REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
Sept 30, 1991	Final Technical Report Apr 1, 89 - 30 JUN 91

4. TITLE AND SUBTITLE New Algorithms for Broad-Band and Narrowband Source Localization and a Separable 2-D IIR Filter Realization (U)

5. FUNDING NUMBERS

61102F
2304/A6

6. AUTHOR(S)

Arnab Kumar Shaw

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Electrical Engineering Department
Wright State University
Dayton, OH-45431
Phone: (513)-873-3527

8. PERFORMING ORGANIZATION REPORT NUMBER

AFOSR-TR- 91 0917

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Department of the Air Force
Air Force Office of Scientific Research (AFSC)
Bolling Air Force Base, DC 20332-6448

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AFOSR-89-0291

11. SUPPLEMENTARY NOTES

No cost extension up to June 30, 91

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release;
distribution unlimited.

12b. DISTRIBUTION CODE

UL

13. ABSTRACT *(Maximum 200 words)*

(i) Optimal Design of ARMA (IIR) filters with arbitrary number of poles and zeros from Impulse Response Data has been developed. The general criterion derived in this report has never been found before. (ii) Optimal synthesis of 2-D IIR filters using 1-D modules have been developed. (iii) Optimal design of a class of 2-D IIR filters from spatial domain data has been developed. (iv) Optimal identification of Multivariable systems from Impulse response data is given. (v) A Periodogram-based Maximum-Likelihood estimator of Narrowband frequencies requiring only off-the-shelf hardware/software has been developed. (vi) A faster Simulated-Annealing method has been developed and applied to frequency estimation. (vii) A coherent one-step angles-of-arrival estimator of multiple broadband sources has been developed. Existing coherent techniques can not localize well separated sources in one step. (viii) An Order-Recursive approach has been given for AR-Bispectrum estimation. (ix) A Time-Delay-Neural-Network has been trained with LPC coefficients for Phoneme/Vowel recognition. (x) Parametric Non-linear prediction algorithms have been introduced for the first time for speech prediction/synthesis/coding.

14. SUBJECT TERMS Digital IIR Filter Design, IIR Filter Synthesis, System Identification, Multiple Frequency Estimation, Spectrum Estimation, Angle-of-Arrival Estimation of narrowband and Broadband Sources, Neural Networks based Speech Recognition.

15. NUMBER OF PAGES 125 pages

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	SAR

Final Technical Report of Grant No. : Grant AFOSR-89-0291

PERIOD : Apr 1, 89 - Mar 31, 91

(No cost extension up to June 30, 91)

PROPOSAL TITLE :

NEW ALGORITHMS FOR BROAD-BAND AND NARROWBAND SOURCE LOCALIZATION AND A SEPARABLE 2-D IIR FILTER REALIZATION

Principal Investigator : *Arnab Kumar Shaw*

Electrical Engineering Department

Wright State University

Dayton, OH-45431

Phone: (513)-873-3527 (Ext.- 2144)

Sponsoring Organization :

Department of the Air Force

Air Force Office of Scientific Research (AFSC)

Bolling Air Force Base, DC 20332-6448

Submission Date : Sept 30, 1991

A-1

TABLE OF CONTENTS

Technical Report of Grant No. : AFOSR-89-0291

PERIOD : Apr 1, 89 - Mar 31, 91

(No cost extension up to June 30, 91)

PROPOSAL TITLE :

NEW ALGORITHMS FOR BROAD-BAND AND NARROWBAND SOURCE LOCALIZATION AND A SEPARABLE 2-D IIR FILTER REALIZATION

Principal Investigator : *Arnab Kumar Shaw*

Electrical Engineering Department

Wright State University

Dayton, OH-45431

Phone: (513)-873-3527 (Ext.- 2144)

Sept 30, 1991

TABLE OF CONTENTS

CHAPTER 1

Introduction

Spectrum Estimation and Digital Filtering have continued to be two of the most important research areas among the signal processing community over the last three decades. As part of the work under this proposal, several research problems of current interest have been addressed and solved satisfactorily. This final report contains the details of all the results of the research that have been accomplished over the entire period. Much of the results contained in this report have either been presented/published or are under review/preparation for future publication. The papers/publications ensuing from this research are listed at the end of this introductory Chapter. Copies of the papers and publications will be included with the Invention Report.

The research conducted under this proposal can be categorized primarily into two broad themes, *viz.*,

(i) **Digital Filtering** : The following problems have been addressed :

- (a) Efficient synthesis of 2-D Digital Filters using 1-D modules
- (b) Optimal design of 1-D and 2-D Digital Filters from Impulse Response
- (c) Optimal Identification of Discrete-time Multivariable Systems from Impulse Response Matrix

(ii) **Spectrum Estimation** :

- (a) Development of efficient algorithms for estimation of the frequencies/arrival-angles of narrowband and wideband sources
- (b) Development of an Order-recursive algorithm for AR-Bispectrum Estimation
- (c) Application of newly emerging non-linear prediction methods for speech analysis and synthesis
- (d) Utilization of Linear Prediction parameters for training Time Delay Neural Networks for speech recognition

The report is organized as follows: In Chapter 2, the work on Digital Filtering is reported whereas in Chapter 3, the Spectral Estimation area is covered in detail. Individual Chapters are divided into several Sections by topics. In the following paragraphs the main results obtained in these Sections are outlined very briefly.

CHAPTER 2. DIGITAL FILTERS : REALIZATION AND SYNTHESIS

Section - 2.1 : Exact Realization of 2-D IIR Filters Using 1-D Modules

A method for exact realization of 2-dimensional digital IIR filters using separable 1-dimensional modules is presented [1]. The proposed design utilizes a theorem on separability of multivariable polynomials for writing a 2-D polynomial as sum-of-product of 1-D polynomials of successively diminishing orders. When compared to existing methods based on singular value decomposition (SVD) and Jordan form decomposition (JD), the proposed approach has reduced hardware complexity for filter implementation. It is also shown that the method has the same complexity as the lower-upper (LU) matrix decomposition based method. But unlike the decomposition based methods, the filter coefficients are found directly from the impulse response matrices by simple and numerically reliable mathematical operations.

It has also been shown that utilizing the inherent modularity in the way the 2-D polynomials has been rewritten, the complete 2-D transfer function can be built with a number of first and second order 1-D filter blocks. Hence, recent advances in VLSI methodologies can be utilized to facilitate mass production of 2-D filters.

Section - 2.2 : Optimal Identification of 1-D Discrete-Time Systems from Impulse Response Data

An optimal algorithm for estimation of the parameters of rational transfer functions from prescribed impulse response data is presented [2]. One of the major contributions of this part of the work is that, *for the first time*, an error minimization criterion has been theoretically derived which is uniformly applicable to rational models with *arbitrary numbers of poles and zeros*. This is a very important result because existing methods either modify the true non-linear error criterion in the theoretical derivation or require the transfer function model to have exactly one less number of zeros than poles. In the proposed algorithm, the transfer function coefficients are estimated by minimizing the ℓ_2-norm of the *exact* model fitting error. It is shown that the complete basis space orthogonal to the model fitting error can be constructed with the coefficients of the denominator polynomial only. The multidimensional non-linear error criterion is decoupled into a purely linear and a purely nonlinear subproblem. Global optimality properties of the decoupled estimators are established. The inherent mathematical structure in the non-linear subproblem is exploited in formulating an efficient iterative computational algorithm for its minimization. The proposed algorithm provides a powerful and comprehensive, theoretical as well as computational framework for modeling general pole-zero (ARMA) and all-pole (AR) systems from prescribed impulse response data. It is shown that the algorithm can be utilized for identifying all-zero (MA) systems also.

Before the general optimal algorithm mentioned above was discovered, we had developed some suboptimal designs for proper transfer functions with equal number of poles and zeros. This work is reported in [3] which will be included with the Invention report but these suboptimal results are not described in Section 2.2.

Section - 2.3 : Design of 2-D Recursive Digital Filters From Spatial Domain Data - Strictly Proper Case

A class of least-squares algorithms for design of two-dimensional digital filters from space domain data is presented [4, 15]. The proposed algorithms iteratively estimate the filter coefficients by minimizing the *true* squared error between the given and the estimated space domain responses. The algorithms are essentially generalization of an existing 1-D design algorithm given by Evans and Fischl (EFM) which is known to be optimal when the number of zeros in the transfer function is one less than the number of poles. Though some work extending EFM had earlier been reported, the full potential of EFM was never made use of because the true reparameterized error criterion was not derived and also the second phase of EFM was never evoked. Also, unlike the earlier methods, the error criterion is simultaneously optimized *w.r.t.* the coefficients in both dimensions. Design algorithms are given for filters with separable and irreducible numerator/denominator polynomials and also for mixed structures.

Section - 2.4 : Identification Of Discrete Time Multivariable Systems from Impulse Response Data

The problem of identification of transfer function matrices of discrete time multivariable systems is addressed [5]. The proposed technique obtains an optimal approximation from the given (possibly noisy) measured *impulse response data*. It is assumed that the measured impulse response data corresponds to a system with a strictly proper transfer function matrix with common denominators and different numerator polynomials. Based on the proposed theoretical basis, an efficient computational algorithm is developed and illustrated by means of several examples. In [16], we propose another algorithm that obtains a common numerator as well as a common denominator polynomial for all the elements of a transfer function matrix. This design essentially produces a common controller with different gains for several plants.

Realization of 2-D State-Space Filters With Fewer Multipliers

This work was published [6] during the proposal period though the major part of the work was completed

before the inception of this research. Hence only a brief summary is being included only in the Introduction and the paper will also be included with the Invention report. In this paper, it has been shown that under certain controllability and observability conditions on the 1-D block diagonal subsystems, a reduction in the number of multipliers for hardware realization can be achieved. Compared to a related existing method which requires $2nm + 3(n + m) + 1$ multipliers, the proposed transformation reduces the multiplier requirement by $(n + m)$, where n and m denote the respective dimensions of individual 1-D blocks. This saving in cost may be substantial if the filter order is high. A systematic procedure for obtaining the coefficients of the minimal number of multipliers is also given in the paper along with a detailed numerical example which illustrate the accuracy of the proposed method.

CHAPTER 3 . SPECTRUM ESTIMATION AND RELATED TOPICS

Section - 3.1 : A Cyclic Algorithm for Maximum Likelihood Frequency Estimation

A simple cyclic algorithm for estimation of multiple frequencies of narrow-band sources from noisy data is given [7]. The algorithm iteratively and recursively updates each unknown frequency by minimizing the model fitting error. For Gaussianly distributed noise, the algorithm produces maximum likelihood estimates, otherwise least-squares estimates are found. At each step of the algorithm, the optimization problem is *w.r.t.* a single frequency only and hence, simple hardware/software (*e.g.*, usage of FFT for the computation of periodogram) will be sufficient for implementation of the proposed cyclic algorithm.

Periodogram is one of the most commonly used spectrum estimation techniques. But it is well known that periodogram can not resolve closely spaced frequencies or angles of arrivals. The main goal of this research was to develop an algorithm that will rely on periodogram but at the same time provide high-resolution estimates by maximizing the maximum-likelihood criterion. The proposed cyclic approach achieves these goals because it requires optimization with respect to only one frequency at every estimation cycle. The method is iterative and recursive and relies on the knowledge of approximate prior estimates of the frequencies (or regions of interest) which may be easily obtained from the periodogram peaks. The estimates obtained using the algorithm are unbiased and follow the Cramer-Rao lower bound up to 0dB SNR.

Section - 3.2 : A Parameter Adaptive Simulated Annealing Algorithm Applied to Frequency Estimation

In this part of the research, a *faster* simulated annealing algorithm is proposed [8] and applied to the frequency estimation problem. The proposed annealing scheme is based on a cooling schedule which is parameter adaptive. In the existing annealing schemes, the temperature parameter is predetermined for every iteration step and is independent of the unknown parameter values. In the proposed scheme, the cooling temperature is made proportional to the deviation of each individual parameter at the earlier iteration step. The other key difference is that the proposed scheme never accepts a higher energy level and remains at the present lower energy position. Instead, the Boltzmann Distribution is used to accept a larger cooling temperature which is same as performing parameter search with a broader search space. This algorithm was applied to the non-linear maximum-likelihood error criterion that arise in frequency estimation and simulations confirm that the proposed scheme converges to the minimum energy level in much fewer iteration steps when compared to an existing fast annealing algorithm due to Szu.

Section - 3.3 : One Step Estimation of Angles-of-Arrival of Wideband Sources

A high resolution algorithm for estimating the angles of arrivals of multiple wideband sources is studied for

this part of the work [9]. The algorithm is effective for a dense and equally spaced array structure where a bilinear transformation is utilized in the frequency domain for combining the signal subspaces at different frequencies for coherent processing. When compared with existing coherent approaches, the algorithm is non-iterative in the sense that *all* the arrival angles can be estimated in only one step of the algorithm. Existing algorithms can only estimate the angles of a cluster of sources in a particular direction. The proposed algorithm, unlike the existing ones, does not need the knowledge of the initial estimates of the arrival angles. The work reported here is a variation of some earlier work by the author. Instead of using generalized eigendecomposition or matrix-pencil method, here we pre- or post-multiply the signal-subspace matrix with the noise matrix. This enables us to use regular eigendecomposition routine to estimate the source angles. It is also shown that it may be numerically more stable if the coherent combination is not focused in the center frequency the numerical value of which could be very large. The new focusing matrix given here allows to focus independent of the center frequency.

The original intention was to utilize structured matrix approximation approach to this problem. This task has been accomplished but the eigendecomposition based method given here performed considerably better.

Section - 3.4 : An Order-Recursive approach for Parametric Bispectrum Estimation

Order recursive computation of AR parameters from cumulants is given [10]. The Cumulant matrix arising in AR-Bispectrum estimation may not be either Toeplitz or symmetric. In such cases, it is shown that using a block matrix inversion formula due to Frobenius and Schur, the inverse of the p-dimensional cumulant matrix can be updated from the $(p-1)$-dimensional inverse with $O(p^2)$ operations. When compared to commonly used standard batch-mode computation, the proposed algorithm reduces the computational requirement for order-recursive calculation of the AR-parameters. When the cumulant matrix is non-symmetric Toeplitz also, further reduction in computation is obtained using an algorithm due to Trench.

Section - 3.5 : Phoneme and Vowel Recognition Using Time-Delay Neural Network

In this part of research, Time-Delay Neural Network (TDNN) architecture has been used for speaker independent recognition of Phonemes and Vowels of isolated words [11, 12]. One of the limiting factors of many existing speech recognition algorithms is the requirement of precise temporal alignment. Segmentation and dynamic time-warping are usually performed to solve this alignment problem. Dynamic time-warping at each speech segment is computation intensive. Also, segmentation may be erroneous in itself and, when in error, will cause recognition failure due to a mismatch. Furthermore, it is advantageous of recognition algorithm to look at multiple time frames at one time in order to make use of the inter-frame relationships and differences in the input features. TDNN has the ability to represent relationships between events in time and at the same time it allows for invariance of these events under translation in time. With this translation invariance, a Time Delay Neural network does not require precise temporal alignment, therefore the network is able to simply scan the input features for clues. This is a necessary requirement for efficient continuous speech recognition addressed in this work.

Our goal was to improve the performance of TDNN by increasing the amount of data supplied to the network. This was achieved by including the LPC coefficients along with the FFT bin energies. We have also trained the network with utterances having variable durations. We feel that it is an important aspect due to the extreme variability in speaking rate of different speakers at different situations. Also, restricting the utterance length to 150ms (as it seems to have been proposed by Waibel *et al*) in order to make a decision about a vowel may limit the network's performance because depending on speaking style and the spoken word the selected 150ms may not contain all the key information to recognize the vowel. We have trained the network with multiple English speakers and have obtained 100% recognition rate.

Section - 3.6 : Speech Analysis and Synthesis with Non-Linear Prediction

One of the common assumptions in speech has been that speech production and perception are essentially linear processes and hence one can accurately model speech data using 'linear prediction' based methods. Physiological evidence indicate that some nonlinear operation does occur in speech production and perception. It is also known that linear models perform poorly for certain types of speech. Based on these observations, for this final part of research we consider the applicability of certain *parametric* nonlinear models for the purpose of analysis/prediction/synthesis/coding of speech signals [13, 14]. To the best of our knowledge, these models have not yet been exploited for speech modeling. Several algorithms for simultaneous estimation of the *nonlinear* as well as the *linear* prediction parameters of speech signals have been studied and more work is under way. These studies indicate that the nonlinear models retain substantially more information when compared to linear-only models. Experiments on telephone quality speech data clearly and consistently indicate that there is a significant reduction in the prediction error when the *bilinear* prediction components are included along with the LPC part. The results in this work may have significant effect on the performance accuracy of any speech recognition/synthesis/coding system that currently relies on linear prediction only.

Publications and Manuscripts under review/preparation during the Proposal Period

1) A. K. Shaw and P. Misra, "An Exact Realization of 2-D IIR Filters Using Separable 1-D Modules," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1293-1296, Albuquerque, New Mexico, April, 1990.

2) A. K. Shaw, "Optimal Identification of Discrete-Time Systems from Impulse Response Data," under review, *IEEE Transaction on Signal Processing*. A brief version, "An Optimal Method for Identification of Pole-Zero Transfer Functions,", under review, *International Symposium on Circuits and Systems*, San Diego, 1992.

3) A. K. Shaw and P. Misra, "Time Domain Identification of Proper Discrete Systems from Measured Impulse Response Data," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, May, 1991.

4) A. K. Shaw, P. Misra and T. Manickam, "Globally Optimum Design of 2-D IIR Filters in Spatial Domain," *International Symposium on Circuits and Systems*, New Orleans, May, 1990.

5) A. K. Shaw, P. Misra and R. Kumaresan, "Identification of Multi-Dimensional Systems From Impulse Response Data," accepted as a full paper, *30th IEEE Conf. on Dec. and Contr.*, Brighton, UK, Dec. 1991. Expanded version, "Multi-Dimensional System Identification From Impulse Response Data," under review, *IEEE Transactions on Aerospace and Electronic Systems.*

6) P. Misra and A. K. Shaw, "Realization of 2-D State Space Filters with Fewer Multipliers", *IEEE Transactions on Circuits and Systems*, pp. 252-256, Jan. 1990. Also in *Asilomar Conference*, California, Sept.,1989.

7) A. K. Shaw, "Maximum Likelihood Frequency Estimation : A Cyclic Algorithm," *IEEE International Conference on Systems Engineering*, Dayton, Ohio, Aug., 1991. Also under Preparation for *IEEE Trans. on Signal Processing.*

8) A. K. Shaw and T. Manickam, "A Parameter Adaptive Fast Simulated Annealing Algorithm," *IEEE International Conference on Systems Engineering*, pp. 213-216, Dayton, Ohio, Aug., 1989.

9) A. K. Shaw, "One-Step Estimation of Angles-of-Arrival of Wideband Sources," under preparation for *IEEE Trans. on Signal Processing.*

10) A. K. Shaw, "Order Recursive Parametric Bispectrum Estimation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, May, 1991.

11) A. K. Shaw and R. Mitchell, "Phoneme Recognition using Time Delay Neural Networks," *International Conference on Neural Networks,* vol.-2, pp. 191-195, San Diego, June, 1990.

12) R. Mitchell and A. K. Shaw, "Vowel Recognition using Time Delay Neural Networks," (with R. Mitchell) *IEEE International Conference on Systems Engineering*, Dayton, Ohio, Aug., 1990.

13) A. K. Shaw and S. Kundu, "Parametric Non-linear Prediction of Speech," (with S. Kundu), under review, *IEEE Automatic Speech Recognition Workshop*, New York, 1991.

14) A. K. Shaw and S. Kundu, "Speech Analysis and Synthesis with Non-linear Prediction of Speech," under review, *IEEE International Conference on Acoustics, Speech and Signal Processing*, San Francisco, California, April, 1992.

15) A. K. Shaw and P. Misra, "Optimal 2-D IIR Filters," under review, *IEEE International Conference on Acoustics, Speech and Signal Processing*, San Francisco, California, 1992.

16) A. K. Shaw and P. Misra, "Optimal Representation of Several Plant Transfer Functions by an Equivalent Single Transfer Function," under review, *IEEE International Conference on Acoustics, Speech and Signal Processing*, San Francisco, California, 1992.

CHAPTER 2

DIGITAL FILTERS : REALIZATION AND SYNTHESIS

SECTION 2.1 : SEPARABLE AND EXACT REALIZATION OF 2-D IIR FILTERS

SUMMARY

A method for exact realization of 2-dimensional digital IIR filters using separable 1-dimensional modules is presented. The proposed design utilizes a theorem on separability of multivariable polynomials for writing a 2-D polynomial as sum-of-product of 1-D polynomials of successively diminishing orders. The proposed realization has lower complexity compared to most existing methods.

I. INTRODUCTION

Exact realization of 2-D IIR filters using separable 1-D filter is one of the most elusive problems in digital signal processing. In development of stability tests and for stabilization of 1-D digital filters, factorization plays an important role. However, it is well recognized that no fundamental theorem of algebra on factorization exists for polynomials in two independent variables. This lack of a corresponding theorem has been aptly phrased as "a fundamental curse" of 2-D filtering. In the last two decades considerable research effort has been devoted to the development of algorithms for 2-D polynomial factorization. In [1], Treitel and Shanks proposed a scheme for *approximate* implementation of 2-D *planar* (2-D FIR) filters with $m \times n$ coefficients in terms of k number of separable blocks each requiring $(m+n)$ coefficients. In [2],[3] a separability theorem of multivariable polynomial is presented and using the separability result for the 2-D case, Suresh and Shenoi [4] presented an *exact* realization of 2-D planar filters by separable 1-D filters. In [5], Venetsanopoulos and Mertzios used a general matrix decomposition theorem for exact decomposition of a general 2-D real rational transfer function. It was shown that in the decomposed form, each rational function depends on only one of the two independent variables. Nikias *et al* [6] showed that LU decomposition can be used for exact implementation of 2-D rational transfer functions. Moreover, from the hardware point of view, the LU decomposition is considerably more efficient compared to the realizations based on Singular Value Decomposition (SVD) and Jordan form Decomposition (JD) [7]. For more references on 2-D filter implementation, see [5] and [6].

