where $m < n$. The problem is loosely tied to multichannel audio compression, as the number of data samples needed to represent the n original signals is reduced by a factor of approximately m/n . Possible applications of this problem include efficient storage of high-quality multichannel audio for scientific or entertainment applications.

We can apply the algorithm described in the previous section to this task. In this case, the n estimated signals are given by $\mathbf{u}(k) = \hat{\mathbf{x}}(k - L)$ in (11), the m stored signals are given by $\mathbf{y}(k)$ in (12), and the reconstruction filter impulse response is given by

$$\mathbf{B}_l = \mathbf{W}_{L-l}^T, \quad 0 \leq l \leq L. \quad (22)$$

We now present numerical experiments evaluating the abilities of this procedure in this task. Data for these experiments were obtained from two sources: (1) a loudspeaker-microphone setup in the Multimedia Systems Laboratory at SMU, and (2) a publicly-available dataset generated for a source separation contest [6]. Figure 1 shows a photograph of the first setup. Three loudspeakers are used to play recordings of talkers—one female and two male—for simulating speech sources. These loudspeakers are located 127 cm away from four omnidirectional microphones and are spaced at angles of -30° , 0° , and 27° from the angle of incidence of the microphone array. The microphone array has a nominal 6cm spacing between the sensors. The room dimensions ($4.6\text{m} \times 4\text{m} \times 2.7\text{m}$) and acoustical treatment correspond to a reverberation time of 300ms. All measurements were made using 7 seconds of data per channel and a 48kHz sampling rate and were downsampled to an 8kHz sampling rate for processing.

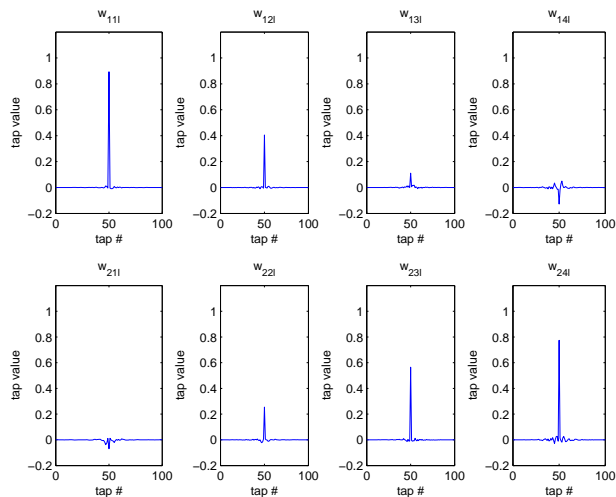


Figure 3: Example analysis/reconstruction filters for the MSLab data set: $m = 2$, $n = 4$, and $L = 100$.

The second data set used for algorithm testing was obtained from the Stereo Audio Source Separation Evaluation Campaign [6]. Although these recordings have been designed to explore two-channel underdetermined source separation, we discovered that several of the recordings in the test data were generated from the same musical sources at two different microphone spacings of 1m and 5cm separation, respectively. We have combined the stereo recordings `wdrums_liverec_5cm_mix.wav` and `wdrums_liverec_1m_mix.wav` from the Jan. 17, 2007 test data set to create a four-channel, 16kHz-sampled data set containing three sources – a male singer’s voice, an electric guitar, and a drum set. As the exact geometry of the sources and sensors is unknown in this data set – in fact, it is claimed that each file was generated using a *different* source geometry [7] – the ability of our algorithm to perform signal reconstruction on this data provides some indication of performance in an unknown setting.

Figure 2 on the top of the next page plots the average reconstruction signal-to-error-ratio

$$SER = \frac{1}{n} \sum_{j=1}^n \left(\frac{\sum_{k=L+1}^N |x_j(k-L)|^2}{\sum_{k=L+1}^N |x_j(k-L) - u_j(k)|^2} \right) \quad (23)$$

as a function of L over the range $L = 5$ to $L = 2000$ with $n = 4$ for four different cases: (a) one source at 0° incident angle with $m = 1$, (b) two sources at -30° and 27° incident angles with $m = 2$, (c) three sources at -30° , 0° , and 27° incident angles, respectively, with $m = 3$, and (d) using the contest data unaltered. Several points about these results are evident:

- For any given data set, the algorithm’s performance improves as the filter length is increased, as is to be expected.
- The algorithm’s ability to reconstruct multichannel data improves as the number of sources and reconstruction channels is increased. For a single source recorded using four microphones with $m = 1$, the best performance achieved in this setting is 21dB for $L = 2000$, whereas a three-source mixture recorded at four microphones with $m = 3$ can be reconstructed with nearly a 33dB average SER. Note that for a single source, performance improves as the value of m is increased in the reconstruction task.