In this work, we utilize the separability theorem of [2],[3] to develop exact realizations of 2-D digital filters with rational transfer function using separable and parallel blocks of 1-D filters of successively reducing orders. For the planar case, the proposed approach is essentially similar to the one given in [4], except that our implementation requires reduced number of delay elements. The proposed approach for the planar case is then extended for implementation of rational transfer functions (2-D IIR filters) by incorporating feedback and cascading. For the full rank planar coefficient matrix case, the given approach has reduced hardware requirement than SVD and JD approaches and same complexity as LU decomposition approach.

In the proposed scheme, the filter coefficients are found by simple operations of matrix addition, subtraction and multiplication. For rank deficient coefficient matrices also, the proposed scheme has reduced hardware complexity than the SVD and JD approaches. Similar to the matrix decomposition based methods [5],[6], the proposed approach also possesses high degree of modularity and modern VLSI technology can be utilized for efficient hardware realization. The results presented in the sequel are a more formal and detailed presentation of [8].

This Section is arranged as follows. In Subsection II, the problem is formulated and the planar full rank case is first treated by considering the numerator of the 2-D rational transfer function. In Subsection III, the separable

IIR design is considered. The rank deficient case is briefly outlined in Subsection IV. Finally, in Subsection V, an examples is given to illustrate the proposed technique.

II. Separable Realization of 2-D FIR Filter

A 2-D rational transfer function has the following general form,

$$H(z_1, z_2) = \frac{q(z_1, z_2)}{p(z_1, z_2)} = \frac{\sum_{i=0}^{n_1-1}\sum_{j=0}^{m_1-1} q(i,j) z_1^{-i} z_2^{-j}}{\sum_{i=0}^{n_2-1}\sum_{j=0}^{m_2-1} p(i,j) z_1^{-i} z_2^{-j}}. \tag{2.1}$$

Let us consider the numerator polynomial $q(z_1, z_2)$ first. Its impulse response can be represented by the following $n_1 \times m_1$ matrix

$$\mathbf{Q} \triangleq \begin{bmatrix} q(n_1-1, m_1-1) & q(n_1-1, m_1-2) & \dots & q(n_1-1, 0) \\ \vdots & \vdots & \ddots & \vdots \\ q(1, m_1-1) & q(1, m_1-2) & \dots & q(1, 0) \\ q(0, m_1-1) & q(0, m_1-2) & \dots & q(0, 0) \end{bmatrix} \tag{2.2}$$

such that the polynomial $q(z_1, z_2)$ can be written as

$$q(z_1, z_2) \triangleq \mathbf{z}_1^T \mathbf{Q} \mathbf{z}_2, \tag{2.3}$$

where $\mathbf{z}_1$ and $\mathbf{z}_2$ are defined as,

$$\mathbf{z}_1 \triangleq [z_1^{-(n_1-1)} \; z_1^{-(n_1-2)} \; \dots z_1^{-1} \; 1]^T \quad \text{and} \quad \mathbf{z}_2 \triangleq [z_2^{-(m_1-1)} \; z_2^{-(m_2-2)} \; \dots z_2^{-1} \; 1]^T, \tag{2.4}$$

and $[\cdot]^T$ denotes the transpose operation. We first consider the case of $\mathbf{Q}$ being a full rank matrix.

In general, the coefficient matrix $\mathbf{Q}$ is not separable, *i.e*, one cannot, in general, factor $q(z_1, z_2)$ in terms of 1-D z-domain polynomials. But following the separability testing criterion [2],[3], the matrix $\mathbf{Q}$ can be expressed as the sum of a separable matrix $\mathbf{Q}_0$ and an error matrix $\mathbf{E}_0$ which again, in general, is not separable, *i.e.*, [4]

$$q(i,j) = q_0(i,j) + e_0(i,j) \tag{2.5}$$

where,

$$q_0(n_1-1, j) = q(n_1-1, j), \; 0 \le j \le m_1 - 1 \tag{2.6a}$$

$$q_0(i, m_1-1) = q(i, m_1-1), \; 0 \le i \le n_1 - 1 \tag{2.6b}$$

$$q_0(i,j) = q(i, m_1-1) q(n_1-1, j), \quad 1 \le j \le m_1-1, \; 1 \le i \le n_1-1. \tag{2.6c}$$

The indices of the matrix $\mathbf{Q}_0$ match those of $\mathbf{Q}$ as in (2.2). Note that according to the separability criterion [2],[3] any $(n_1 \times m_1)$ 2-D planar filter $\mathbf{G}$ is separable if

$$g(i,0) g(0,j) = g(i,j) \qquad 0 \le i \le n_1 - 1, \; 0 \le j \le m_1 - 1. \tag{2.7}$$

Since the elements $q_0(i,j)$'s follow (2.7), the 2-D filter having impulse response matrix $\mathbf{Q}_0$ is separable by construction. Let the separable form of the transfer function be expressed as

$$q_0(z_1, z_2) = \sum_{i=0}^{n_1-1}\sum_{j=0}^{m_1-1} q_0(i,j) z_1^{-i} z_2^{-j} = c_0^n(z_1) r_0^n(z_2), \tag{2.8}$$

where,

$$c_0^n(z_1) \triangleq \sum_{i=0}^{n_1-1} c_0^n(i) z_1^{-i} \quad \text{and} \quad r_0^n(z_2) \triangleq \sum_{j=0}^{m_1-1} r_0^n(j) z_2^{-j}. \tag{2.9}$$

The coefficients $c_0^n(i)$ for $i = 0, \ldots, n_1 - 1$ and $r_0^n(j)$ for $j = 0, \ldots, m_1 - 1$ are constructed as,

$$c_0^n(n_1 - 1) = 1, \quad \text{and} \quad c_0^n(i) = q(i, m_1 - 1), \text{ for } i = 0, \ldots, n_1 - 2 \tag{2.10a}$$

and

$$r_0^n(m_1 - 1) = 1, \quad \text{and} \quad r_0^n(i) = q(i, n_1 - 1), \text{ for } j = 0, \ldots, m_1 - 2, \tag{2.10b}$$

where the superscript n indicates that the decomposition is for the denominator polynomial. Note that $q(n_1 - 1, m_1 - 1)$ is assumed to be non-zero without any loss of generality. If it happens to be zero then the first column (row) of $\mathbf{Q}$ may be interchanged with any column (row) with non-zero leading coefficient and at the same time interchanging the corresponding powers of z_2 (z_1) in the vector $\mathbf{z}_2$ (z_1) . Further, the matrix $\mathbf{Q}$ can be normalized such that the (1,1) element is equal to 1. In (2.10) above, we have assumed that the necessary column (row) permutation and/or normalization has already been performed.

The error matrix $\mathbf{E}_0 \in \mathbf{R}^{n_1 \times m_1}$, formed with the elements of $e_0(i,j)$ in (2.5) has the following form,

$$\mathbf{E}_0 \triangleq \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & e_0(n_1 - 2, m_1 - 2) & \ldots & e_0(n_1 - 2, 0) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & e_0(0, m_1 - 2) & \ldots & e_0(0, 0) \end{bmatrix}. \tag{2.11}$$

Now, let us call the $n_1 - 1 \times m_1 - 1$ non-zero submatrix at the lower right hand corner of $\mathbf{E}_0$ as $\tilde{\mathbf{E}}_0$ and let $\tilde{e}_0(z_1, z_2)$ be the system function corresponding to it. Then,

$$\tilde{e}_0(z_1, z_2) = \sum_{i=0}^{n_1-2} \sum_{j=0}^{m_1-2} z_1^{-i} z_2^{-j} \tilde{e}_0(i,j). \tag{2.12}$$

Hence, combining (2.8) and (2.12), the numerator can now be written as

$$q(z_1, z_2) = c_0^n(z_1) r_0^n(z_2) + \tilde{e}_0(z_1, z_2). \tag{2.13}$$

Now starting from the matrix $\tilde{\mathbf{E}}_0$ which, in general, is non-separable, we can proceed similarly as we did for the matrix $\mathbf{Q}$ and form another summation of a separable and a possibly non-separable filter, *i.e.*, similar to (2.8) and (2.13), we can express $\tilde{e}_0(z_1, z_2)$ as,

$$\tilde{e}_0(z_1, z_2) = q_1(z_1, z_2) + \tilde{e}_1(z_1, z_2) = c_1^n(z_1) r_1^n(z_2) + \tilde{e}_1(z_1, z_2), \tag{2.14}$$

where, $c_1^n(z_1)$ and $r_1^n(z_2)$ are defined similarly as in (2.9) and (2.10) except that their orders are reduced by one. Continuing that process, eventually we will get,

$$q(z_1, z_2) = \sum_{i=0}^{m_1-1} q_i(z_1, z_2) = \sum_{i=0}^{m_1-1} c_i^n(z_1) r_i^n(z_2). \tag{2.15}$$

This is the separable form we were seeking for the numerator polynomial. Note that we have assumed $n_1 \geq m_1$ (= rank of $\mathbf{Q}$), without any loss of generality. Also note that $c_i^n(z_1)$ and $r_i^n(z_2)$ are 1-D filters with orders $n_1 - 1 - i$ and $m_1 - 1 - i$, respectively, *i.e.*, the filter orders are successively reducing with increasing values of i. The

separable form of $q(z_1, z_2)$ in equation (2.15) implies that the planar 2-D filter $q(z_1, z_2)$ can be implemented with parallel blocks of separable 1-D filters. The modularity of this decomposition, for the purpose of implementation is obvious. Note also that in SVD and JD approaches [5], each parallel block would usually have two cascaded 1-D filters of orders $n_1 - 1$ and $m_1 - 1$, respectively, whereas in the proposed approach, the orders of the 1-D filters in successive parallel branches diminish, thereby reducing the complexity and hardware requirement in the filter implementation. Specifically, SVD and JD approaches would require $m_1(m_1 + n_1)$ coefficients whereas the present approach would require $m_1(n_1 + 1)$ coefficients which is the same as that of the requirement for the LU decomposition case. Hence the hardware requirement for LU decomposition approach [6] is exactly same as in the present case.

It may be pointed out here that the above derivation is essentially similar to the one given in [4], except that they considered only the planar case and started with a permuted form of the **Q** matrix. Specifically, in [4], the planar matrix is written as

$$\hat{\mathbf{Q}} \triangleq \begin{bmatrix} q(0,0) & q(0,1) & \dots & q(0,m_1-1) \\ q(1,0) & q(1,1) & \dots & q(1,m_1-1) \\ \vdots & \vdots & \ddots & \vdots \\ q(n_1-1,0) & q(n_1-1,1) & \dots & q(n_1-1,m_1-1) \end{bmatrix}. \tag{2.16}$$

such that the polynomial $q(z_1, z_2)$ may be expressed as

$$q(z_1, z_2) \triangleq \hat{\mathbf{z}}_1^T \hat{\mathbf{Q}} \hat{\mathbf{z}}_2, \tag{2.17}$$

where $\hat{\mathbf{z}}_1$ and $\hat{\mathbf{z}}_2$ are defined as,

$$\hat{\mathbf{z}}_1 \triangleq [1 \; z_1^{-1} \; z_1^{-2} \; \dots \; z_1^{-(n_1-1)}]^T \quad \text{and} \quad \hat{\mathbf{z}}_2 \triangleq [1 \; z_2^{-1} \; z_2^{-2} \; \dots \; z_2^{-(m_1-1)}]^T. \tag{2.18}$$

Eventually, the separable form equivalent to (2.15) was found to be,

$$q(z_1, z_2) = \sum_{i=0}^{m_1-1} \hat{q}_i(z_1, z_2) z_1^{-i} z_2^{-i} = \sum_{i=0}^{m_1-1} z_1^{-i} z_2^{-i} \hat{c}_i(z_1) \hat{r}_i(z_2). \tag{2.19}$$

For the realization in (2.15), the highest powers of delays are accounted for in the first recursion and subsequent recursion have lower powers of delay. However, the realization (2.19) retains the highest powers of delay till the last recursion. Therefore, the realization in (2.19) would require $2m_1$ extra delays when compared to the one in (2.15). This is evident from the $z_1^{-i} z_2^{-i}$ term in (2.19).

The separable form of the numerator polynomial is now complete and the extension to the denominator realization is given next.

III. Separable Implementation of 2-D IIR Filter

Let us first write the denominator polynomial as

$$p(z_1, z_2) = K + f(z_1, z_2) \tag{3.1}$$

where the polynomial $f(z_1, z_2)$ has no constant term. Next the all-pole part of $H(z_1, z_2)$ in (2.1) is rewritten as,

$$H_p(z_1, z_2) = \frac{1}{p(z_1, z_2)} = \frac{1}{K + f(z_1, z_2)} = \frac{\frac{1}{K}}{1 + \frac{1}{K} f(z_1, z_2)} \tag{3.2}$$

which is a feedback network with $\frac{1}{K}$ in the forward branch and $f(z_1, z_2)$ polynomial in the feedback branch as shown in Fig. 1. Our contention here is that $f(z_1, z_2)$ which, in general, is not separable, can again be expressed as a sum of separable polynomials following the same steps used in obtaining (2.15) *i.e.*,

$$f(z_1, z_2) = \sum_{i=0}^{m_2-1} c_i^d(z_1) r_i^d(z_2), \tag{3.3}$$

where d indicates that the decomposition is for the denominator polynomial. Next, the contributions from the numerator and the denominator can be cascaded as shown in Fig. 2 for an exact realization of the 2-D rational transfer function in terms of only separable 1-D modules. For realizable decomposition, it should be ensured that there must not be any delay-free loops in the feedback path. This, in turn, can be ensured by selecting $c_i^d(z_1)$ and $r_i^d(z_2)$ such that for any value of i, both $c_i^d(z_1)$ and $r_i^d(z_2)$ do not have a constant term. This is discussed later in Subsection V.

IV. The Rank Deficient Case

If the coefficient matrix $\mathbf{Q}$ is not full rank, then SVD and JD approaches would require $k(m_1 + n_1)$ coefficients for the implementation, where k is the rank of $\mathbf{Q}$. Whereas the LU decomposition would require $k(n_1 + m_1 - k + 1)$ coefficients. However the proposed direct approach would still require $m_1(n_1 + 1)$ coefficients unless at some stage $\mathbf{E}_i$ is a zero matrix. Hence, the direct implementation of the proposed method may not always be economical when compared to SVD or JD approaches and it will almost always have more hardware requirement when compared to the LU approach. Hence, it is recommended that elementary row and column operations be performed on the $\mathbf{Q}$ matrix so that its first k principal minors are non-zero [6]. If $\mathbf{Q}$ can be reduced to LU decomposable form after the row/column operations, then we can still apply the proposed algorithm. Note that in this case $\mathbf{E}_{k-1}$ so obtained will be a zero matrix. Therefore, the number of coefficients will again be $k(n_1 + m_1 - k + 1)$ *i.e.*, it will have lower hardware requirement than SVD and JD approaches and the same hardware complexity as the LU approach. It should be pointed out that the main advantage of the proposed algorithm over the LU approach is that it produces the filter of same complexity as the LU decomposition without resorting to LU decomposition of $\mathbf{Q}$.

V. An Example

In this Subsection, we will illustrate the proposed scheme by means of a numerical example. The transfer function used for this example has been taken form [9]. $H(z) = \frac{q(z_1,z_2)}{1+f(z_1,z_2)/K}$ with

$$\begin{aligned} q(z_1, z_2) &= 3z_1^{-2}z_2^{-2} + 7z_1^{-1}z_2^{-2} + 2z_2^{-2} + 7z_1^{-2}z_2^{-1} + 9z_1^{-1}z_2^{-1} + 3z_2^{-1} + 3z_1^{-2} + 3z_1^{-1} + 1 \\ f(z_1, z_2) &= 3z_1^{-1} + z_1^{-2} + 4z_2^{-1} + 6z_1^{-1}z_2^{-1} + 2z_1^{-2}z_2^{-1} + 3z_2^{-2} + 3z_1^{-1}z_2^{-2} + z_1^{-2}z_2^{-2}. \end{aligned}$$

Following the steps outlined in the previous Subsection, we can write

$$q(z_1, z_2) = \mathbf{z}_1^T \begin{bmatrix} 3 & 7 & 3 \\ 7 & 9 & 3 \\ 2 & 3 & 1 \end{bmatrix} \mathbf{z}_2 \triangleq \mathbf{z}_1^T \mathbf{Q} \mathbf{z}_2$$

$$\begin{aligned} \mathbf{Q} &= \mathbf{Q}_0 + \mathbf{E}_0 \\ &= 3 \begin{bmatrix} 1 & 7/3 & 1 \\ 7/3 & 49/9 & 7/3 \\ 2/3 & 14/9 & 2/3 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & -22/3 & -4 \\ 0 & -5/3 & -1 \end{bmatrix}. \end{aligned}$$

Defining $\tilde{\mathbf{E}}_0 = \begin{bmatrix} -22/3 & -4 \\ -5/3 & -1 \end{bmatrix}$, we can write the numerator polynomial as

$$q(z_1, z_2) = \left(z_1^{-2} + 7/3z_1^{-1} + 2/3\right)\left(3z_2^{-2} + 7z_2^{-1} + 3\right) + \tilde{e}_0(z_1^{-1}, z_2^{-1})$$

where $\tilde{e}_0(z_1, z_2) = -(22/3z_1^{-1}z_2^{-1} + 4z_1^{-1} + 5/3z_2^{-1} + 1$. Following the same steps on $\tilde{\mathbf{E}}_i$, it can be shown that

$$q(z_1, z_2) = \left(z_1^{-2} + 7/3z_1^{-1} + 2/3\right)\left(3z_2^{-2} + 7z_2^{-1} + 3\right) - \left(z_1^{-1} + 5/22\right)\left(22/3z_2^{-1} + 4\right) - 1/11$$

Next, we perform a similar reduction on the denominator. The polynomial $f(z_1, z_2)$ can be written in the matrix vector form as

$$f(z_1, z_2) = \mathbf{z}_1^T \begin{bmatrix} 1 & 2 & 1 \\ 3 & 6 & 3 \\ 3 & 4 & 0 \end{bmatrix} \mathbf{z}_2 \triangleq \mathbf{z}_1^T \mathbf{F} \mathbf{z}_2. \tag{4.1}$$

Note that due to realizability constraints [6],[10], we cannot have any delay-free loops in the feedback path. Accordingly, we must make some minor modifications to the method used for realization of the numerator polynomial. In particular, it is necessary to permute the rows (columns) of the matrix $\mathbf{F}$ such that the resulting realization obtained by applying the proposed method does not contain any delay-free loop. It should be pointed out that such realizations always exist provided that the element $p(0,0)$ in (2.1) is non-zero [6].

Applying the necessary permutation (interchanging the first and the third columns of the matrix $\mathbf{F}$), we can rewrite (4.1) as

$$f(z_1, z_2) = \begin{bmatrix} z_1^{-2} & z_1^{-1} & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 3 & 6 & 3 \\ 0 & 4 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ z_2^{-1} \\ z_2^{-2} \end{bmatrix} \tag{4.2}$$

Next, using the proposed technique, we can obtain a realizable form for $f(z_1, z_2)$ given by

$$f(z_1, z_2) = \left(3z_1^{-1} + z_1^{-2}\right)\left(1 + z_2^{-1}\right)^2 + \left(4z_2^{-1} + 3z_2^{-2}\right).$$

Note that for this example several other realizable forms do exist. However, the above form requires the least number of delay elements.

Next, we compare the hardware requirement for the proposed realization with several existing ones.

Table 4.1: Comparison of Hardware Requirements (delay elements)

Method	Numerator	Denominator	Total
LU [6]	8	6	14
SVD [5]	18	8	26
JD [5]	18	6	26
SS [4]	14	*	*
Proposed	8	6	14

* Note that the method in [4] is applicable to FIR filters only.

VI. Discussion

The modularity in $q(z_1, z_2)$ can be observed in (2.15) which can be built with separate 1-D modules $c_i^n(z_1)$'s and $r_i^n(z_2)$'s only. The same will be true for the feedback path to implement $f(z_1, z_2)$. The individual modules $c_i^n(z_1)$'s and $r_i^n(z_2)$'s can again be factored into first and second (to have real coefficients only) order polynomials because they are functions of single independent variables only. This implies that the complete 2-D transfer function can

be built with a number of first and second order 1-D blocks. Recent advances in VLSI methodologies can be utilized for building such 2-D filters.

REFERENCES

[1] S. Treitel and J. Shanks, "The Design of Multistage Separable Planar Filters," *IEEE Transactions on Geoscience Electronics*, vol. GE-9, pp. 10-27, 1971.

[2] M. Lal, "On The Separability of Multivariable Polynomials," *Proceeding of the IEEE*, vol. 63, pp. 718-719, 1975.

[3] P. Misra and R.V. Patel, "A Necessary and Sufficient Condition for Decomposability of 2-Dimensional Polynomials", *IEEE Int. Conf. on Circuits and Systems*, New Orleans, LA, May 1990.

[4] B. R. Suresh and B. A. Shenoi, "Exact Realization of 2-Dimensional Digital Filters by Separable Filters," *Electronic Letters*, pp. 242-244, 1976.

[5] A. N. Venetsanopoulos and B. G. Mertzios, "A Decomposition Theorem and its Implications to the Design and Realization of Two-Dimensional Filters," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-33, pp. 1562-1575, 1985.

[6] C. L. Nikias, A. P. Chrysafis and A. N. Venetsanopoulos, "The LU Decomposition Theorem and its Implications to the Realization of Two-Dimensional Digital Filters," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-33, pp. 694-711, 1985.

[7] F. R. Gantmacher, *The Theory of Matrices*, Vol. 1, New York, Chelsa.

[8] A. K. Shaw and P. Misra, "An Exact Realization of 2-D Filters using Separable 1-D Modules," *Proceedings of IEEE Conf. on Acoust. Speech and Sig. Proc.*, Albuquerque, NM, April, 1990.

[9] S. K. Mitra, A. D. Sagar and N. A. Pendergrass, "Realization of Two-Dimensional Recursive Filters", *IEEE Transactions of Circuits and Systems*, vol. CAS-22, pp. 177-184, 1975.

[10] L. T. Bruton, "Low-Sensitivity Digital Ladder Filters", *IEEE Transactions of Circuits and Systems*, vol. CAS-22, pp. 171, 1975.

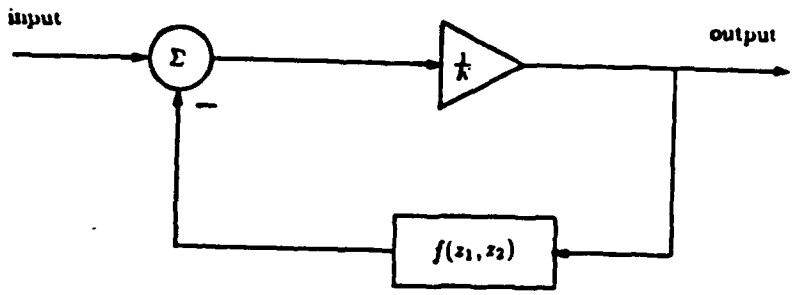

FIGURE 1: Realization of $H(z_1, z_2) = \frac{1}{p(z_1, z_2)}$

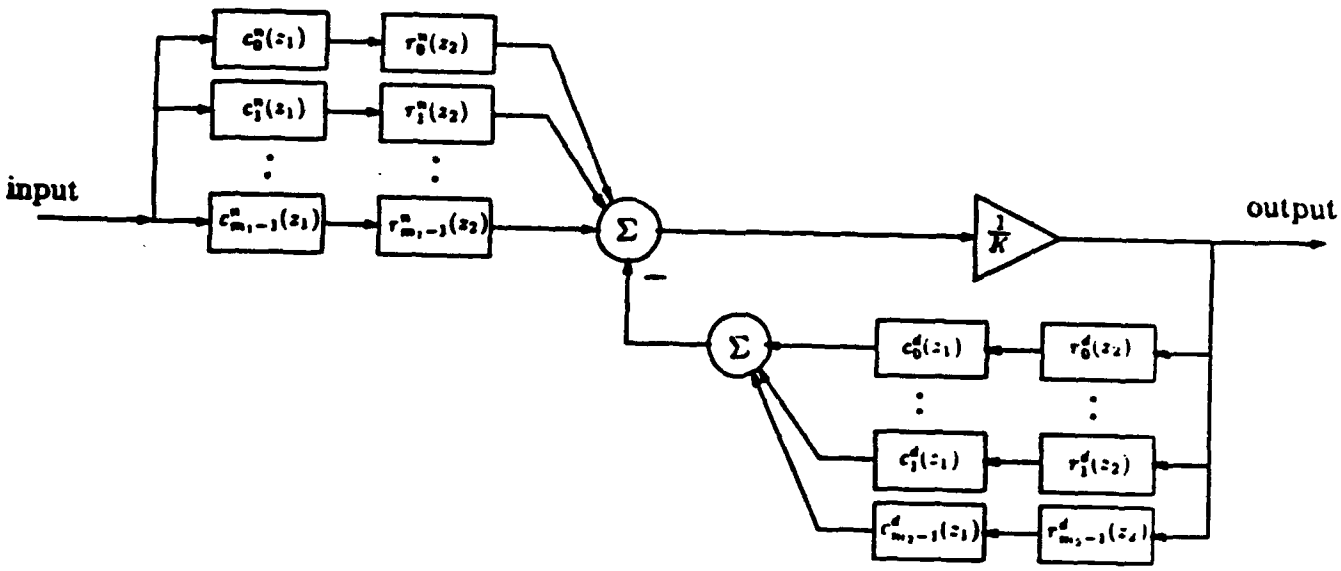

FIGURE 2: Realization of $H(z_1, z_2) = \frac{g(z_1, z_2)}{p(z_1, z_2)}$

SECTION 2.2 : OPTIMAL IDENTIFICATION OF DISCRETE-TIME SYSTEMS FROM IMPULSE RESPONSE DATA

SUMMARY

An optimal algorithm for estimation of the parameters of rational transfer functions from prescribed impulse response data is presented. A major contribution of this work is that, for the first time, an error minimization criterion has been theoretically derived which is uniformly applicable to rational models with arbitrary numbers of poles and zeros. Existing methods either modify the true non-linear error criterion in the theoretical derivation or require the transfer function model to have exactly one less number of zeros than poles. In the proposed algorithm, the transfer function coefficients are estimated by minimizing the ℓ_2-norm of the *exact* model fitting error. It is shown that the complete basis space orthogonal to the model fitting error can be constructed with the coefficients of the denominator polynomial only. The multidimensional non-linear error criterion is decoupled into a purely linear and a purely nonlinear subproblem. Global optimality properties of the decoupled estimators are established. The inherent mathematical structure in the non-linear subproblem is exploited in formulating an efficient iterative computational algorithm for its minimization. The proposed algorithm provides a powerful and comprehensive, theoretical as well as computational framework for modeling general pole-zero and all-pole systems from prescribed impulse response data. It is shown that the algorithm can be utilized for identifying all-zero systems also. The effectiveness of the algorithm is demonstrated with several simulation examples.

1. INTRODUCTION

Identification of unknown discrete-time linear systems is a fundamental problem in signal processing. Over the last several decades this problem has remained among the most active and important research areas [1-7, 10, 13-17, 24-34, 37, 55]. Depending on the application, design specifications or the information available about the unknown system, linear system modeling can be broadly categorized into two classes, namely, non-parametric and parametric. Historically, non-parametric modeling had gained early popularity due mainly to relatively simpler mathematical and computational complexity. Classical non-parametric models include impulse response sequence, Fourier-domain representation or frequency response, Power Spectral Density or Periodogram, autocorrelation sequence, Characteristic Function and others [8-12, 57, 58]. The major drawback of classical non-parametric models is that, more often than not, these representations are theoretically infinite in extent. Hence, though sound and simple in theory, these models may encounter serious data-handling problems in practice. Furthermore, in various non-parametric representations, the utilization of finite amount of data may cause limited resolution, spurious estimates, overshoot/oscillation (*e.g.*, Gibb's Phenomenon), broadening of main-lobe width and other well-known problems associated with data truncation [7, 9, 11, 12, 57, 59].

Parametric models overcome the infinite dimensionality problem of non-parametric models by representing the system in terms of only a finite number of parameters or coefficients. Pole-zero models or transfer functions, linear differential and difference equations, Wiener/Kalman filters, exponential models, Markov models, finite automata, state-space representations etc. are some of the well known parametric system models. The primary advantage of these models is their ability to describe a possibly infinite dimensional system accurately and completely in terms of a parsimonious representation that is dependent only on a small number of parameters. Some of the problems encountered in parametric modeling include the proper choice of the type of the model that can appropriately describe the underlying system, accurate determination of model order as well as the estimation of the parameters defining the model. Over the last few decades these problems have been addressed in a large body of work [1-17, 24-50, 52, 53, 55, 57-59].

Among the parametric models mentioned above, the pole-zero or rational transfer function model is one of

the most effective and practical representations in digital signal processing literature. Over the years, research on parameter identification of rational transfer functions has evolved in four major directions. The first among the following four approaches deals with the modeling of random processes whereas the later three address deterministic problems. Firstly, in statistical and modern spectrum analysis literature, unknown systems or rather their Power Spectral Densities (PSD) are modeled using AR, MA and ARMA model identification methods [8, 9, 10, 11, 52]. The system parameters in these cases are estimated from the observed output data record only. Since the input to such systems can not be observed, white noise is assumed to be the input for modeling convenience. Secondly, in Control Systems, the input and output signals of a plant are usually measurable. Hence, the system parameters are estimated from the known input/output record [1-3, 13-17, 32, 55]. Thirdly, in Digital Filter design, the filter specifications may be given in the frequency domain in terms of the magnitude and phase responses in the pass/stop bands. Given such clear-cut specifications, standard methods are available [7, 12, 59] for designing IIR filters with classical structures such as Butterworth, Chebyshev, Elliptic and others. For arbitrary or non-classical frequency domain specifications, least-squares matching leads to general non-linear optimization algorithms [25, 26, 31, 33]. Fourthly, and again within the deterministic design context, in some applications the least-squares fit of a desired time-domain impulse response may be required or non-classical filter specifications with arbitrarily shaped pass/stopband response may be desired of an IIR filter or the unknown system may be known to contain unequal numbers of poles and zeros. In such cases, it is more advantageous and appropriate to fit the impulse response data directly to a rational transfer function model by estimating the unknown parameters of the model.

The modeling problem addressed in this Section belongs to the fourth category mentioned above. Specifically, given a desired impulse response the goal here is to estimate the coefficients of the numerator and denominator polynomials of the unknown rational transfer function by minimizing the model fitting error in the least-squares sense. It is well-known that this is a multidimensional nonlinear optimization problem [1-7, 10, 24-31, 34-37, 41-43, 55]. There have been a substantial amount of work on this particular problem starting, most probably, with the work of Kalman [1] where a linearized and approximate 'equation error' was minimized. In [3], Steiglitz and McBride (SM) proposed a linearized 'fitting error' minimization criterion whereas Shanks in [2] and Burrus and Parks in [30] proposed a couple of two-step procedures where the denominator polynomial was first estimated by minimizing an 'equation error' and then that denominator was utilized to obtain the numerator by minimizing a linearized 'fitting error'. These methods can not be expected to produce optimal estimates because the exact model fitting error norm is not minimized. For the special case of *strictly proper* transfer function models, *i.e.*, when the order of the numerator polynomial is exactly one less than the denominator polynomial order, an optimal solution that minimizes the exact fitting error criterion has been introduced by Evans and Fischl (EF) [5, 6]. Some of these approaches will be mathematically explained briefly in Subsection II so that their relationships with the proposed algorithm may be better appreciated.

Among other important works on this subject, a quasi-linearization method similar to SM was given in [24] and a modification of SM method was given in [63]. A least-squares Taylor approach was proposed in [27] and an algorithm relying on general non-linear optimization methods such as Gauss-Newton method was given in [62]. In [28] a two-step modified least-squares criterion was optimized using Pade synthesis technique and the standard Newton-Raphson method. In [29], the first two terms of the Taylor series expansion of the non-linear criterion was minimized and in [36] gradient based algorithms have been studied. The list of work cited here is not exhaustive, only some of the more important research are mentioned. Other related references may be found in the publications cited. A thorough treatment on 'filter design by modeling' based on a unified framework is presented in the book by Jackson [7, see also 10].

Most of the approaches cited above may be considered sub-optimal in the sense that theoretically they do not

minimize the exact model fitting error. In this work the ℓ_2-norm of the exact fitting error between the desired and the estimated impulse responses is minimized. The optimization is performed with respect to the denominator and the numerator coefficients. The proposed algorithm is closely related to the optimal technique proposed by Evans and Fischl [5-7]. It is well-known that the optimal EF method is applicable *only* for strictly proper rational transfer functions. In certain applications, such as exponential modeling, the strictly proper model is indeed the appropriate choice and a complex and constrained version of the EF method has been very effective [44-50, 60]. But the EF method has found limited applications in rational transfer function modeling problems because, in general, the highest degree of the numerator polynomial need not necessarily be exactly one less than the highest denominator degree [3, 4, 25]. The optimal method proposed here has no such restrictions and may be considered a generalization of the EF method. Furthermore, in contrast to the methods presented in [1-4, 24, 30] no linearization or modification of error criterion is introduced in the theoretical derivation of the least-squares model fitting criterion.

One of the critical steps in the theoretical derivation of the optimal criterion has been to identify the complete basis space orthogonal to the model fitting error. It will be shown later that even for general pole-zero (ARMA) models with arbitrary numerator and denominator orders, the orthogonal basis space can be completely defined in terms of the denominator coefficients only. The fitting error is then shown to be related to an equation error that turns out to be somewhat different than the one that appears in the EF method. But the form of the equation error is mathematically more appropriate for the general case considered here. The final error criterion possesses similar mathematical structure as in EF method but in the present generalized version, the dimensions of both the 'prefilter' matrix and the 'data' matrix vary according to the numerator and denominator orders. The inherent matrix prefiltering structure of the error criterion directly leads to formulating an efficient iterative computational algorithm for minimization.

The all-pole (AR) filter design problem can be considered a special case of fitting general pole-zero (ARMA) transfer functions models. Hence, given a desired impulse response, the proposed algorithm is also shown to produce the *optimum* least-squares estimates of the parameters of an all-pole (AR) model. Furthermore, for all-zero (MA) models, the well known Durbin's method basically relies on two AR model identification steps. By utilizing the proposed optimum AR-algorithm in one or both steps of Durbin's method, a modified Durbin-type algorithm is also presented for estimation of MA model parameters. The attractive feature of this work is that the proposed algorithm provides a unified and general framework for *optimal* identification of discrete-time systems from impulse response data encompassing a broad class of IIR and FIR structures.

This Section is arranged as follows : in Subsection II, the problem is defined and some related works are briefly outlined. In Subsection III the problem is formulated for the pole-zero (ARMA) case and the error optimization criterion is derived in detail. In Subsection IV, the criterion is appropriately modified to solve the all-pole (AR) case. In Subsection V, the all-zero (MA) case is addressed and in Subsection VI, several simulation examples are given. Finally, in Subsection VII, some discussion on the proposed algorithm is given and the Section is then concluded with some directions on possible future research.

II. Problem Statement and Related Methods

The rational transfer function model of a general recursive IIR digital filter can be represented in the z-domain as,

$$H(z) = \frac{a_0 + a_1 z^{-1} + a_{q-1} z^{-(q-1)} + \cdots + a_q z^{-q}}{1 + b_1 z^{-1} + \cdots + b_{p-1} z^{-(p-1)} + b_p z^{-p}} \triangleq \frac{N(z)}{D(z)}, \tag{1}$$

where the coefficient of z^0 term in denominator has been assumed to be unity without any loss of generality.

Equivalently, the transfer function $H(z)$ can be also written in terms of its impulse response as,

$$H(z) = h(0) + h(1)z^{-1} + \cdots + h(N-2)z^{-(N-2)} + h(N-1)z^{-(N-1)} + \cdots. \tag{2}$$

Stacking the first N 'significant' samples of $H(z)$, define,

$$\mathbf{h} \underline{\underline{\Delta}} [h(0) \quad h(1) \quad \cdots \quad h(N-1)]^T. \tag{3}$$

Next, denote vector containing the N samples of the prescribed or desired impulse response data as,

$$\mathbf{h}_d \underline{\underline{\Delta}} [h_d(0) \quad h_d(1) \quad \cdots \quad h_d(N-1)]^T. \tag{4}$$

Paraphrasing from Steiglitz's paper [4], 'in the best of worlds', given a desired impulse response $\mathbf{h}_d$, 'the ideal problem' of optimal estimation of the parameters a_i and b_i can be stated as :

$$\min_{\mathbf{a},\mathbf{b}} \|\mathbf{e}\|^2 \underline{\underline{\Delta}} \min_{\mathbf{a},\mathbf{b}} \sum_{i=0}^{N-1} \left[h_d(i) - \frac{N(z)}{D(z)} \{\delta(i)\} \right]^2, \qquad \text{where,} \tag{5}$$

$$\delta(i) = \begin{cases} 1, & i = 0 \\ 0, & i \neq 0, \end{cases} \tag{5a}$$

$$\mathbf{e} \underline{\underline{\Delta}} \mathbf{h}_d - \mathbf{h} \tag{5b}$$

$$\mathbf{a} \underline{\underline{\Delta}} [a_0 \quad a_1 \quad \cdots \quad a_q]^T \qquad \text{and} \tag{5c}$$

$$\mathbf{b} \underline{\underline{\Delta}} [1 \quad b_1 \quad \cdots \quad b_p]^T. \tag{5d}$$

This problem is known to be nonlinear in $\mathbf{b}$ and standard nonlinear optimization algorithms have been suggested [62, 23, 25-29, 36]. Several linearization approaches that exploit the inherent structure in this problem have also evolved [1-6, 24]. In order to motivate the proposed approach, some of the important related results are briefly outlined next.

Kalman's Method

In one of the earliest work on this subject, the solution of the following linear problem was suggested [1] :

$$\min_{\mathbf{a},\mathbf{b}} \sum_{i=0}^{N-1} \left[D(z)\{h_d(i)\} - N(z)\{\delta(i)\} \right]^2. \tag{6}$$

The advantage of this modified error criterion is that it can be easily minimized in the least-squares sense *w.r.t.* the unknown coefficients in $\mathbf{a}$ and $\mathbf{b}$ by solving a set of simultaneous linear equations. Apart from its simplicity, this approach is not known to possess any optimality property.

Shanks' Method

This is a two-step approach where the denominator coefficients are first estimated by minimizing an equation error at the tail end of the impulse response, *i.e.*,

$$\min_{\mathbf{b}} \sum_{i=p}^{N-1} [D(z)\{h_d(i)\}]^2. \tag{7a}$$

The minimization of this optimization criterion is also known as the 'covariance method' of linear prediction [7-11, 52, 53]. Once an estimate $\hat{D}(z)$ of the denominator polynomial is found, the numerator coefficients are estimated by minimizing the following modified fitting error norm :

$$\min_{\mathbf{a}} \sum_{i=0}^{N-1} \left[h_d(i) - \frac{N(z)}{\hat{D}(z)} \{\delta(i)\} \right]^2. \tag{7b}$$

Note that the estimation of **a** in (7b) is again a linear problem. Burrus and Parks [30] had presented another method that is closely related to Shank's work. The denominator was found in exactly the same manner as in (7a). To obtain the numerator, the first q error samples were forced to be zeros, *i.e.*, the first q samples of $h_d(n)$ were used as the best available estimates. The elements of **a** are then found from,

$$a_k = \sum_{i=0}^{k} \hat{b}_i h(i-k). \tag{7c}$$

The noteworthy feature of both these two-step procedures is that an essentially non-linear problem is converted into linear problems, but the methods are not known to produce optimal estimates.

Steiglitz-McBride's Iterative Prefiltering Method

In this method, an initial estimate $\hat{D}(z)$ of the denominator coefficients is first found by either Kalman's method (6) or Shanks' first step (7a). Then the following modified fitting error criterion is optimized iteratively,

$$\min_{\mathbf{a},\mathbf{b}} \sum_{i=0}^{N-1} \left[\frac{D(z)}{\hat{D}(z)} \{h_d(i)\} - \frac{N(z)}{\hat{D}(z)} \{\delta(i)\} \right]^2. \tag{8}$$

The estimate $\hat{D}(z)$ obtained at the i-th iteration step is used as prefilters for obtaining the estimates at the next iteration step. Note that (8) closely approximates (5) and both are exactly same if $D(z) = \hat{D}(z)$. But using (8), the unknown parameters in **a** and **b** can be estimated by solving a set of simultaneous linear equations. Further details on this method and its application in AR and ARMA model-based filter design may be found in [7].

It should be mentioned here that, very recently, McClellan and Lee [34] have shown that, instead of optimizing the original SM criterion in (8), it is possible to split the optimization problem into a linear and a non-linear problem. But more interestingly, for the strictly proper case ($p = q + 1$), they have also demonstrated that if **a** is estimated in a particular manner, then the corresponding *non-linear criterion* for estimating **b** has *exact mathematical equivalence* with the *iterative algorithm* of the optimal EF criterion (outlined next). This equivalence proof appears to explain why the SM method, which was originally proposed as a logical extension to Kalman's linearized approach, has been found to be effective in many applications over so many years. It should also be noted here that another decoupled version of the SM method may be found in [7] where **a** is estimated by minimizing the fitting error norm in (7b). It will be shown later that, for a given **b**, the criterion in (7b) produces the *optimal* least-squares estimate of **a**.

Evans-Fischl's Exact Fitting Error Minimization Method

The criteria in (6)-(8) attempt to but do not exactly solve the ideal problem stated in (5). But, as shown in the previous paragraph, the SM method does the best job of closely approximating the fitting error and it has found wide applications in 1-D and 2-D filtering [4, 7, 10, 24, 39, 40, 55, 61, 63]. As mentioned above, for strictly proper $H_{SP}(z)$ with $p = q + 1$, an optimal approach that minimizes the exact fitting error has been presented

in [5, 6]. In their approach, the orthogonality between the modeling error and the vector space spanned by the denominator coefficients was used to show that the following criteria are exactly equivalent to (5) :

$$\min_{\mathbf{a},\mathbf{b}} \|\mathbf{e}_{SP}(\mathbf{a},\mathbf{b})\|^2 \equiv \min_{\mathbf{b}} \|\mathbf{B}_{SP}(\mathbf{B}_{SP}^T\mathbf{B}_{SP})^{-1}\mathbf{B}_{SP}^T\mathbf{h}_d\|^2$$
$$= \min_{\mathbf{b}} \|\mathbf{h}_d^T\mathbf{B}_{SP}(\mathbf{B}_{SP}^T\mathbf{B}_{SP})^{-1}\mathbf{B}_{SP}^T\mathbf{h}_d\|^2 \qquad (9a)$$

where,

$$\mathbf{B}_{SP} \triangleq \begin{bmatrix} b_p & 0 & \cdots & 0 \\ b_{p-1} & b_p & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & b_1 & \ddots & b_p \\ 0 & 1 & \ddots & b_{p-1} \\ \vdots & \vdots & \ddots & b_1 \\ 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{N\times N-p}, \qquad (9b)$$

where the subscript 'SP' denotes the strictly proper case considered by in [5]. Note that, in deriving (9a), no linearization or approximation had been introduced at the outset. The criterion in (9a) is non-linear and an iterative minimization scheme was also given in [5, 6]. The initial estimate of $\mathbf{b}$ is found by setting the 'prefilter' matrix $(\mathbf{B}_{SP}\mathbf{B}_{SP}^T)^{-1} = \mathbf{I}_{(N-p)}$. This again results in the so-called 'covariance method' of (7a). At convergence of the EF iterations, the optimum $\mathbf{b}^*$ and the corresponding minimized error $\mathbf{e}_{SP}^*$ are obtained. Using the optimal error, the 'cleaned up' or optimum impulse response is found from,

$$\mathbf{h}^* = \mathbf{h}_d - \mathbf{e}_{SP}^*. \qquad (9c)$$

The first $p-1$ terms of this optimum impulse response is used for calculating the optimum $\mathbf{a}_{SP}^*$ as,

$$a_k^* = \sum_{i=0}^{k} \hat{b}_i^* h^*(i-k), \qquad \text{for, } k = 0,1,\ldots,(p-1). \qquad (9d)$$

More details of the algorithm may be found in [5, 6]. Also, the EF case can be considered a special case of the general algorithm presented here.