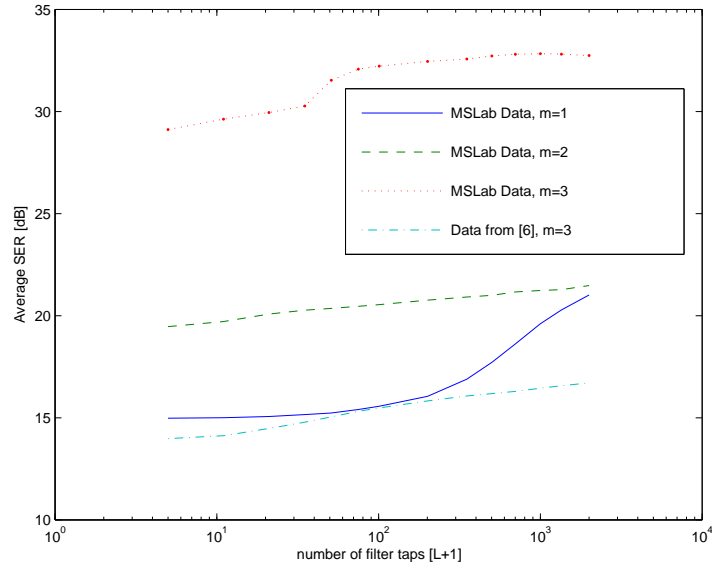


Figure 2: SER vs. reconstruction filter length ($L + 1$) for the various data sets in the examples.

- For the contest data, the reconstruction SER is between 14dB and 16dB for all filter lengths. Comparing the reconstructed signals $\{u_j(k)\}$ with the original recordings $\{x_j(k)\}$, the reconstructed signals have somewhat reduced high frequency energy as compared to the original signals.

Figure 3 shows the impulse responses generated from the algorithm for the MSLab data set containing two sound sources, where $L = 100$ has been chosen. Looking horizontally across the figure, each of the four impulse responses is used to generate two signals $y_1(k)$ and $y_2(k)$ from the original four-channel data by linear combination and convolution. Looking vertically along the figure, one observes the four pairs of time-reversed two-channel filters that are used to reconstruct the four original microphone signals. Because of the nearly-uniform spacing of the array, the algorithm adapts the system coefficients such that $y_1(k)$ and $y_2(k)$ are most-closely related to $x_1(k)$ and $x_4(k)$, whereas estimates of $x_2(k)$ and $x_3(k)$ are found to contain significant portions of both $y_1(k)$ and $y_2(k)$. This intuitive interpretation of the algorithm's behavior has little to do with the actual algorithm operation, as the procedure works for unstructured arrays and with sources at arbitrary positions in the room, as has been experimentally verified.

4. DETERMINING THE NUMBER OF SOURCES IN MULTICHANNEL AUDIO RECORDINGS

An important problem in multichannel audio applications is determining the number of sound sources currently active in multichannel recordings. Clearly, such a task is closely related to the signal reconstruction method described in this paper. In fact, one could consider developing a procedure for estimating the number of sound sources m in an n -channel data set by simply

1. finding all i -dimensional signal reconstructions of the data set for $i \in \{1, 2, \dots, n\}$, and
2. determining the value of i that accurately trades off reconstruction error for descriptive complexity.

The main drawback of such an approach is evident in Figure 2, in which it is observed that the SER can *increase* with the number of sources, making an absolute determination of the number of

sources from reconstruction error alone challenging. An alternative is to develop a procedure that finds the specific filters corresponding to an eigen-decomposition of the spatio-temporal autocorrelation matrix, which is the spatio-temporal equivalent of principal component analysis. Such methods are extremely important for successful solutions to the convolutive blind source separation problem [2], as many algorithms require knowledge of the number of sources in order to work properly. Efforts on developing such methods are ongoing.

5. CONCLUSIONS

In this paper, we have described a spatio-temporal extension of a well-known iterative method for finding principal signal subspaces from a symmetric autocorrelation matrix. Our procedure finds a paraunitary filter bank that can accurately encode a reduced-rank n -dimensional multichannel signal using m -dimensional signal representations, where $m < n$. The approach combines two iterative procedures in a nested way to perform the optimization. Application of the method to real-world multichannel audio recordings shows the ability of the method to perform this signal reconstruction task. Extensions of the method to rank determination of multichannel signal sets in audio contexts is ongoing.

6. REFERENCES

- [1] M. Brandstein and D. Ward, eds. *Microphone Arrays: Techniques and Applications* (New York: Springer, 2001).
- [2] S. Makino, T.-W. Lee, and H. Sawada, eds. *Blind Speech Processing* (New York: Springer, in press).
- [3] K.I. Diamantaras and S.Y. Kung, *Principal Component Neural Networks* (New York: Wiley, 1996).
- [4] S.C. Douglas, "An adaptive constraint method for paraunitary filter banks with applications to spatiotemporal subspace analysis," *EURASIP J. Advances in Signal Processing*, vol. 2007, Article ID 80301, 11 pages, 2007.
- [5] A. Tkacenko, P.P. Vaidyanathan, and T.Q. Nguyen, "On the eigenfilter design method and its applications: A tutorial," *IEEE Trans. Circuits Syst. II: Anal. Dig. Sig. Proc.*, vol. 50, pp. 497-517, Sept. 2003.
- [6] Data collected in March, 2007 from <http://sassec.gforge.inria.fr/>.
- [7] E. Vincent, personal communication.