The EF method has been covered in detail in the books by Jackson [7] and Scharf [10]. It has also been extended for separable-denominator 2-D filter design in a series of papers [41-43]. A modified complex version of the EF method (with complex-conjugate symmetric constraints imposed on $\mathbf{b}$) has also been developed for maximum-likelihood estimation of multiple frequencies or angles-of-arrival from noisy observation data [44, 45, 47, 49, 60]. The complex version of the EF method has also been extended to 2-D for simultaneous frequency-wavenumber estimation from array data [46-48, 50]. It is briefly shown next that, for frequency/wavenumber estimation problems, the original EF method with $q = p-1$ is exactly appropriate. A general exponential model can be defined as,

$$x(n) \triangleq \sum_{k=1}^{p} \alpha_k e^{s_k n}, \qquad \text{for, } n = 0,1,2,\ldots$$

where, $s_k = \sigma_k + j\omega_k$. In this model, there are p unknown s_k's as well as p unknown α_k's. The z-Transform of $x(n)$ is

$$X(z) = \sum_{k=1}^{p} \frac{\alpha_k z}{(z - s_k)}.$$

Clearly, after summation of the p terms in the right hand side, the numerator order will indeed be one less than the denominator order.

Though the EF method is optimal for the $p = q + 1$ cases, its usefulness in rational transfer modeling has been very limited because the number of zeros can not always be restricted only to one less than the number of poles [see 4, for example]. It does not solve the exact fitting error optimization stated in (5) for the general ARMA ($q \neq p - 1$) and the AR ($q = 0$) cases. The primary objective of the present work is to fill this void.

III. Problem Formulation and Algorithm Development

In this Subsection, the problem is formulated for the ARMA case and the computational algorithm for the fitting error minimization is derived. The development and the execution may seem similar to the EF method, but it will be demonstrated that the proposed algorithm is uniformly applicable in a broad class of filter design problems.

III.1 : General ARMA Case

In this Subsection the general ARMA case with $q \leq p - 1$ is considered. Other ARMA cases with $q \geq p$ will be addressed later. From (1) and (2), equating the coefficients of equal powers of z, the transfer function coefficients can be related to the impulse response samples in $H(z)$ as

$$\begin{bmatrix} \mathbf{a} \\ \cdots \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_1 \\ \cdots \\ \mathbf{H}_2 \end{bmatrix} \mathbf{b} \tag{10}$$

where, $\mathbf{a}$ and $\mathbf{b}$ have been defined in (5c) and 5(d), respectively, and

$$\mathbf{H}_1 \triangleq \begin{bmatrix} h(0) & 0 & \cdots & 0 & 0 & \cdots & 0 \\ h(1) & h(0) & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h(q) & h(q-1) & \cdots & h(0) & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{(q+1)\times(p+1)}, \tag{10a}$$

$$\mathbf{H}_2 \triangleq \begin{bmatrix} h(q+1) & h(q) & \cdots & h(0) & 0 & \cdots & 0 \\ h(q+2) & h(q+1) & \cdots & h(1) & h(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h(p) & h(p-1) & \cdots & \cdots & \cdots & h(1) & h(0) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h(N-1) & h(N-2) & \cdots & \cdots & \cdots & \cdots & h(N-p-1) \end{bmatrix} \in \mathbb{R}^{(N-q-1)\times(p+1)}. \tag{10b}$$

If $\mathbf{b}$ and $\mathbf{H}_1$ are known exactly, $\mathbf{a}$ can be found simply by the following matrix-vector multiplication,

$$\mathbf{a} = \mathbf{H}_1\mathbf{b}. \tag{10c}$$

However, neither the exact $\mathbf{h}$ nor the matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ are known in practice and only the desired impulse response $\mathbf{h}_d$ is available. The elements of $\mathbf{H}_1$ and $\mathbf{H}_2$ in (10) can be replaced by the corresponding elements in the desired response $\mathbf{h}_d$ to form the matrices $\mathbf{H}_{d1}$ and $\mathbf{H}_{d2}$, respectively. But with $\mathbf{H}_{d1}$ and $\mathbf{H}_{d2}$, the equality in (10) will not hold and the lower $(N - q - 1)$ equations of (10) can then be written as :

$$\mathbf{H}_{d2}\mathbf{b} = \begin{bmatrix} h_d(q+1) & h_d(q) & \cdots & h_d(0) & 0 & \cdots & 0 \\ h_d(q+2) & h_d(q+1) & \cdots & h_d(1) & h_d(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_d(p) & h_d(p-1) & \cdots & \cdots & \cdots & h_d(1) & h_d(0) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_d(N-1) & h_d(N-2) & \cdots & \cdots & \cdots & \cdots & h_d(N-p-1) \end{bmatrix} \begin{bmatrix} 1 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} = \mathbf{d}(\mathbf{b}), \tag{11}$$

where $\mathbf{d(b)}$ is an "equation error". It may be pointed out again that, for $q = p - 1$, the minimization of $\|\mathbf{d(b)}\|^2$ produces the so-called "covariance method" of linear prediction [7-11, 52, 53]. It may also be recalled that the covariance method was the choice in [2] and [30] for linearly estimating $\mathbf{b}$, and also in the EF method [5, 6] for obtaining the initial estimate of $\mathbf{b}$. For the present general ARMA case with $q \leq p - 1$, the form of the equation error appearing in (11) is mathematically more appropriate even though it can not be called either "covariance method" or "auto-correlation method". As briefly outlined next, the initial estimate of $\mathbf{b}$ will be computed in the proposed algorithm, by minimizing $\|\mathbf{d(b)}\|^2$ *w.r.t.* the denominator coefficients. Equation (11) can be rewritten as,

$$\begin{bmatrix} h_d(q) & \cdots & h_d(0) & 0 & \cdots & 0 \\ h_d(q+1) & \cdots & h_d(1) & h_d(0) & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_d(p-1) & \cdots & \cdots & \cdots & h_d(1) & h_d(0) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_d(N-2) & \cdots & \cdots & \cdots & \cdots & h_d(N-p-1) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = - \begin{bmatrix} h_d(q+1) \\ h_d(q+2) \\ \vdots \\ h_d(p) \\ \vdots \\ h_d(N-1) \end{bmatrix} + \mathbf{d(b)}. \qquad (12)$$

Now, letting,

$$\mathbf{G} \triangleq \begin{bmatrix} h_d(q) & \cdots & h_d(0) & 0 & \cdots & \\ h_d(q+1) & \cdots & h_d(1) & h_d(0) & \cdot & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h(p-1) & \cdots & \cdots & \cdot\cdot & h(1) & h(0) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_d(N-2) & \cdots & \cdots & \cdots & \cdots & h_d(N-p-1) \end{bmatrix} \quad \text{and} \quad \mathbf{g} \triangleq \begin{bmatrix} h_d(q+1) \\ h_d(q+2) \\ \vdots \\ h_d(p) \\ \vdots \\ h_d(N-1) \end{bmatrix}, \qquad (13)$$

the minimization of $\|\mathbf{d(b)}\|^2$ with respect to $\hat{\mathbf{b}} = [b_1 \;\; b_2 \;\; \cdots \;\; b_n]^T$ results in the following initial estimate of $\mathbf{b}$,

$$\mathbf{b}^{(0)} = \begin{bmatrix} 1 \\ \cdots\cdots \\ -\mathbf{G}^{\#}\mathbf{g} \end{bmatrix}, \qquad (14)$$

where $\mathbf{G}^{\#} \triangleq (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T$ denotes the pseudo-inverse of $\mathbf{G}$. Since this estimate is obtained by minimization of an equation error only, it does not necessarily minimize the norm of the true model fitting error and that remains the primary objective. Next, the equation error $\mathbf{d(b)}$ is related to the model fitting error $\mathbf{e}$.

III.2 : Fitting Error Minimization

The lower partition of (10) is reproduced below :

$$\mathbf{H}_2\mathbf{b} = \mathbf{0}, \qquad (15a)$$

or,

$$\begin{bmatrix} h(q+1) & h(q) & \cdots & h(0) & 0 & \cdots & 0 \\ h(q+2) & h(q+1) & \cdots & h(1) & h(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h(p) & h(p-1) & \cdots & \cdots & \cdots & h(1) & h(0) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h(N-1) & h(N-2) & \cdots & \cdots & \cdots & \cdots & h(N-p-1) \end{bmatrix} \begin{bmatrix} 1 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} = \mathbf{0}. \qquad (15b)$$

These equations can also be expressed in a rearranged form as,

$$\begin{bmatrix} b_{q+1} & b_q & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ b_{q+2} & b_{q+1} & \cdots & b_1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ b_p & b_{p-1} & \cdots & \cdots & \cdots & b_1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & b_p & b_{p-1} & \cdots & \cdots & \cdots & b_1 & 1 \end{bmatrix} \begin{bmatrix} h(0) \\ h(1) \\ \vdots \\ h(N-1) \end{bmatrix} = \mathbf{0}. \tag{16a}$$

$$\text{or, by definition,} \qquad \mathbf{B}^T\mathbf{h} = \mathbf{0}, \tag{16b}$$

where the $(N-q-1) \times N$ matrix appearing in (16*a*) is defined as $\mathbf{B}^T$ in (16*b*). This is a key equation. It clearly demonstrates that the $(N-q-1)$ rows of $\mathbf{B}^T$ (*i.e.*, the columns of $\mathbf{B}$) are *orthogonal* to the impulse response vector $\mathbf{h}$. This relationship will be useful in developing the optimization criterion for the general ARMA case. Using the definition of $\mathbf{B}$ in (16) the equation error $\mathbf{d(b)}$ can also be written in a rearranged form as,

$$\mathbf{d(b)} \triangleq \mathbf{H}_{d2}\mathbf{b} \tag{17a}$$

$$= \begin{bmatrix} b_{q+1} & b_q & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ b_{q+2} & b_{q+1} & \cdots & b_1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ b_p & b_{p-1} & \cdots & \cdots & \cdots & b_1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & b_p & b_{p-1} & \cdots & \cdots & \cdots & b_1 & 1 \end{bmatrix} \begin{bmatrix} h_d(0) \\ h_d(1) \\ \vdots \\ h_d(N-1) \end{bmatrix} \tag{17b}$$

$$\triangleq \mathbf{B}^T\mathbf{h}_d. \tag{17c}$$

From (5b), the desired impulse response can be written as,

$$\mathbf{h}_d = \mathbf{h} + \mathbf{e}. \tag{18}$$

Plugging (18) into (17) and using the orthogonality result in (16),

$$\mathbf{d(b)} = \mathbf{B}^T[\mathbf{h} + \mathbf{e}] \tag{19a}$$

$$= \mathbf{B}^T\mathbf{h} + \mathbf{B}^T\mathbf{e} \tag{19b}$$

$$= \mathbf{B}^T\mathbf{e}. \tag{19c}$$

This equation establishes a key relationship between the equation error and the fitting error. But in order to facilitate the minimization of the fitting error norm in (5), an inverse relationship between $\mathbf{e}$ and $\mathbf{d(b)}$ is developed next.

For a given $\mathbf{b}$, let $\mathbf{a}^o$ denote the optimum numerator coefficient vector and $\mathbf{e}(\mathbf{a}^o, \mathbf{b})$ be the corresponding minimized fitting error. Then according to the orthogonality principle [10 (page-325), 22], this fitting error $\mathbf{e}(\mathbf{a}^o, \mathbf{b})$ must be orthogonal to the 'estimate' $\mathbf{h}(\mathbf{a}^o, \mathbf{b})$ which corresponds to the optimum $\mathbf{a}^o$ and the given $\mathbf{b}$. For a given $\mathbf{b}$, if this orthogonality does not hold, the non-zero projection of the error $\mathbf{e}(\mathbf{a}^o, \mathbf{b})$ onto $\mathbf{h}$ would contain further information about $\mathbf{a}$. This implies that by changing $\mathbf{a}^o$ appropriately, one could still minimize the length of the error $\mathbf{e}(\mathbf{a}^o, \mathbf{b})$. This contradicts the original assumptions about $\mathbf{a}^o$ and $\mathbf{e}(\mathbf{a}^o, \mathbf{b})$. Now the questions are how to construct this $\mathbf{e}(\mathbf{a}^o, \mathbf{b})$ which will be orthogonal to the estimate of $\mathbf{h}$ and how that error may be related to the

unknown denominator coefficients in $\mathbf{b}$. Fortunately, the answers to both these questions can be found in equation (16) which clearly shows that all the $(N-q-1)$ linearly independent columns of $\mathbf{B}$ are indeed orthogonal to $\mathbf{h}$ and, furthermore, $\mathbf{B}$ is constructed using denominator coefficients only. Since the number of unknowns in $\mathbf{a}$ is $q+1$, it can be concluded that, by construction, the $(N-q-1)$ column vectors of $\mathbf{B}$ constitute the complete basis space of $\mathbf{e}(\mathbf{a}^o, \mathbf{b})$ that is orthogonal to $\mathbf{h}(\mathbf{a}^o, \mathbf{b})$. This argument implies that one can construct the orthogonal $\mathbf{e}(\mathbf{a}^o, \mathbf{b})$ by a linear combination of all its basis vectors in the columns of $\mathbf{B}$, namely,

$$\mathbf{e}(\mathbf{a}^o, \mathbf{b}) = \mathbf{B}\mathbf{c}, \tag{20}$$

where, $\mathbf{c} \underline{\underline{\Delta}} [c_1 \; c_2 \; \cdots \; c_{N-q-1}]^T$ denotes a vector of constants which needs to be determined. In order to find $\mathbf{c}$, it may be observed that $\mathbf{e}(\mathbf{a}^o, \mathbf{b})$ must also satisfy (19). Hence, plugging the expression of $\mathbf{e}(\mathbf{a}^o, \mathbf{b})$ from (20) into (19b),

$$\mathbf{d}(\mathbf{b}) = (\mathbf{B}^T\mathbf{B})\mathbf{c}. \tag{21}$$

But $(\mathbf{B}^T\mathbf{B})$ is a square, invertible, banded Toeplitz matrix and hence, $\mathbf{c}$ has a unique solution :

$$\mathbf{c} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{d}(\mathbf{b}), \tag{22}$$

and using (17c),

$$\mathbf{c} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{h}_d. \tag{23}$$

Plugging this unique value of $\mathbf{c}$ back into the fitting error expression of (20),

$$\mathbf{e}(\mathbf{a}^o, \mathbf{b}) = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{h}_d \tag{24a}$$

$$\underline{\underline{\Delta}} \; \mathbf{P}_\mathbf{B}\mathbf{h}_d, \tag{24b}$$

where, $\mathbf{P}_\mathbf{B}$ denotes the projection matrix of $\mathbf{B}$. The fitting error $\mathbf{e}(\mathbf{a}^o, \mathbf{b})$ in (24) is exactly same as the one in (5b), except that it corresponds to the optimum $\mathbf{a}^o$. Interestingly, the dependence on $\mathbf{a}$ has also been removed in the process from the expression of the fitting error. The problem then is to estimate the optimum $\mathbf{b}^o$ by minimizing the ℓ_2-norm of $\mathbf{e}(\mathbf{a}^o, \mathbf{b})$ *w.r.t.* $\mathbf{b}$. For an optimum $\mathbf{a}^o$, the minimization problem of (5) is exactly equivalent to the following series of expressions :

$$\min_\mathbf{b} \|\mathbf{e}(\mathbf{a}^o, \mathbf{b})\|^2 = \min_\mathbf{b} \|\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{h}_d\|^2, \tag{25a}$$

$$= \min_\mathbf{b} \|\mathbf{P}_\mathbf{B}\mathbf{h}_d\|^2, \tag{25b}$$

$$= \min_\mathbf{b} \mathbf{h}_d^T\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{h}_d, \tag{25c}$$

$$= \min_\mathbf{b} \mathbf{b}^T\mathbf{H}_{d2}^T(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{H}_{d2}\mathbf{b}, \tag{25d}$$

$$= \min_\mathbf{b} \mathbf{d}^T(\mathbf{b})(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{d}(\mathbf{b}). \tag{25e}$$

Equation (25e) demonstrates that the fitting error minimization is equivalent to a "weighted equation error minimization" where $(\mathbf{B}^T\mathbf{B})^{-1}$ acts as the "matrix prefilter" or "weight matrix". At minimum, (25) produces the optimum denominator vector $\mathbf{b}^o$ and from this the minimized error can be found from,

$$\mathbf{e}(\mathbf{a}^o, \mathbf{b}^o) = \mathbf{P}_{\mathbf{B}^o}\mathbf{h}_d, \tag{26}$$

where, $\mathbf{B}^o$ is constructed from the optimum $\mathbf{b}^o$. The optimum estimate of the impulse response is then,

$$\mathbf{h}^o = \mathbf{h}_d - \mathbf{e}(\mathbf{a}^o, \mathbf{b}^o). \tag{27}$$

Using the optimum values of $\mathbf{h}^o$ and $\mathbf{b}^o$ in the top partition of (10), the optimum numerator can be found as follows,

$$\mathbf{a}^o = \mathbf{H}_1^o \mathbf{b}^o. \tag{28}$$

Equations (25) and (28) are the two key formulae for estimating the coefficients of the denominator and numerator polynomials, respectively. Close similarity between the final optimization criterion in (25) and the one in (9) is obvious. It must be emphasized though that (9) deals with an important but specific case of strictly proper transfer function with $q = p - 1$. But here, the derivation of (25) was based on appropriate choice of the orthogonal basis vectors in $\mathbf{B}$ for any $q \leq p-1$. Hence, using (25) a larger class of general pole-zero transfer function model fitting problems are solvable.

Intuitively, it appears that the decoupled estimation of $\mathbf{a}$ and $\mathbf{b}$, as found above, falls within a special class of non-linear optimization problems which have been studied extensively by numerical analysts [18-21]. It has been shown in [18] that in a non-linear error criterion, if some of the unknown variables are linearly related to the error and the other variables are non-linearly related and if these variables appear in a separable form, as happens to be the case here, then the two decoupled estimators in (25) and (28) should be the ***globally optimum estimators*** for both sets of variables. The derivations leading to (25) is based on the orthogonality principle and the simple solution of $\mathbf{a}$ in (28) is a direct consequence of (10c). But it is not exactly obvious if these results do possess the global optimality properties. An alternate derivation for indirectly arriving at (25) and (28) is given in Appendix A, where the desired optimality properties are clearly established. A computational algorithm for minimization of the criterion in (25) is briefly outlined next.

III.3 : Computational Algorithm

The criterion in (25) is non-linear in $\mathbf{b}$ and hence it can not be minimized directly. But instead of using standard non-linear optimization techniques the inherent mathematical structure of the criterion will be utilized to develop an iterative computational algorithm. The final form of the error vector in (24) is rewritten as,

$$\begin{aligned}
\mathbf{e} &= \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{h}_d && (29a)\\
&\underline{\underline{\Delta}}\ \mathbf{W}\mathbf{B}^T\mathbf{h}_d && (29b)\\
&= \mathbf{W}\mathbf{H}_{d2}\mathbf{b} && (29c)\\
&= \mathbf{W}\left[\mathbf{g} \vdots \mathbf{G}\right]\mathbf{b} && (29d)\\
&= \mathbf{W}\mathbf{g} + \mathbf{W}\mathbf{G}\hat{\mathbf{b}}, && (29e)
\end{aligned}$$

where,

$$\mathbf{W} \underline{\underline{\Delta}}\ \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}. \tag{29f}$$

If the matrix $\mathbf{W}$ is treated as independent of $\hat{\mathbf{b}}$, an expression for $\hat{\mathbf{b}}$ can be easily obtained by minimizing $\|\mathbf{e}\|^2$ *w.r.t.* $\hat{\mathbf{b}}$ as follows :

$$\begin{aligned}
\hat{\mathbf{b}} &= -(\mathbf{W}\mathbf{G})^{\#}\mathbf{W}\mathbf{g}\\
&= -\left(\mathbf{G}^T\mathbf{W}^T\mathbf{W}\mathbf{G}\right)^{-1}\mathbf{G}^T\mathbf{W}^T\mathbf{W}\mathbf{g}. && (30)
\end{aligned}$$

But since $\mathbf{W}$ does depend on the elements in $\hat{\mathbf{b}}$, (30) can only be computed iteratively. At the $(i+1)$-th step of iteration, $\mathbf{W}^{(i)}$ is formed using the estimate of $\mathbf{b}$ found in the i-th iteration step. This leads to the following iterative algorithm for computing $\mathbf{b}^{i+1}$:

$$\mathbf{b}^{(i+1)} = \begin{bmatrix} 1 \\ \cdots\cdots\cdots\cdots \\ -[\mathbf{X}^{(i)}\mathbf{G}]^{-1}[\mathbf{X}^{(i)}]\mathbf{g} \end{bmatrix} \tag{31}$$

where,

$$\mathbf{X}^{(i)} \triangleq \mathbf{G}^T\mathbf{W}^{T(i)}\mathbf{W}^{(i)} \tag{31a}$$

$$= \mathbf{G}^T(\mathbf{B}^{T(i)}\mathbf{B}^{(i)})^{-1}. \tag{31b}$$

The iterations are continued until $\|\mathbf{b}_{i+1} - \mathbf{b}_i\|^2 < \delta$, where δ is an arbitrarily small number. It must be noted here that the iterations in (31) may not always converge to the absolute minimum of the error criterion in (5) and hence the estimated $\mathbf{b}$ may not be the optimum one. This is because in (31) the variability of $\mathbf{W}$ *w.r.t.* $\mathbf{b}$ had been ignored while minimizing $\|\mathbf{e}\|^2$. To achieve the optimum, the gradient of the complete expression of $\|\mathbf{e}\|^2$ must be set to zero. If desired, this can be done in a second phase of the algorithm which is outlined in Appendix B. It may be noted here that the simulation studies indicate that the Phase-1 of iterations using (31) does an excellent job of bringing the estimate very close to the optimum. It wil. be shown in Subsection VI that the Phase-2, if invoked, only causes slight changes in the $\mathbf{b}$ vector and the minimized error norm. In simulations, the convergence was found to be quite rapid in both the phases. Once the estimates of $\mathbf{b}$ converge, $\mathbf{a}$ is computed by following the steps (26)-(28).

III.4 : ARMA Transfer Functions with $\mathbf{q \geq p}$

In this case, corresponding to the lower partition of (10), the equation error appearing in (11) has the following form,

$$\mathbf{H}_{d2}\mathbf{b} = \begin{bmatrix} h_d(q+1) & h_d(q) & \cdots & h_d(q-p) \\ h_d(q+2) & h_d(q+1) & \cdots & h_d(q-p+1) \\ \vdots & \vdots & \ddots & \vdots \\ h_d(p) & h_d(p-1) & \cdots & h_d(0) \\ \vdots & \vdots & \ddots & \vdots \\ h_d(N-1) & h_d(N-2) & \cdots & h_d(N-p-1) \end{bmatrix} \begin{bmatrix} 1 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} = \mathbf{d}(\mathbf{b}). \tag{32}$$

Note that the initial $(q-p)$ elements of $\mathbf{h}_d(n)$ do not play any role in this equation error. But again, when $q \geq p$, the minimization of the norm of this equation error is more appropriate than the use of either covariance or autocorrelation methods or (11). Similarly, for this case the equations equivalent to (17)-(19) are,

$$\mathbf{d}(\mathbf{b}) \triangleq \mathbf{H}_{d2}\mathbf{b} \tag{33a}$$

$$= \begin{bmatrix} 0 & \cdots & 0 & b_p & b_{p-1} & \cdots & b_1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & b_p & b_{p-1} & \cdots & b_1 & 1 \end{bmatrix} \begin{bmatrix} h_d(0) \\ h_d(1) \\ \vdots \\ h_d(N-1) \end{bmatrix} \tag{33b}$$

$$\triangleq \mathbf{B}^T\mathbf{h}_d \tag{33c}$$

$$= \mathbf{B}^T\mathbf{e}. \tag{33d}$$

Following similar arguments leading to (25), the optimization criterion can also be shown to be,

$$\min_{\mathbf{a}^o,\mathbf{b}} \|\mathbf{e}(\mathbf{a}^o,\mathbf{b})\|^2 \equiv \min_{\mathbf{b}} \|\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{h}_d\|^2 \tag{34a}$$

$$= \min_{\mathbf{b}} \|\mathbf{h}_d^T\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{h}_d\|^2 \tag{34b}$$

$$= \min_{\mathbf{b}} \mathbf{b}^T\mathbf{H}_{d2}^T(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{H}_{d2}\mathbf{b}. \tag{34c}$$

Observe that the matrix $\mathbf{B}^T$ appearing in (33) contains $(q-p)$ leading zero columns. This has the net effect of zeroing out the first $(q-p)$ elements of $\mathbf{h}_d$ and $\mathbf{e}$ in equations (33c) and (33d), respectively. Hence, the criterion

in (34) essentially minimizes the norm of a shorter error vector $\mathbf{e}' \triangleq [e(q-p)\ e(q-p+1)\ \cdots\ e(N-1)]^T$. The iterative algorithm described earlier in Subsection III.3 can again be utilized for estimating $\mathbf{b}$. But, if the estimated $\mathbf{b}$ is used in (26), the leading $(q-p)$ elements of the $\mathbf{e}$ vector will turn out to be zeros. Next, if (27) is used for calculating $\mathbf{h}^o$, the first $(q-p)$ elements of the estimated impulse response will have exactly same values as in the desired response $\mathbf{h}_d$. Finally, while estimating $\mathbf{a}$ using (28), the leading $(q-p)$ elements of $\mathbf{h}_d$ must be used without any error reduction.

The observations in the previous paragraph, if made casually, may lead one to conclude that the estimates obtained in this manner may not be optimal because the leading $(q-p)$ error samples are *not minimized* at all. But in Appendix A it has been proved that (25)-(28) do indeed produce the optimal estimates for *any* values of p and q, as long as $(p+q+1) \leq N$. That some of the error samples can not be minimized is only due to the mathematical nature inherent in the problem itself which enforces the leading $(q-p)$ elements of $\mathbf{h}_d$ and $\mathbf{h}$ to be equal when $q \geq p$. This should not be viewed as a limitation of the proposed algorithm or the optimal criterion. Furthermore, once the denominator is calculated using (34), one may consider minimizing the modified fitting error criterion in (7b) in order to estimate $\mathbf{a}^o$. But it has also been shown in Appendix A that for a given $\mathbf{b}$, (7b) and (26)-(28) produce *identical* estimates.

IV. An All-Pole Filter Design Algorithm

The $H(z)$ in this case has the following form,

$$H(z) = \frac{a_0}{1 + b_1 z^{-1} + \cdots + b_{p-1} z^{-(p-1)} + b_p z^{-p}} \triangleq \frac{a_0}{D(z)} \tag{37}$$

where, $q = 0$. The optimization problem of (5) can be restated as,

$$\min_{a_0, \mathbf{b}} \|\mathbf{e}\|^2 \triangleq \min_{a_0, \mathbf{b}} \sum_{i=0}^{N-1} \left[h_d(i) - \frac{a_0}{D(z)} \{\delta(i)\} \right]^2. \tag{38}$$

It has been shown in Subsection III and in Appendix A that the equations (25)-(28) developed in Subsection III is equally applicable for $\forall q \leq p-1$. Accordingly, the optimal AR-algorithm for determining a_0 and $\mathbf{b}$ can be treated as a special case of the iterative approach derived earlier for the general ARMA case. Hence only a brief summary will be given by defining the appropriate matrices involved in this case. The denominator coefficient vector $\mathbf{b}$ is estimated by optimizing,

$$\min_{a_0, \mathbf{b}} \|\mathbf{e}_{AR}(a_0, \mathbf{b})\|^2 \equiv \min_{\mathbf{b}} \|\mathbf{B}_{AR}(\mathbf{B}_{AR}^T \mathbf{B}_{AR})^{-1} \mathbf{B}_{AR}^T \mathbf{h}_d\|^2 \tag{39a}$$

$$= \min_{\mathbf{b}} \mathbf{b}^T \mathbf{H}_{d2}^{AR^T} (\mathbf{B}_{AR}^T \mathbf{B}_{AR})^{-1} \mathbf{H}_{d2}^{AR} \mathbf{b}, \tag{39b}$$

where,

$$\mathbf{B}_{AR} \triangleq \begin{bmatrix} b_1 & \cdots & b_p & 0 & \cdots & 0 \\ 1 & \cdots & b_{p-1} & b_p & \cdots & 0 \\ 0 & \ddots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \cdots & 1 & b_1 & \ddots & b_p \\ 0 & \cdots & 0 & 1 & \ddots & b_{p-1} \\ \vdots & \cdots & \vdots & \vdots & \ddots & b_1 \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{N \times N-1} \tag{39c}$$

and

$$\mathbf{H}_{d2}^{AR} \triangleq \begin{bmatrix} h_d(1) & h_d(0) & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_d(q+1) & h_d(q) & \cdots & h_d(0) & 0 & \cdots & 0 \\ h_d(q+2) & h_d(q+1) & \cdots & h_d(1) & h_d(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \vdots \\ h_d(p) & h_d(p-1) & \cdots & \cdots & \cdots & \cdots & h_d(0) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \vdots \\ h_d(N-1) & h_d(N-2) & \cdots & \cdots & \cdots & \cdots & h_d(N-p-1) \end{bmatrix} \in \mathbb{R}^{(N-1)\times(p+1)}. \qquad (39d)$$

Minimization of $\|\mathbf{e}_{AR}\|^2$ *w.r.t.* $\hat{\mathbf{b}}$ is carried out iteratively and at the $(i+1)$-th step of iteration, $\mathbf{b}$ is found from,

$$\mathbf{b}^{(i+1)} = \begin{bmatrix} 1 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ -[\mathbf{X}_{AR}^{(i)}\mathbf{G}_{AR}]^{-1}[\mathbf{X}_{AR}^{(i)}]\mathbf{g}_{AR} \end{bmatrix} \qquad (40a)$$

where,

$$\mathbf{X}_{AR}^{(i)} \triangleq \mathbf{G}_{AR}^{T}(\mathbf{B}_{AR}^{T(i)}\mathbf{B}_{AR}^{(i)})^{-1}, \qquad (40b)$$

$$\mathbf{G}_{AR} \triangleq \begin{bmatrix} h_d(0) & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_d(q) & \cdots & h_d(0) & 0 & \cdots & 0 \\ h_d(q+1) & \cdots & h_d(1) & h_d(0) & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_d(p-1) & \cdots & \cdots & \cdots & \cdots & h_d(0) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_d(N-2) & \cdots & \cdots & \cdots & \cdots & h_d(N-p-1) \end{bmatrix} \quad \text{and} \quad \mathbf{g}_{AR} \triangleq \begin{bmatrix} h_d(1) \\ \vdots \\ h_d(q+1) \\ h_d(q+2) \\ \vdots \\ h_d(p) \\ \vdots \\ h_d(N-1) \end{bmatrix}. \qquad (40c)$$

As in (14), the iterations are started using the initial estimate of $\mathbf{b}$ obtained from,

$$\mathbf{b}^{(0)} = \begin{bmatrix} 1 \\ \cdots\cdots\cdots \\ -\mathbf{G}_{AR}^{\#}\mathbf{g}_{AR} \end{bmatrix}. \qquad (41)$$

Once the iterations in (40a) converge, the optimal $\mathbf{e}$ and $\mathbf{h}$ are found similarly as in (26) and (27), respectively. Since $b_0 = 1$, the optimal numerator coefficient can be calculated from (28) as, $a_0 = h^o(0)$. Simulation runs indicate that the iterations converge rapidly in this case also. Though the all-pole case is considered here as a special case for the general pole-zero case, it should be emphasized that the all-pole design itself is an important problem in many applications. To the best of the knowledge of the author, the *optimal* solution for the all-pole case had remained unsolved.

V. An All-Zero Filter Design Algorithm

An all-zero transfer function is given by,

$$H(z) = a_0 + a_1 z^{-1} + a_{q-1}z^{-(q-1)} + \cdots + a_q z^{-q} \triangleq N(z). \qquad (42)$$

The problem then is to estimate **a** for a given $\mathbf{h}_d$ where $N > q+1$. The algorithms presented in Subsection III and IV are suitable for ARMA and AR filter design cases, respectively, and can not utilized directly for estimating MA filter coefficients. But the most effective algorithm known for MA modeling is the well-known Durbin's algorithm [64] which relies on two steps of AR parameter estimation. Traditionally, the *autocorrelation method* of linear prediction is utilized [7-11] in both the steps of Durbin's method, because it produces minimum-phase polynomials. But the estimates obtained from autocorrelation method may not be optimal. The algorithm given in Subsection IV is optimal for determining AR filter coefficients from prescribed impulse response. Hence, it can be reasonably hoped that the introduction of the proposed AR algorithm in one or both stages of Durbin's algorithm may produce more accurate results. The proposed modification on Durbin's algorithm is briefly outlined below.

Step 1 : Let $\frac{a_0}{D_L(z)}$ be an AR-model of $h_d(n)$ with 'sufficiently' large order L such that, $q << L < N$, and

$$H(z) \approx \frac{a_0}{D_L(z)}. \tag{43a}$$

The AR coefficients in, $\mathbf{b}_L \triangleq [1 \;\; b_1 \;\; \cdots \;\; b_L]^T$, can be estimated using the optimal algorithm given in Subsection IV.

Step 2 : If L is chosen large enough then the approximation in Step 1 will be very close. In that case,

$$\begin{aligned} \hat{D}_L(z) &\approx \frac{a_0}{H(z)} \\ &= \frac{a_0}{N(z)}. \end{aligned} \tag{43b}$$

Hence, using $\hat{\mathbf{b}}_L$ as 'data', the MA parameter vector **a** can again be estimated using the optimal algorithm presented in Subsection IV.

It must be noted here that the usage of the autocorrelation method at both the steps ensures that the final $N(z)$ is minimum phase. Instead, if the proposed algorithm is used, the minimum phase property can not be guaranteed. If minimum phase property is indeed desired of the final $\hat{N}(z)$, then autocorrelation method can be used (instead of (41)) to obtain the initial estimates for starting the iterative AR-algorithm. If the estimates obtained from the iterative algorithm becomes maximum phase at any iteration step of the AR-algorithm, the iterations can be terminated at that stage. The estimate found at the preceding iteration may be considered to be the best possible minimum phase estimate that the iterative AR-algorithm can produce. The drawback of this scheme is that one needs to root a polynomial (with possibly large order in Step-1) at each stage of iteration. But in some applications, this extra computational burden may turn out to be an acceptable trade-off in order to gain higher accuracy in the resulting estimates.

VI. Simulation Results

In this Subsection, the performance of the proposed algorithms are evaluated by means of several ARMA(p,q) and AR(p) model identification examples with various p and q values. $\delta = 10^{-3}$ was used as the stopping criterion in both phases of the algorithms for all the four examples below.

Simulation 1 : The desired impulse response has a Triangular form as shown by the solid lines in Fig.1a and 1b. The algorithm described in Subsection III was used with $p = 7$ and $q = 4$. The resulting impulse response fit at the end of each of the two phases are shown in circles in Fig. 1a and Fig. 1b. The minimized error norm and the closeness of the fit to the desired signal $\mathbf{h}_d$ are listed in Table 1. The number of iterations for convergence

are also listed. It can be seen that there is no significant difference between the 1st and the 2nd phase of the proposed algorithm.

Simulation 2 : An arbitrary impulse response was generated with $p = 8$ and $q = 4$. If the algorithm in Subsection III is used directly to match the true response it will give perfect answers. Instead, some perturbations or 'noise' was added to the true response to come up with the desired response $\mathbf{h}_d$. For 30dB 'noise', the true and the desired responses are shown in Fig. 2a. The impulse response match of the algorithm at the end of Phase-1 and Phase-2 are shown if Fig. 2b and Fig. 2c, respectively. The minimized error norms and the closeness to the true response ($\underline{\underline{\Delta}}$ $\mathbf{h}_t$) are listed in the first two columns of Table 2a, respectively. In the third column, the closeness to $\mathbf{h}_d$ is also given. It can be observed that the closeness figure to $\mathbf{h}_t$ may decrease from Phase-1 to Phase-2 because the algorithm is not attempting to match that directly. In Figures 2d, 2e and 2f and Table 2b, the corresponding results with 20dB perturbation are given.

Simulation 3 : In this case, the denominator and the numerator coefficients of Simulation 2 were switched to obtain an impulse response with $p = 4$ and $q = 8$. For 30dB 'noise', the true and the desired responses are shown in Fig. 3a. The algorithm in Subsection III.4 was used. The impulse response fit at the end of Phase-1 and Phase-2 are shown if Fig. 3b and Fig. 3c, respectively. Other results are listed in Table 3.

Simulation 4 : With the Triangular desired impulse response of Simulation 1, the AR-algorithm presented in Subsection IV was employed with $p = 5$. The resulting impulse response fit at the end of each of the two phases are shown in Fig. 4a and Fig. 4b, respectively. The minimized error norm, the closeness of the fit to the desired signal $\mathbf{h}_d$ and the number of iterations for convergence are listed in Table 4.

It can be fairly concluded from these simulations that the Phase-1 of the algorithm does an excellent job of error minimization. Hence, the Phase-2 of the algorithm need not be invoked for most applications.

VII. Discussion and Concluding Remarks

In this Subsection, a classical rational model identification problem has been addressed and, for the most parts, appears to have been *solved*. The major focus is to demonstrate that, given a desired impulse response corresponding to an unknown transfer function with *arbitrary* number of poles and zeros, it is possible to obtain the estimates of the parameters of the transfer function that are *optimal in the least-squares sense*. Unlike many well-known existing results, no linearization or approximation has been done while deriving the theoretical optimization criterion. The proposed technique is applicable in a comprehensive class of ARMA, AR and MA model identification problems. It is shown that the multidimensional non-linear problem can be decoupled into two smaller problems of which one is a linear problem and the other one is a non-linear problem. The inherent mathematical structure of the non-linear part is utilized to formulate an efficient iterative computational algorithm for estimating the denominator parameters. The numerator is then found by a simple matrix-vector multiplication. Global optimality properties of the estimators have been verified by relating the multi-dimensional optimization problem to certain well-known results in numerical analysis. In simulation studies also, the method has been shown to be highly effective. Some important aspects related to the algorithm but are not covered above are addressed next :

Model Order Selection :

Model order selection of rational transfer function models from impulse response data remains an open problem. This aspect of the problem has not been here. It appears that for this essentially deterministic problem, Akaike Information Criterion (AIC) or Minimum Description Length Criterion (MDL) may not be applicable. In

deriving the optimal criterion, the model orders p and q have been tacitly assumed to be known or the estimates are assumed to be available. But in defense of this work, it may be noted that model order selection was not addressed in any of the previous results on this problem. In simulation studies, it was observed that if the model orders are increased, then, in some cases, the estimated response gets closer to the desired response. But over-determination of model orders may lead to unstable and/or useless solutions because the matrices may tend to be singular. Furthermore, increasing model orders also raises the computational load on the resulting filter. Hence, there is a trade-off that has to be considered in deciding the proper choice of the model order.

Relationship with SM method and Convergence Analysis of the iterative algorithm :

As already noted in Subsection II, in a recent paper [34], McClellan and Lee have shown that for the *strictly proper case* ($p = q + 1$), if the original SM method is decoupled into a linear and a non-linear estimation problem in a certain manner, the resulting non-linear SM criterion has *exact mathematical equivalence* with the *optimal* EF criterion which is only applicable for the strictly proper transfer functions. The SM method is also known to be very effective for general ARMA and AR cases. But in this Subsection the *optimal* criteria for both ARMA and AR models have been derived. Hence it would certainly be interesting to investigate whether the equivalence also holds for the general cases. It appears that by using the new definitions of the matrices $\mathbf{H}_1$, $\mathbf{H}_2$ and $\mathbf{B}$, appearing in (10), (16), (32), (33) and (39), McClellan-Lee's derivation can be appropriately modified for any other values of p and q also. With these modifications, it can be easily shown that the original general SM method for arbitrary p and q can also be decoupled as suggested in [34]. In the decoupled form, the *non-linear criterion* of the SM method can also be shown (proof omitted for space limitation) to be *mathematically equivalent* to the *iterative algorithm* given in (31) for minimizing the *optimal* criterion in (25). This equivalence proof may have another important consequence for the proposed algorithm. There already exists a convergence analysis of the *original* SM method in [55]. One can reasonably hope that the convergence analysis will also apply to the decoupled form of SM method given in [34]. If that happens to be the case, as alluded to in [34], the convergence analysis in [55] should also apply to the iterative computational algorithm presented here.

Computational Requirements :

The major computational load of the algorithm is in performing the iterative refinement in (29) - (31), where, at each iteration step, one needs to invert an $(N - q - 1) \times (N - q - 1)$ matrix $(\mathbf{B}^T\mathbf{B})$. It may appear that this inversion should require $O[(N - q - 1)^3]$ operations. But $(\mathbf{B}^T\mathbf{B})$ is a symmetric-banded-Toeplitz matrix and many efficient algorithms are available for inverting such matrices [9-11, 45, 56, 67]. Specifically, in [56] it has been shown that inversion of such banded Toeplitz matrices only requires $O[(N - q - 1)\log(N - q - 1)] + O[p^2]$ operations. Furthermore, in the SM method, the calculation of the impulse response of the inverse filter and data filtering are required at every step of iteration, whereas the proposed method uses the estimated $\mathbf{b}$ directly to form the $\mathbf{B}$ matrix.

Future Work :

Many of the 1-D algorithms discussed in the Introduction have been extended to 2-D for estimating 2-D filter coefficients from spatial domain data [29, 35, 36, 39-43]. It appears that by identifying the appropriate orthogonal subspaces, it should be possible to formulate an *optimal* 2-D filter design algorithm by extending the optimal 1-D method proposed in this Section. It is also possible to extend this work for identification of Multidimensional systems from multidimensional impulse response data [38]. Work on these topics are under progress [65, 66] and the preliminary results look encouraging.

REFERENCES

[1] R. E. Kalman, "Design of a Self Optimizing Control System," *Trans. ASME*, Vol. 80, pp. 468-478, 1958.

[2] J.L.Shanks, "Recursion Filters for Digital Processing", *Geophysics*, Vol. 32, pp. 33-51, 1967.

[3] K. Steiglitz and L.E. McBride, "A Technique for Identification of Linear Systems", *IEEE Transactions on Automatic Control*, Vol. AC-10, pp. 461-464, 1965.

[4] K. Steiglitz, "On the Simultaneous Estimation of Poles and Zeros in Speech Analysis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-25, no. 3, pp. 229-234, June, 1977.

[5] A.G. Evans and R. Fischl, "Optimal Least Squares Time-Domain Synthesis of Recursive Digital Filters", *IEEE Transactions on Audio and Electro-Acoustics*, Vol. AU-21, pp. 61-65, 1973.

[6] A. G. Evans, *Least-Squares System Parameter Identification and Time Domain Approximation*, Ph. D. Dissertation, Drexel University, 1972.

[7] L.B. Jackson, *Digital Filters and Signal Processing*, Kluwer, Boston, 1986.

[8] S. M. Kay and L. Marple, "Spectrum Analysis - A Modern Perspective," *Proceedings of the IEEE*, vol. 69, no. 11, pp. 1380-1419, Nov. 1981.

[9] S.M. Kay, *Modern Spectral Estimation: Theory and Applications*, Prentice Hall, Englewood Cliffs, NJ, 1988.

[10] L. L. Scharf, *Statistical Signal Processing - Detection, Estimation and Time Series Analysis*, Addison-Wesley, Reading, MA, 1990.

[11] S. L. Marple, *Digital Spectral Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[12] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Englewood Cliffs, NJ, Prentice-Hall, 1975.

[13] K.J. Åström and P. Eykhoff, "System identification: a survey", *Automatica*, vol. 7, pp. 123-167, 1971.

[14] L. Ljung, *System Identification: Theory for the Users*, Prentice Hall, NJ, 1987.

[15] A.P. Sage and J.L. Melsa, *System Identification*, Academic Press, NY, 1971.

[16] K.J. Åström and P. Eykhoff, "System identification: a survey", *Automatica*, vol. 7, pp. 123-167, 1971.

[17] T. Söderström and P. Stoica, *System Identification*, Prentice Hall, NJ, 1987.

[18] G. H. Golub and V. Pereyra, "The Differentiation of Pseudoinverses and Nonlinear Problems Whose Variables Separate," *SIAM Journal on Numerical Analysis*, vol. 10, no. 2, pp. 413-432, Apr., 1973.

[19] H. D. Scolnik, "On the Solution of Nonlinear Least Squares Problems," *Proceedings of IFIP-1971*, Numerical Mathematics, North Holland, Amsterdam, pp. 18-23, 1971.

[20] A. Perez and H. D. Scolnik, "Derivatives of Pseudoinverses and constrained non-linear regression problems," *Numer. Math.*

[21] I. Guttman, V. Pereyra, "Least Squares Estimation for a Class of Non-Linear Models," *Technometrics*, vol. 15, no. 2, pp. 209-218, May, 1973.

[22] D.G. Luenberger, *Optimization by Vector Space Method*, New York: Wiley, 1969.

[23] R. Fletcher and M.J.D. Powell, "A Rapidly Convergent Descent Method for Minimization", *Computer Journal*, Vol. 6, pp. 163-168, 1963.

[24] E. R. Schulz, "Estimation of Pulse Transfer Function Parameters by Quasilinearization," *IEEE Transactions on Automatic Control*, pp. 424-426, 1968.

[25] K. Steiglitz, "Computer-Aided Design of Recursive Digital Filters," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-18, no. 2, pp. 123-129, June, 1970.

[26] A. G. Deczky, "Synthesis of Recursive Digital Filters Using the Minimum p-Error Criterion," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-20, no. 4, pp. 257-263, June, 1970.

[27] R. Fischl, "Optimal l_p-Approximation of Prescribed Impulse Response Functions on a Finite Point Set," in *Proc. IEEE Int. Symp. on Circuit Theory*, pp. 155-156, 1970.

[28] F. Brophy and A. C. Salazar, "Recursive Filter Synthesis in the Time Domain," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-22, no. 1, pp. 45-55, Feb., 1974.

[29] M. S. Bertran, "Approximation of Digital Filters in One and Two Dimensions," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-23, no. 5, pp. 438-443, Oct., 1975.

[30] C. S. Burrus and T. W. Parks, "Time Domain Design of Recursive Digital Filters," *IEEE Trans. on Aud. Elect.*, vol. AU-18, pp. 137-141, June, 1970.

[31] C. Charalambous, "Minimax Optimization of Recursive Digital Filters Using Recent Minimax Results," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-23, no. 4, pp. 333-345, June, 1977.

[32] M. J. Levin, "Estimation of a System Pulse Transfer Function in the Presence of Noise," *IEEE Transactions on Automatic Control*, pp. 229-235, 1964.

[33] G. Miller, "Least-Squares Rational Z-Transform Approximation," *Journal of the Franklin Institute*, vol. 295, no. 1, pp. 1-7, Jan., 1973.

[34] J. H. McClellan and D. Lee, "Exact Equivalence of the Steiglitz-McBride Iteration and IQML," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-39, no. 2, pp. 509-512, Feb., 1991.

[35] J.L. Shanks, S. Treitel and J.H. Justice, "Stability and Synthesis of Two-Dimensional Recursive Filters" *IEEE Trans. Audio Electroacoust.*, Vol. AU-20, pp. 115-128, 1972.

[36] J. A. Cadzow, "Recursive Digital Filter Synthesis via Gradient Based Algorithms", *IEEE Transaction on Acoustic, Speech and Signal Processing*, Vol. ASSP-24, pp. 349-355, 1976.

[37] A. K. Shaw and P. Misra, "Time Domain Identification of Proper Discrete Systems from Measured Impulse Response Data,", *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1689-1692, Toronto, Canada, May, 1991.

[38] A. K. Shaw, P. Misra and R. Kumaresan, "Identification of Multi-Dimensional Systems From Impulse Response Data," accepted as a full paper, *30th IEEE Conf. on Dec. and Contr.*, Brighton, UK, Dec. 1991.

[39] G. A. Shaw and R. M. Mersereau, "Design, Stability and Performance of Two-Dimensional Recursive Digital Filters", *Tech. Report E21-B05-1*, Georgia Inst. of Technology School of Electrical Engg., 1979.

[40] D.E. Dudgeon and R.M. Mersereau, *Multidimensional Digital Signal Processing*, Englewood Cliffs, NJ: Prentice Hall, 1984.

[41] T. Hinamoto and S. Maekawa, "Spatial-Domain Design of a Class of Two-Dimensional Recursive Digital Filters", *IEEE Transaction on Acoustic, Speech and Signal Processing*, Vol.- ASSP-32, no.1, Feb.,1984.

[42] T. Hinamoto and S. Maekawa, "Separable-Denominator 2-D Rational Approximation via 1-D Based Algorithm ", *IEEE Transactions on Circuits and Systems*, Vol. CAS-32, pp. 989-999, Nov. 1985.

[43] T. Hinamoto, "Design of 2-D Separable-Denominator Recursive Digital Filters", *IEEE Transactions on Circuits and Systems*, Vol. CAS-31, pp. 925-933, Nov. 1984.

[44] R. Kumaresan and A. K. Shaw, "High Resolution Bearing Estimation Without Eigendecomposition," *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, Florida, April, 1985.

[45] R. Kumaresan, L. L. Scharf and A. K. Shaw, "An Algorithm for Pole-Zero Modeling and Spectral Estimation," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol.ASSP-34, pp. 637-640, June, 1986.

[46] A. K. Shaw and R. Kumaresan, "Frequency-Wavenumber Estimation by Structured Matrix Approximation," *Proceedings of the 3rd. IEEE-ASSP Workshop on Spectral Estimation*, Boston, pp. 81-84, Nov. 1986.

[47] A.K. Shaw, *Structured Matrix Problems in Signal Processing*, Ph.D. Dissertation, Univ. of Rhode Island, RI, 1987.

[48] A.K. Shaw and R. Kumaresan, "Some Structured Matrix Approximation Problems", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, NY, pp. 2324-2327, April, 1988.

[49] R. Kumaresan and A.K. Shaw, "Superresolution by Structured Matrix Approximation", *IEEE Transactions on Antennas and Propagation*, Vol. AP-36, pp. 34-44, 1988.

[50] R. Kumaresan and A. K. Shaw, "An Exact Least Squares Fitting Technique for Two-Dimensional Frequency Wavenumber Estimation," *Proceedings of the IEEE*, vol. 74, no. 4, pp. 606-607, April, 1986 .

[51] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, New Jersey, 1984.

[52] J. Makhoul, "Linear Prediction : A Tutorial Review," *Proceedings of the IEEE*, vol. 63, pp. 561-580, April, 1975.

[53] R. Prony, "Essai Experimental et Analytique etc.," L'Polytechnique, Paris, 1 Cahier 2, pp. 24-76, 1795.

[54] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and its Applications.* New York; Wiley, 1971.

[55] P. Stoica and T. Söderström, "The Steiglitz-McBride Identification Algorithm Revisited - Convergence Analysis and Accuracy Aspects," *IEEE Transactions on Automatic Control*, vol. AC-26, no. 3, pp. 712-717, June, 1981.

[56] A. K. Jain, "Fast Inversion of Banded Toeplitz Matrices by Circular Decompositions," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-26, no. 2, pp. 121-126, April, 1978.

[57] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[58] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, 1977.

[59] L. R. Rabiner and C. M. Rader, eds., *Digital Signal Processing*, New York, IEEE Press, 1972.

[60] Y. Bressler and A. Macovski, "Exact Maximum Likelihood Parameter Estimation of Superimposed Exponential Signals in Noise," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 10, pp. 1081-1089, Oct., 1986.

[61] J. S. Lim, *Two-Dimensional Signal and Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[62] C. S. Burrus, T. W. Parks, and T. B. Watt, "A Digital Parameter-Identification Technique Applied to

Biological Signals," *IEEE Trans. on Bio-Medical Engineering*, vol. BME-18, pp. 35-37, Jan., 1971.

[63] A. Stefanski and C. Weygandt, "Extension of the Steiglitz and McBride Identification Technique," *IEEE Transactions on Automatic Control*, pp. 503-504, Oct., 1971.

[64] J. Durbin, "Efficient Estimation of Parameters in Moving-Average Models," *Biometrika*, vol. 46, 1959, pp. 306-316.

[65] A. K. Shaw, "Optimal Design of Two-Dimensional Filters from Spatial Domain Data," under progress.

[66] A. K. Shaw, "Optimal Identification of General Multi-Dimensional Systems," under progress.

[67] S. Zohar, "Toeplitz Matrix Inversion : The Algorithm of W. F. Trench," *Journal of the Association of Computing Machinery*, vol. 16, pp. 592-601, Oct. 1969.

APPENDIX A

An Alternative Derivation of the Error Criterion

In this Appendix, an alternate derivation is given for the error criterion in (25). This derivation reaffirms that the optimal criterion for estimating **b** using (25) and the estimate of the numerator in (28) are indeed exactly equivalent to the original criterion appearing in (5). The criterion in (5) needs to be optimized *w.r.t.* two sets of parameters in **a** and **b**, where **a** is linearly related to the error whereas **b** has a non-linear relationship. In this appendix, it is shown that utilizing filtering interpretation it is possible to split this multidimensional optimization problem into a linear estimation problem for **a** and a non-linear optimization problem for **b**. The filtering interpretation also makes it possible to relate this problem to certain non-linear optimization problems studied by numerical analysts [18-21].

Let $H_b(z)$ be the inverse filter corresponding to $D(z)$, *i.e.*,

$$D(z)H_b(z) = 1. \tag{A.1}$$

This is a convolution operation where the discrete sequence b_k's in $D(z)$ are finite whereas the $h_b(n)$'s in $H_b(z)$ are infinite in extent. In matrix notation this convolution operation may be expressed as,

$$\mathbf{D}\mathbf{H}_b = \mathbf{I}_N \tag{A.2a}$$

where

$$\mathbf{D} \triangleq \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ b_1 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ b_{q+1} & b_q & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ b_{q+2} & b_{q+1} & \cdots & b_1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ b_p & b_{p-1} & \cdots & \cdots & \cdots & b_1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & b_p & b_{p-1} & \cdots & \cdots & \cdots & b_1 & 1 \end{bmatrix} \in \mathbb{R}^{N \times N} \quad \text{and} \tag{A.2b}$$

$$\mathbf{H}_b \triangleq \begin{bmatrix} h_b(0) & 0 & \cdots & 0 & 0 & \cdots & 0 \\ h_b(1) & h_b(0) & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_b(q) & h_b(q-1) & \cdots & h_b(0) & 0 & \cdots & 0 \\ h_b(q+1) & h_b(q) & \cdots & h_b(1) & h_b(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_b(N-1) & h_b(N-2) & \cdots & \cdots & \cdots & h_b(1) & h_b(0) \end{bmatrix} \in \mathbb{R}^{N \times N} \qquad (A.2c)$$

where, $\mathbf{I}_N$ denotes an $N \times N$ identity matrix. Now, rewriting (1),

$$\begin{aligned} H(z) &= \frac{N(z)}{D(z)} \\ &\triangleq H_b(z)N(z). \end{aligned} \qquad (A.3)$$

The right hand side in (A.3) again represents convolution of $h_b(n)$ with the numerator coefficient sequence a_n. In matrix-vector notation, the vector $\mathbf{h}$ containing the impulse response values in (A.3) can be represented as,

$$\mathbf{h} \triangleq \mathbf{H}'_b\mathbf{a} \qquad (A.4)$$

where, $\mathbf{H}'_b$ contains the first $q+1$ columns of $\mathbf{H}_b$, *i.e.*,

$$\mathbf{H}'_b \triangleq \begin{bmatrix} h_b(0) & 0 & \cdots & 0 \\ h_b(1) & h_b(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_b(q-1) & h_b(q-2) & \cdots & 0 \\ h_b(q) & h_b(q-1) & \cdots & h_b(0) \\ \vdots & \vdots & \ddots & \vdots \\ h_b(N-1) & h_b(N-2) & \ddots & h_b(N-q-1) \end{bmatrix} \in \mathbb{R}^{N \times (q+1)}. \qquad (A.4)$$

With these definitions, the problem stated in (5) can be rephrased as,

$$\min_{\mathbf{a},\mathbf{b}} \|\mathbf{e}\|^2 \triangleq \min_{\mathbf{a},\mathbf{b}} \|\mathbf{h}_d - \mathbf{H}'_b\mathbf{a}\|^2, \qquad (A.5)$$

where, using (A.4) in (5b), the error vector is formed as,

$$\mathbf{e} \triangleq \mathbf{h}_d - \mathbf{H}'_b\mathbf{a}. \qquad (A.5a)$$

This equation clearly shows the linear relationship between the error $\mathbf{e}$ and $\mathbf{a}$ and also the non-linear relationship between $\mathbf{e}$ and $\mathbf{b}$ through the matrix $\mathbf{H}'_b$. In this form, it is apparent that the present problem belongs to a larger class of mixed optimization problems where the linear and nonlinear variables appear separately. This class of problems have been studied extensively in numerical analysis literature [18-21]. The main objective is to optimize the two sets of variables independently. If $\mathbf{H}'_b$ (*i.e.* $\mathbf{b}$) is exactly known, then the minimization of (A.5) will produce the linear least-squares estimate of $\mathbf{a}$ as follows,

$$\hat{\mathbf{a}} \triangleq \mathbf{H}'^{\#}_b\mathbf{h}_d. \qquad (A.6)$$

In practice, $\mathbf{b}$ will not be known and needs to be estimated. Plugging $\hat{\mathbf{a}}$ back in (A.5), the optimization criterion for $\mathbf{b}$ is given by,

$$\min_{\mathbf{a},\mathbf{b}} \|\mathbf{h}_d - \mathbf{H}'_b\mathbf{a}\|^2 \equiv \min_{\mathbf{b}} \|\mathbf{h}_d - \mathbf{H}'_b\mathbf{H}'^{\#}_b\mathbf{h}_d\|^2 \tag{A.7a}$$

$$= \min_{\mathbf{b}} \|\mathbf{h}_d - \mathbf{P}_{\mathbf{H}'_b}\mathbf{h}_d\|^2 \tag{A.7b}$$

$$= \min_{\mathbf{b}} \|(\mathbf{I}_N - \mathbf{P}_{\mathbf{H}'_b})\mathbf{h}_d\|^2. \tag{A.7c}$$

For a larger class of problems, it has been proved in Theorem 2.1 of [18] that if $\hat{\mathbf{b}}$ is estimated by minimizing the criterion in (A.7) and if that estimate is utilized in computing $\hat{\mathbf{a}}$ using (A.6), then the resulting estimates are the *unique and global minimizers* of the criterion in (A.5).

The derivation of the optimization criterion in (A.6)-(A.7) is concise, though rigorous, but direct optimization of (A.7) would require taking resort to standard non-linear optimization techniques. This is because the parameters in $\mathbf{b}$ are related to the error criterion in a complicated manner through $\mathbf{P}_{\mathbf{H}'_b}$. Next, the criterion in (A.7) is reparameterized so as to relate it directly to the coefficients in $\mathbf{b}$. Appropriately partitioning the matrices $\mathbf{D}$ and $\mathbf{H}_b$ as follows,

$$\mathbf{D} \triangleq \left[\begin{array}{cccccccccc} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ b_1 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ b_q & b_{q-1} & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ \hline b_{q+1} & b_q & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\ b_{q+2} & b_{q+1} & \cdots & b_1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ b_p & b_{p-1} & \cdots & \cdots & \cdots & b_1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & b_p & b_{p-1} & \cdots & \cdots & \cdots & b_1 & 1 \end{array}\right] \triangleq \left[\begin{array}{c} \mathbf{B}_u^T \\ \hline \mathbf{B}^T \end{array}\right] \quad \text{and} \tag{A.8a}$$

$$\mathbf{H}_b \triangleq \left[\begin{array}{cccc|ccc} h_b(0) & 0 & \cdots & 0 & 0 & \cdots & 0 \\ h_b(1) & h_b(0) & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_b(q) & h_b(q-1) & \cdots & h_b(0) & 0 & \cdots & 0 \\ h_b(q+1) & h_b(q) & \cdots & h_b(1) & h_b(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ h_b(N-1) & h_b(N-2) & \cdots & \cdots & \cdots & h_b(1) & h_b(0) \end{array}\right] \triangleq [\mathbf{H}'_b|\mathbf{H}''_b], \tag{A.8b}$$

the equation (A.2) can be written as,

$$\left[\begin{array}{c} \mathbf{B}_u^T \\ \hline \mathbf{B}^T \end{array}\right] [\mathbf{H}'_b|\mathbf{H}''_b] = \mathbf{I}_N \tag{A.8c}$$

$$\text{or,} \quad \left[\begin{array}{c|c} \mathbf{B}_u^T\mathbf{H}'_b & \mathbf{B}_u^T\mathbf{H}''_b \\ \hline \mathbf{B}^T\mathbf{H}'_b & \mathbf{B}^T\mathbf{H}''_b \end{array}\right] = \left[\begin{array}{c|c} \mathbf{I}_{(q+1)} & \mathbf{0}_{(q+1)\times(N-q-1)} \\ \hline \mathbf{0}_{(N-q-1)\times(q+1)} & \mathbf{I}_{(N-q-1)\times(N-q-1)} \end{array}\right]. \tag{A.8d}$$

It should be noted that in the partitioning example above, only the $q \leq p-1$ case is shown, but the results are equally applicable for any values of p and q. The bottom-left corner element shows that the $N \times (N-q-1)$

matrix $\mathbf{B}$ and the $N \times (q+1)$ matrix $\mathbf{H}_b'$ are orthogonal, *i.e.*, $\mathbf{B}^T\mathbf{H}_b' = 0_{(N-q-1)\times(q+1)}$ and since by construction both are full-rank matrices, it is also true that,

$$rank(\mathbf{B}) + rank(\mathbf{H}_b') = N. \tag{A.9}$$

Hence, using a theorem on projection matrices [54],

$$\mathbf{P}_{\mathbf{B}} + \mathbf{P}_{\mathbf{H}_b'} = I_N. \tag{A.10}$$

Using this result in (A.7c), the following reparameterized form of the optimization criterion is obtained,

$$\min_{\mathbf{b}} \|\mathbf{P}_{\mathbf{B}}\mathbf{h}_d\|^2. \tag{A.11}$$

Interestingly, this criterion is identical to the one in (25b) which was derived by relating the fitting error to the equation error. It should be mentioned here that for the strictly proper case with $p = q + 1$, an analogous derivation appears in [7]. Equation (A.11) clearly shows that if the criterion in (25) is optimized to estimate $\mathbf{b}$ and (A.6) is used to estimate $\mathbf{a}$, then these estimates *are* the optimal solutions for the problem stated in (5). But it may be recalled that the optimal numerator coefficients were found in Subsection III using equation (28), *i.e.*,

$$\mathbf{a}^o = \mathbf{H}_1^o\mathbf{b}^o, \tag{A.12}$$

where, the superscript ^{o} denotes optimized values obtained from (25)-(27). Hence, all that is left is to show that, once $\mathbf{b}^o$ is estimated from (25), the equations (28) and (A.6) produce identical estimates. This is proved next.

Similar to (17), the equation (A.12) can also be rewritten as,

$$\begin{aligned}
\mathbf{a}^o &= \mathbf{B}_u^{T^o}\mathbf{h}^o, && (A.13a)\\
&= \mathbf{B}_u^{T^o}(\mathbf{h}_d - \mathbf{e}^o), \qquad \text{using (27)}, && (A.13b)\\
&= \mathbf{B}_u^{T^o}(\mathbf{h}_d - \mathbf{h}_d + \mathbf{H}_b^{o\prime}\mathbf{H}_b^{o\prime\#}\mathbf{h}_d), \qquad \text{using (A.7a)}, && (A.13c)\\
&= (\mathbf{B}_u^{T^o}\mathbf{H}^{o\prime}{}_b)\mathbf{H}_b^{o\prime\#}\mathbf{h}_d, && (A.13d)\\
&= \mathbf{H}^{o\prime}{}_b^{\#}\mathbf{h}_d, && (A.13e)
\end{aligned}$$

where, the last equality uses the fact that, $\mathbf{B}_u^{T^o}\mathbf{H}^{o\prime}{}_b = \mathbf{I}_{(q+1)}$, which appears in the upper-left partition of (A.8c). This completes the proof of equivalence of both derivations. The derivation given in this Appendix reveals the globally optimal properties of the estimates obtained Subsection III. It may be observed that (28) may be preferred over (A.6) in computing $\mathbf{a}$ because the computation of $\mathbf{h}_b$ and the pseudo-inverse solution can be avoided, whereas, *computation of* $\mathbf{h}^o$ *may be a necessary step.*

It should be emphasized here that the derivation in this Appendix did not make any assumption about the relation between q and p. Hence, the optimal algorithms derived here and the equivalent results in Subsection III are equally applicable for any AR and ARMA cases with arbitrary model orders p and q. The only restriction is that the number of unknowns, $(p + q + 1)$ be less than or equal to the number of observations, N. This can be seen from (10), where the $(q + 1)$ equations in the upper partition are utilized in estimating the $(q + 1)$-length $\mathbf{a}$-vector. In the bottom partition, there are $(N - q - 1)$ equations to solve for the p unknowns in $\mathbf{b}$. Uniqueness of the estimated $\mathbf{b}$ requires $(N - q - 1) \geq p$. This appears to be a powerful result because it substantiates that the applicability of proposed optimal algorithm encompasses a wide range of filter design problems.

APPENDIX B

Computational Algorithm : Phase II

In this appendix the second phase of the iterative algorithm is described in detail. In this phase, the derivative of the matrix $\mathbf{W}$ *w.r.t.* $\hat{\mathbf{b}}$ is taken into consideration while minimizing the fitting error norm. The complete error expression is rewritten below,

$$\|\mathbf{e}(\mathbf{a}^o, \mathbf{b})\|_2^2 = \mathbf{e}^T(\mathbf{a}^o, \mathbf{b})\mathbf{e}(\mathbf{a}^o, \mathbf{b}). \tag{B.1}$$

By setting the derivative of this squared norm to zero, we obtain the updated $\hat{\mathbf{b}}^{(i+1)}$ at the $(i+1)$-th iteration as,

$$\hat{\mathbf{b}}^{(i+1)} = -[\mathbf{U}^{(i)}\mathbf{G}]^{-1}[\mathbf{U}^{(i)}]\mathbf{g} \tag{B.2}$$

where (suppressing the superscript $^{(i)}$),

$$\mathbf{U} \triangleq \mathbf{L}^T\mathbf{W} + \mathbf{G}^T\mathbf{W}^T\mathbf{W}, \tag{B.2a}$$

$$\mathbf{L} \triangleq \left[\frac{\partial \mathbf{W}}{\partial b_1}\mathbf{d}(\mathbf{b}) \;\middle|\; \ldots \middle| \frac{\partial \mathbf{W}}{\partial b_p}\mathbf{d}(\mathbf{b})\right], \tag{B.2b}$$

$$\frac{\partial \mathbf{W}}{\partial b_k} \triangleq \frac{\partial \mathbf{W}}{\partial b_k}, \tag{B.2c}$$

$$\frac{\partial \mathbf{W}}{\partial b_k} \triangleq \frac{\partial}{\partial b_k}[\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}] = \frac{\partial \mathbf{B}}{\partial b_k}(\mathbf{B}^T\mathbf{B})^{-1} - \mathbf{W}\left[\left[\frac{\partial \mathbf{B}^T}{\partial b_k}\right]\mathbf{B} + \mathbf{B}^T\left[\frac{\partial \mathbf{B}}{\partial b_k}\right]\right](\mathbf{B}^T\mathbf{B})^{-1} \quad \text{and} \tag{B.2d}$$

$\frac{\partial \mathbf{B}}{\partial b_k}$ has the same form as the $\mathbf{B}$ matrix defined in (16a) but filled with all zeros except at the locations where b_k appear. For example,

$$\frac{\partial \mathbf{B}}{\partial b_p} = \begin{bmatrix} 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 0 \\ 0 & \ddots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \cdots & 0 & 0 & \ddots & 1 \\ 0 & \cdots & 0 & 0 & \ddots & 0 \\ \vdots & \cdots & \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{N \times N-q} \tag{B.2e}$$

Once $\hat{\mathbf{b}}^{(i+1)}$ is found, $\mathbf{b}^{(i+1)}$ can be formed as,

$$\mathbf{b}^{(i+1)} = \begin{bmatrix} 1 \\ \cdots\cdots \\ \hat{\mathbf{b}}^{(i+1)} \end{bmatrix} \tag{B.3a}$$

$$= \begin{bmatrix} 1 \\ \cdots\cdots\cdots\cdots\cdots \\ -[\mathbf{U}^{(i)}\mathbf{G}]^{-1}[\mathbf{U}^{(i)}]\mathbf{g} \end{bmatrix}. \tag{B.3b}$$

This minimization phase continues until $\mathbf{b}^{i+1} \simeq \mathbf{b}^i$ is reached and this optimum $\mathbf{b}^*$ vector corresponds to a minimum of the error surface of $\|\mathbf{e}(\mathbf{a}^*, \mathbf{b})\|_2^2$.

Iteration Phases	Minimized error norm	Closeness with $\mathbf{h_t}$, in dB	Number of Iterations
Phase 1	0.5948	30.5173	6
Phase 2	0.5735	30.6740	3

Table. 1 : Results of Simulation 1. A triangular Impulse response is modeled by an ARMA model with $p = 7$ and $q = 4$. $\delta = 10^{-3}$ was used as the stopping criterion.

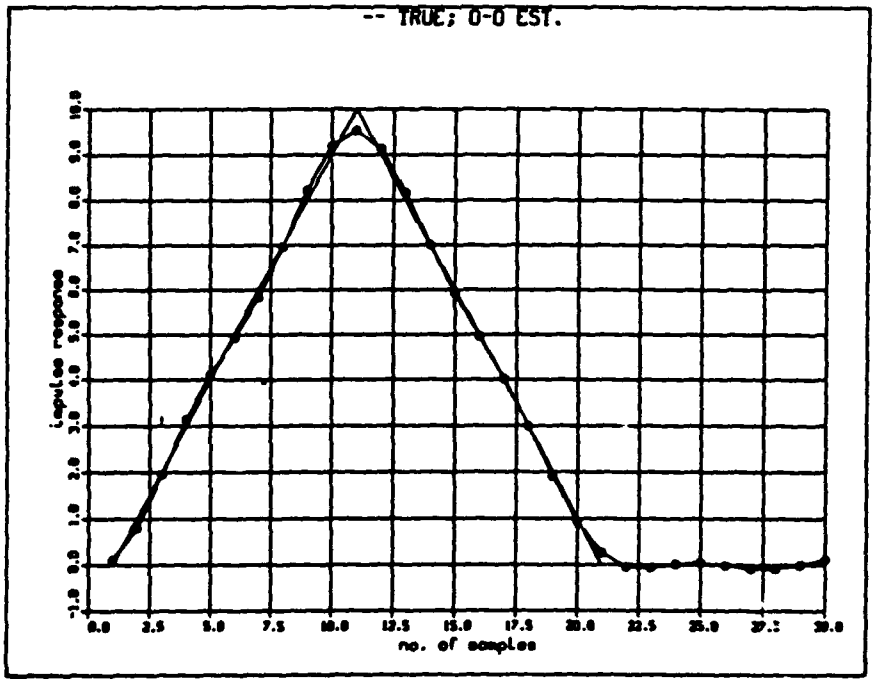

Fig. 1a : Simulation 1 :- A triangular Impulse response is modeled by an ARMA(7,4) model. $\delta = 10^{-3}$. Result shows fit after Phase-1 convergence.

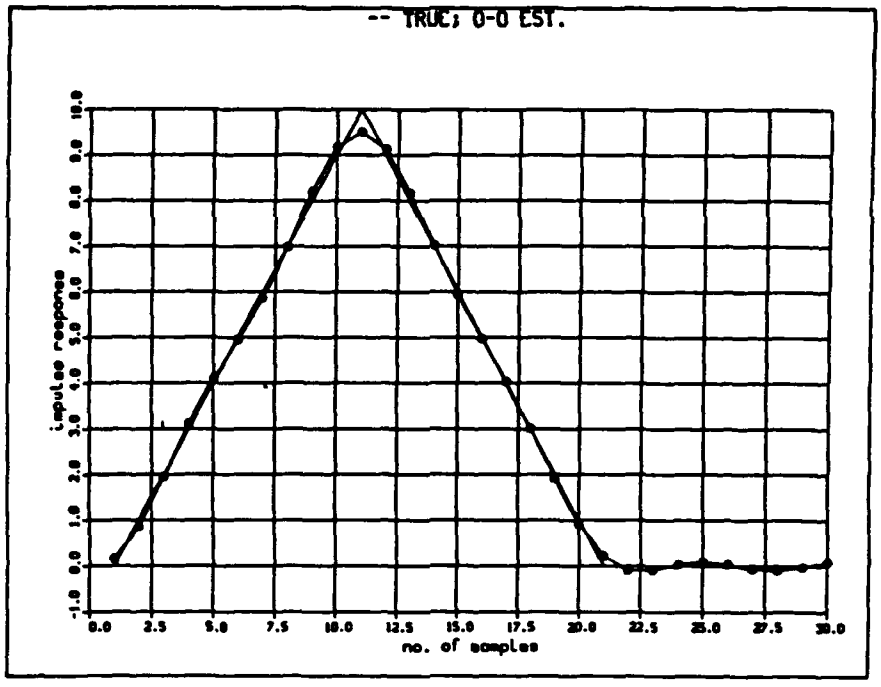

Fig. 1b : Simulation 1 :- A triangular Impulse response is modeled by an ARMA(7,4) model. $\delta = 10^{-3}$. Result shows fit after Phase-2 convergence.

Iteration Phases	Minimized error norm	Closeness with $\mathbf{h}_t$, in dB	Closeness with $\mathbf{h}_d$, in dB	Number of Iterations
Phase 1	0.010177	25.8552	22.6084	1
Phase 2	0.010079	25.30777	22.05049	2

Table 2a : Results of Simulation 2. 30dB perturbation was added to a true ARMA(8,4) impulse response ($\mathbf{h}_t$) to form the desired response ($\mathbf{h}_d$). $\delta = 10^{-3}$ was used as stopping criterion.

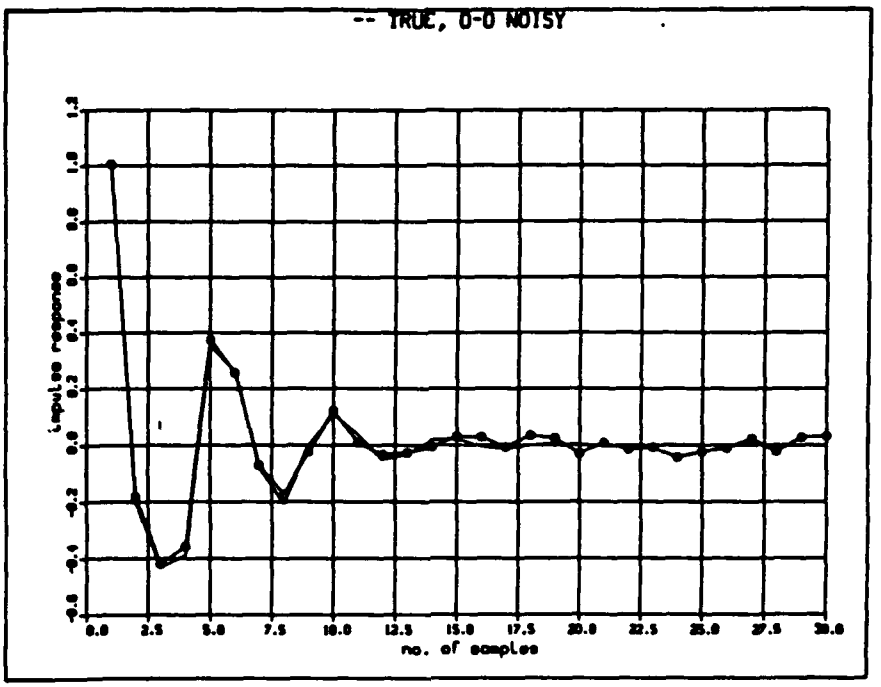

Fig. 2a : Simulation 2 :- True and 30dB perturbed ARMA(8,4) impulse responses

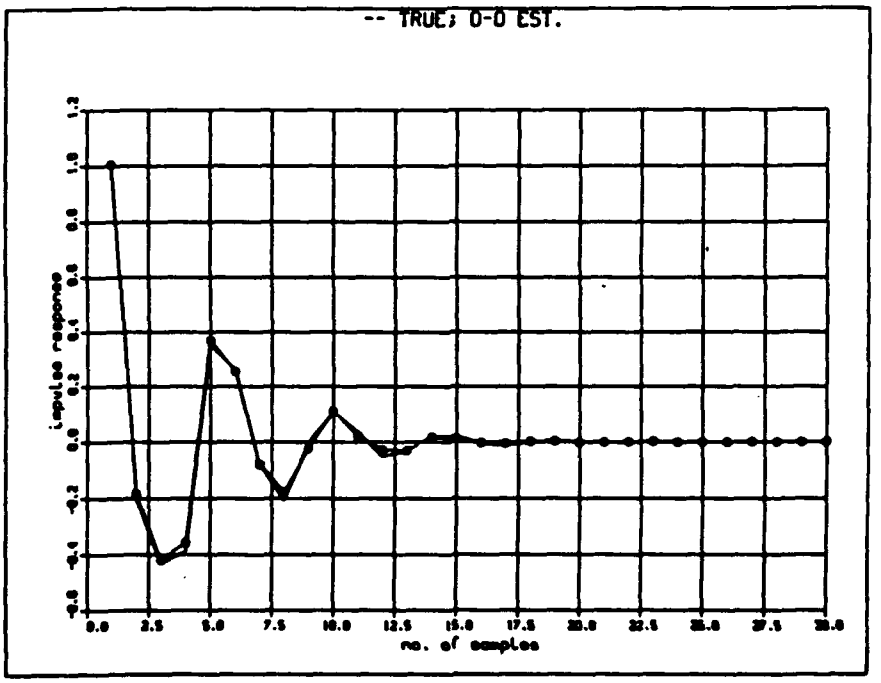

Fig. 2b : Simulation 2 :- ARMA(8,4) True and estimated impulse responses after Phase 1 convergence. SNR=30dB, $\delta = 10^{-3}$.

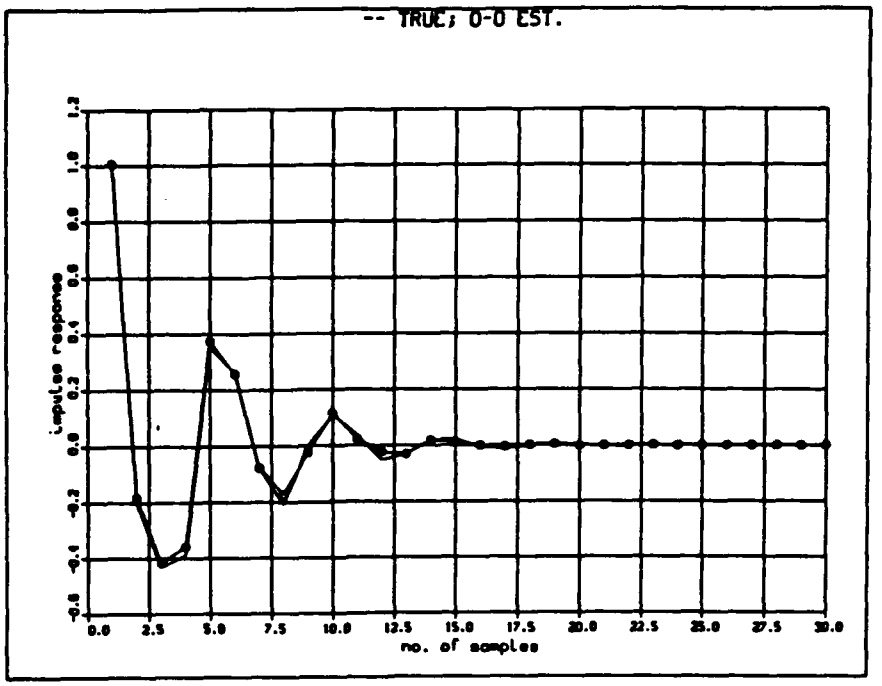

Fig. 2c : Simulation 2 :- ARMA(8,4) True and estimated impulse responses after Phase 2 convergence. SNR=30dB, $\delta = 10^{-3}$.

Iteration Phases	Minimized error norm	Closeness with h_t, in dB	Closeness with h_d, in dB	Number of Iterations
Phase 1	0.1052	16.91	12.00	1
Phase 2	0.1010	15.22	12.178	2

Table 2b : Results of Simulation 2. 20dB perturbation was added to a true ARMA(8,4) impulse response (h_t) to form the desired response (h_d). $\delta = 10^{-3}$ was used as stopping criterion.

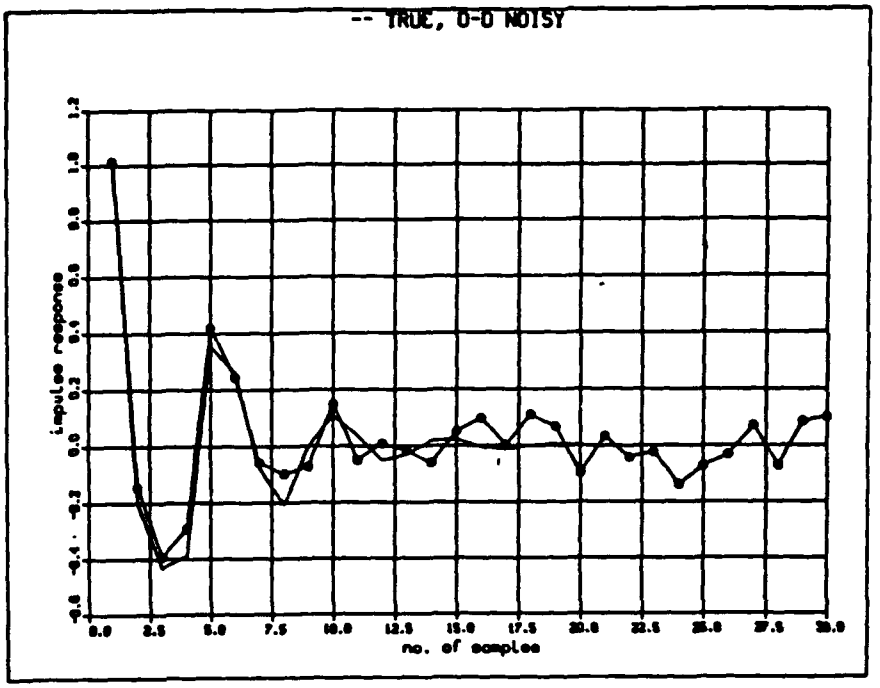

Fig. 2d : Simulation 2 :- True and 20dB perturbed ARMA(8,4) impulse responses

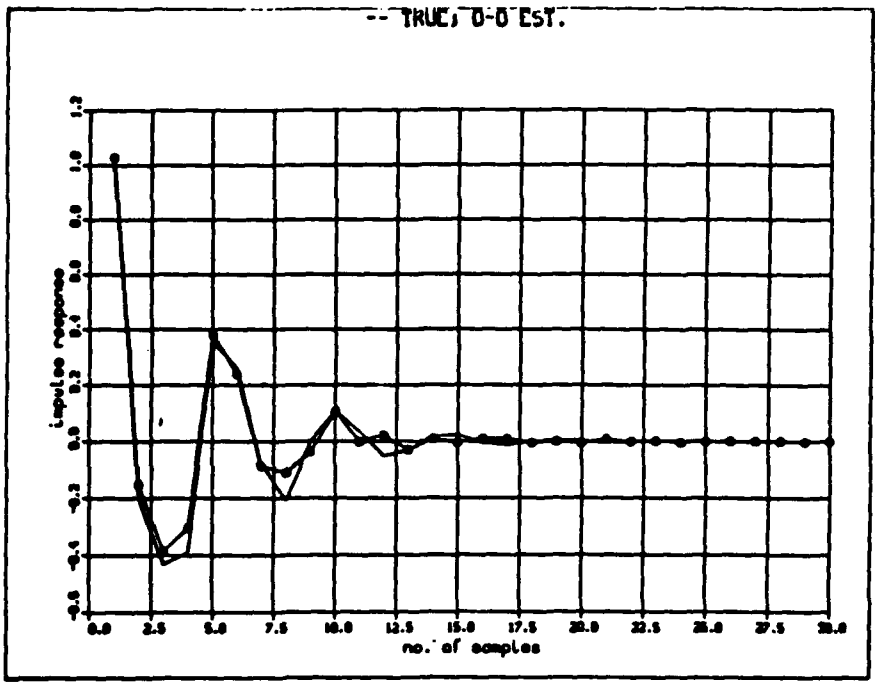

Fig. 2e : Simulation 2 :- ARMA(8,4) True and estimated impulse responses after Phase 1 convergence. SNR=20dB, $\delta = 10^{-3}$.

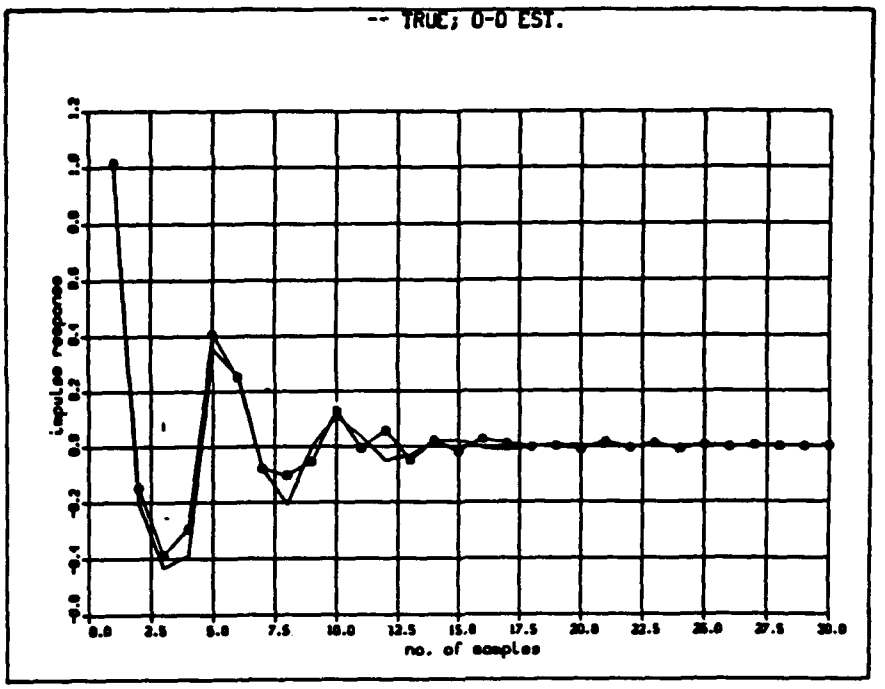

Fig. 2f : Simulation 2 :- ARMA(8,4) True and estimated impulse responses after Phase 2 convergence. SNR=20dB, $\delta = 10^{-3}$.

Iteration Phases	Minimized error norm	Closeness with $\mathbf{h}_t$, in dB	Closeness with $\mathbf{h}_d$, in dB	Number of Iterations
Phase 1	0.008752	24.837	22.8754	5
Phase 2	0.008399	23.869	23.054	3

Table 3 : Results of Simulation 3. 30dB perturbation was added to a true ARMA(4,8) impulse response ($\mathbf{h}_t$) to form the desired response ($\mathbf{h}_d$). $\delta = 10^{-3}$ was used as stopping criterion.

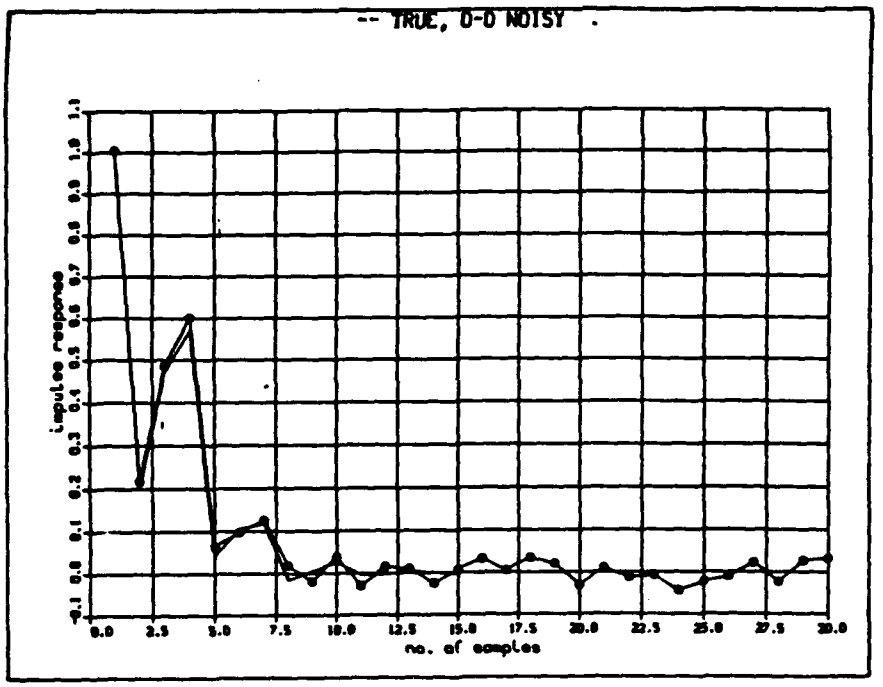

Fig. 3a : Simulation 3 :- True and 30dB perturbed ARMA(4,8) impulse responses

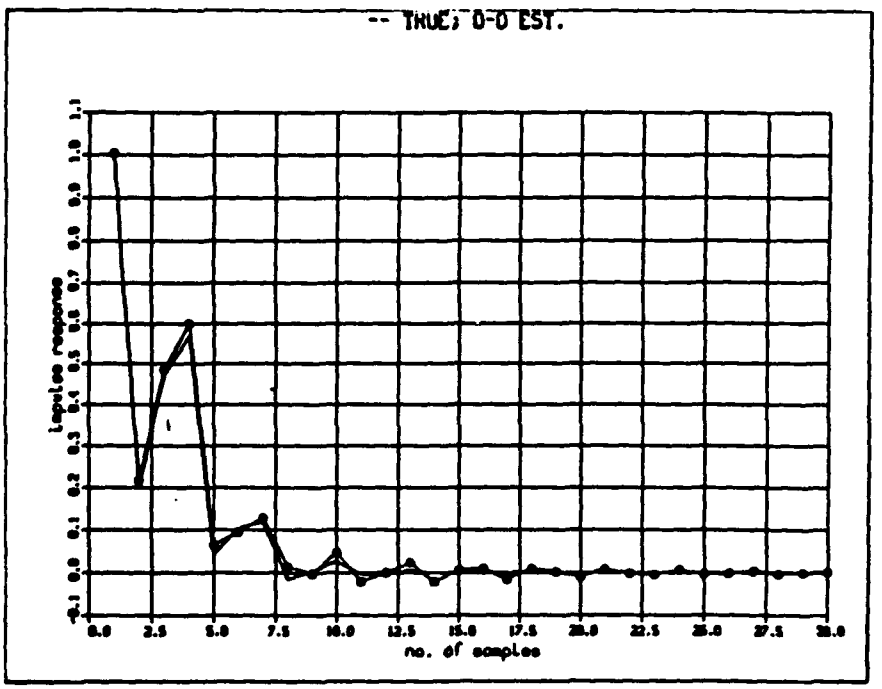

Fig. 3b : Simulation 3 :- ARMA(4,8) True and estimated impulse responses after Phase 1 convergence. SNR=30dB, $\delta = 10^{-3}$.

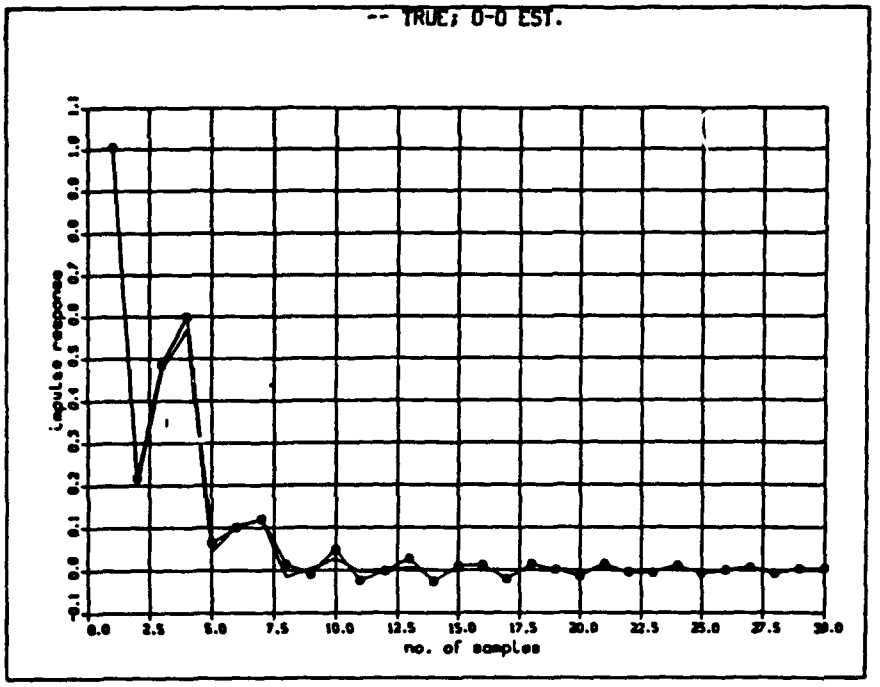

Fig. 3c : Simulation 3 :- ARMA(4,8) True and estimated impulse responses after Phase 2 convergence. SNR=30dB, $\delta = 10^{-3}$.

Iteration Phases	Minimized error norm	Closeness with $\mathbf{h_t}$, in dB	Number of Iterations
Phase 1	3.0368	23.4366	4
Phase 2	3.0348	23.4395	3

Table. 4 : Results of Simulation 4. A triangular Impulse response is modeled by an AR model with $p = 5$. $\delta = 10^{-3}$ was used as the stopping criterion.

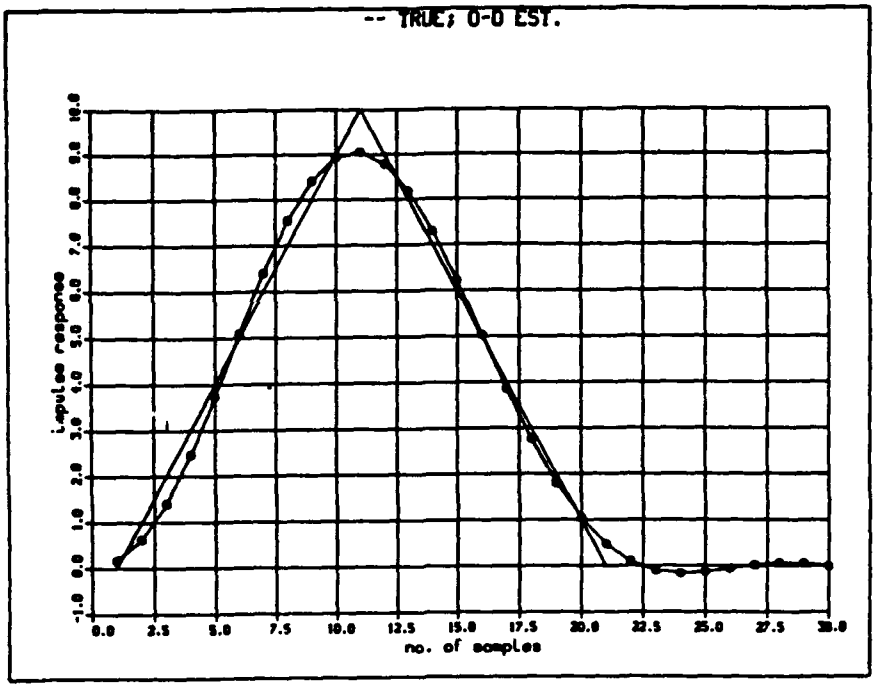

Fig. 4a : Simulation 4 :- A triangular Impulse response is modeled by an AR(5) model. $\delta = 10^{-3}$. Result shows fit after Phase-1 convergence.

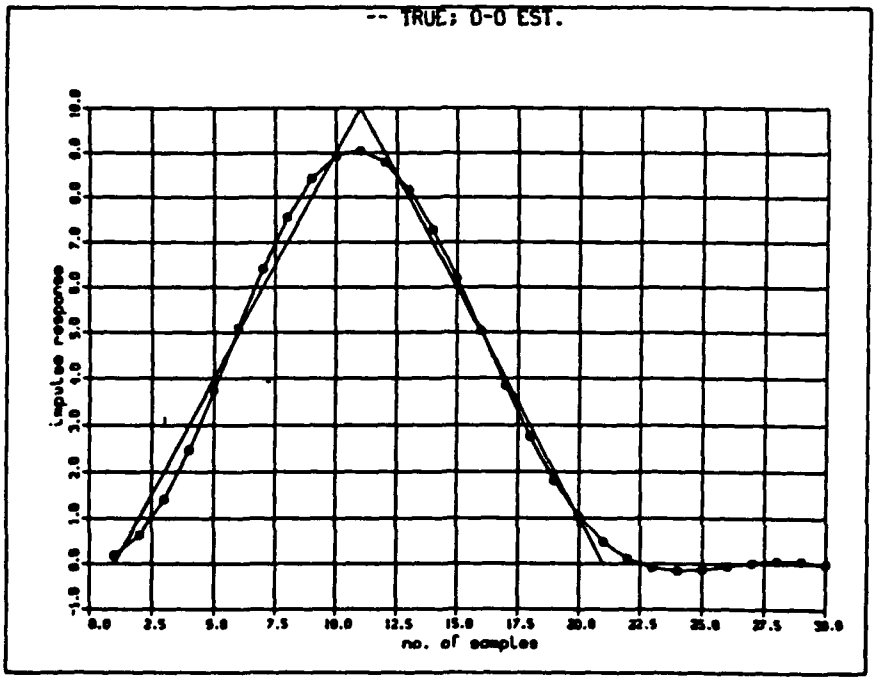

Fig. 4b : Simulation 4 :- A triangular Impulse response is modeled by an AR(5) model. $\delta = 10^{-3}$. Result shows fit after Phase-2 convergence.

SECTION 2.3 : DESIGN OF 2-D RECURSIVE DIGITAL FILTERS IN THE SPATIAL DOMAIN

SUMMARY

A class of least-squares algorithms for design of two-dimensional digital filters from space domain data is presented. The proposed algorithms iteratively estimates the filter coefficients by minimizing the true squared error between the given and the estimated space domain responses. The algorithms are essentially generalization of the optimal 1-D design algorithm given by Evans and Fischl [7]. The error criterion is simultaneously optimized *w.r.t.* the coefficients in both dimensions. Design algorithms are given for filters with separable and irreducible numerator/denominator polynomials and also for mixed structures. The effectiveness of the algorithms is illustrated with several simulation examples.

I. Introduction

Given the space domain unit sample response (2-D impulse response) data, an important design problem is to obtain the optimal coefficients for the two-dimensional infinite impulse response (IIR) filters. This problem has received considerable attention in the recent years *e.g.*, see [1-4,10]. The relationship between the time domain impulse response data and corresponding 1-D recursive filter parameters are well established. Analogous relationships can also be established between the space domain data and the corresponding 2-D recursive filter parameters. Exploiting these similarities, many 2-D filter synthesis algorithms have been developed by modification and extension of existing algorithms for 1-D filter design. Specifically, Shanks *et al* [1] extended the work of Shanks [5]; Cadzow [2] and Shaw and Mersereau [3] utilized many of the general non-linear optimization methods; and Shaw and Mersereau [3] also extended the work of Steiglitz and McBride [6]. But these methods are *suboptimal* in the sense that they do not optimize the exact error criterion. In contrast to these approaches, the iterative method (EFM) proposed by Evans and Fischl [7] is *optimal* and it does optimize the true error criterion. Recently some work [10] has been done on the 2-D extension of EFM for spatial domain design, but we believe that the full potential of EFM has not been utilized for 2-D recursive filter design. The main drawbacks of the approach in [10] is that the complete error criterion was not optimized and the second phase of EFM was not invoked. Also, the error criterion was not optimized *w.r.t.* the filter coefficients in two domains simultaneously.

The Evans-Fischl method is extremely effective in 1-D filter design. A modified complex version of the EFM with certain symmetry constraints has recently been shown to be equally effective for maximum-likelihood 1-D and 2-D frequency-wavenumber estimation [8,9]. In this work we develop a 2-D version of the EFM in order to establish a general framework for *optimal* and *sub-optimal design* of 2-D recursive filters from the given space domain data.

This Section is arranged as follows : In Subsection II, the least-squares problem is formulated. In Subsection III, the case with separable numerator and separable denominator polynomials is considered in detail and the error criterion is related to the case with irreducible numerator. In Subsection IV, the irreducible case is outlined briefly. In Subsection V, simulation results are provided to illustrate the effectiveness of the proposed algorithm. Finally, in Subsection VI some concluding remarks are given.

II. Formulation of the 2-D Least-Squares Synthesis Problem

In the general form, a 2-D rational function $H(z_1, z_2)$, with non-decomposable numerator and denominator polynomials is described by:

$$H(z_1, z_2) = \frac{Q(z_1, z_2)}{P(z_1, z_2)} = \frac{\sum_{i=0}^{n_1} \sum_{j=0}^{n_2} q(i,j) z_1^{-i} z_2^{-j}}{\sum_{i=0}^{m_1} \sum_{j=0}^{m_2} p(i,j) z_1^{-i} z_2^{-j}}. \quad (1)$$

Equivalently, $H(z_1, z_2)$ may also be written as,

$$H(z_1, z_2) = \mathbf{z}_1^T \mathbf{H} \mathbf{z}_2 \tag{2}$$

where, $\mathbf{z}_1 \triangleq [1 \; z_1^{-1} \cdots z_1^{-(k_1-1)}]^T$, $\mathbf{z}_2 \triangleq [1 \; z_2^{-1} \cdots z_2^{-(k_2-1)}]^T$ and

$$\mathbf{H} \triangleq \begin{bmatrix} h(0,0) & h(0,1) & \cdots & h(0,k_2-1) \\ h(1,0) & h(1,1) & \cdots & h(1,k_2-1) \\ \vdots & \vdots & \ddots & \vdots \\ h(k_1-1,0) & h(k_1-1,1) & \cdots & h(k_1-1,k_2-1) \end{bmatrix}. \tag{3}$$

In (3), $\mathbf{H}$ contains the $k_1 \times k_2$ significant impulse response values. By stacking the columns of $\mathbf{H}$, the impulse response may be represented in vector form as,

$$\mathbf{h}_c \triangleq [\mathbf{h}_1^T \; \mathbf{h}_2^T \; \cdots \; \mathbf{h}_{k_2}^T]^T \tag{4}$$

where, $\mathbf{h}_i$ denotes the i^{th} column of $\mathbf{H}$. Next, let the *given* space-domain impulse response matrix be,

$$\mathbf{X} \triangleq \begin{bmatrix} x(0,0) & x(0,1) & \cdots & x(0,k_2-1) \\ x(1,0) & x(1,1) & \cdots & x(1,k_2-1) \\ \vdots & \vdots & \ddots & \vdots \\ x(k_1-1,0) & x(k_1-1,1) & \cdots & x(k_1-1,k_2-1) \end{bmatrix} \tag{5}$$

and the corresponding vector be:

$$\mathbf{x}_c \triangleq [\mathbf{x}_1^T \; \mathbf{x}_2^T \; \cdots \; \mathbf{x}_{k_2}^T]^T. \tag{6}$$

In this work, we address the 2-D least-squares space-domain synthesis problem as stated below :

> Given the space-domain 2-D impulse response matrix $\mathbf{X}$, estimate $\mathbf{q} \triangleq [q(0,0) \; q(0,1) \cdots q(n_1,n_2)]^T$ and $\mathbf{p} \triangleq [p(0,0) \; p(0,1) \cdots p(m_1,m_2)]^T$ by optimizing the following ℓ_2-norm based error criterion,
>
> $$\min_{\mathbf{q},\mathbf{p}} \|\mathbf{e}\|^2 \triangleq \|\mathbf{x}_c - \mathbf{h}_c\|^2 \qquad \text{with } p(0,0) = 1. \tag{7}$$

Next, we formulate and outline the optimization procedures for several special cases of interest.

III. Design With Separable Denominator and Separable/Irreducible Numerator

In this Subsection, we will first perform the derivation for the separable numerator case and then relate our results to the irreducible numerator case. It may be noted that the optimal separable denominator polynomials are same for both cases.

The 2-D rational transfer function with *separable* numerator and denominator polynomials can be written as,

$$H(z_1, z_2) = \frac{\sum_{i=0}^{n_1} a(i) z_1^{-i} \sum_{j=0}^{n_2} b(j) z_2^{-j}}{\sum_{i=0}^{m_1} c(i) z_1^{-i} \sum_{j=0}^{m_2} d(j) z_2^{-j}}. \tag{8}$$

Note that for strictly proper case, $m_1 = n_1 + 1$ and $m_2 = n_2 + 1$. The numerator and denominator coefficients in equations (1) and (8) are related as follows:

$$q(i,j) = a(i)b(j) \qquad \text{for } 0 \le i \le n_1, \quad 0 \le j \le n_2 \tag{9a}$$

$$p(i,j) = c(i)d(j) \qquad \text{for } 0 \le i \le m_1, \quad 0 \le j \le m_2 \tag{9b}$$

The space-domain transfer function $H(z_1, z_2)$ can also be written in the separable form as,

$$H(z_1, z_2) = \mathbf{z}_1^T \mathbf{H} \mathbf{z}_2 = \mathbf{z}_1^T \mathbf{f} \mathbf{g}^T \mathbf{z}_2 = \sum_{i=0}^{k_1-1} f(i) z_1^{-i} \sum_{j=0}^{k_2-1} g(j) z_2^{-j} \quad (10)$$

where, $\mathbf{f} \triangleq [f(0)\ f(1) \cdots f(k_1-1)]^T$,

$$\mathbf{z}_1^T \mathbf{f} = \sum_{i=0}^{k_1-1} f(i) z_1^{-i} \triangleq \frac{\sum_{i=0}^{n_1} a(i) z_1^{-i}}{\sum_{i=0}^{m_1} c(i) z_1^{-i}} \quad (11a)$$

and

$$\mathbf{g}^T \mathbf{z}_2 = \sum_{j=0}^{k_2-1} g(j) z_2^{-j} \triangleq \frac{\sum_{j=0}^{n_2} b(j) z_2^{-j}}{\sum_{j=0}^{m_2} d(j) z_2^{-j}}. \quad (11b)$$

From (10), the 2-D space-domain impulse response values in (3) are related to the separable 1-D impulse response values as,

$$h(i,j) = f(i)g(j), \qquad \text{for} \quad 0 \le i \le k_1 - 1, 0 \le j \le k_2 - 1 \quad (11c)$$

From (11a), the transfer function coefficients in $\mathbf{a} \triangleq [a(0)\ a(1) \cdots a(n_1)]^T$ and $\mathbf{c} \triangleq [c(0)\ c(1) \cdots c(m_1)]^T$ are related to the impulse response values in $\mathbf{f}$ as [7]

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \end{bmatrix} \mathbf{c} \quad (12)$$

where,

$$\mathbf{F}_1 \triangleq \begin{bmatrix} f(0) & 0 & \cdots & 0 & 0 \\ f(1) & f(0) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f(m_1-1) & f(m_1-2) & \cdots & f(0) & 0 \end{bmatrix} \quad \text{and} \quad (13a)$$

$$\mathbf{F}_2 \triangleq \begin{bmatrix} f(m_1) & f(m_1-1) & \cdots & f(1) & f(0) \\ f(m_1+1) & f(m_1) & \cdots & f(2) & f(1) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ f(k_1-1) & f(k_1-2) & \cdots & f(k_1-m_1) & f(k_1-m_1-1) \end{bmatrix}. \quad (13b)$$

Note that the lower partition of (12) can also be written as,

$$\mathbf{0} = \mathbf{F}_2 \mathbf{c} = \mathbf{C}^T \mathbf{f} \quad (14)$$

where, $\mathbf{C}$ is $k_1 \times (k_1 - m_1)$ banded Toeplitz matrix defined as,

$$\mathbf{C} \triangleq \begin{bmatrix} c(m_1) & 0 & \cdots & 0 \\ c(m_1-1) & c(m_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ c(0) & c(1) & \cdots & c(m_1) \\ 0 & c(0) & \cdots & c(m_1-1) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c(0) \end{bmatrix}. \quad (15)$$

Proceeding in a similar manner from (11*b*), we get the relationships between $\mathbf{b} \triangleq [b(0)\ b(1) \cdots b(n_2)]^T$, $\mathbf{d} \triangleq [d(0)\ d(1) \cdots d(m_2)]^T$ and the impulse response values in $\mathbf{g}$ as

$$\begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \end{bmatrix} \mathbf{d}. \tag{16}$$

Again, the lower partition of (16) can be written as,

$$\mathbf{0} = \mathbf{G}_2\mathbf{d} = \mathbf{D}^T\mathbf{g}, \tag{17}$$

where, $\mathbf{G}_1$, $\mathbf{G}_2$ and $\mathbf{D}$ are defined in a manner similar to $\mathbf{F}_1$, $\mathbf{F}_2$ and $\mathbf{C}$, respectively, with appropriate dimensions. Now, for the 2-D transfer function $H(z_1, z_2)$ as defined in equations (8)-(10), the coefficients are related to the 2-D impulse response values as follows:

$$\begin{aligned} \begin{bmatrix} \mathbf{a} \otimes \mathbf{b} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} &= \begin{bmatrix} \mathbf{F}_1 \otimes \mathbf{G}_1 \\ \mathbf{F}_1 \otimes \mathbf{G}_2 \\ \mathbf{F}_2 \otimes \mathbf{G}_1 \\ \mathbf{F}_2 \otimes \mathbf{G}_2 \end{bmatrix} [\mathbf{c} \otimes \mathbf{d}] \\ &= \begin{bmatrix} \mathbf{C}_1^T \otimes \mathbf{D}_1^T \\ \mathbf{C}_1^T \otimes \mathbf{D}^T \\ \mathbf{C}^T \otimes \mathbf{D}_1^T \\ \mathbf{C}^T \otimes \mathbf{D}^T \end{bmatrix} [\mathbf{f} \otimes \mathbf{g}] = \begin{bmatrix} \mathbf{C}_1^T \otimes \mathbf{D}_1^T \\ \mathbf{C}_1^T \otimes \mathbf{D}^T \\ \mathbf{C}^T \otimes \mathbf{D}_1^T \\ \mathbf{C}^T \otimes \mathbf{D}^T \end{bmatrix} \mathbf{h}_c \end{aligned} \tag{18}$$

where, $\otimes$ denotes the Kronecker product. Following (9a), for the case of irreducible numerator design, the terms $a(i)b(j)$, for $i = 1, \ldots, n_1$ and $j = 1, \ldots, m_1$ in $\mathbf{a} \otimes \mathbf{b}$ on the left hand side should be replaced by $q(i, j)$. Also in (18), $\mathbf{C}_1$ and $\mathbf{C}_2$ become $k_1 \times m_1$ and $k_2 \times m_2$ matrices, respectively, defined as,

$$\mathbf{C}_1 \triangleq \begin{bmatrix} c(0) & c(1) & \cdots & c(m_1-1) \\ 0 & c(0) & \cdots & c(m_1-2) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & c(0) \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \tag{19a}$$

$$\mathbf{D}_1 \triangleq \begin{bmatrix} d(0) & d(1) & \cdots & d(m_2-1) \\ 0 & d(0) & \cdots & d(m_2-2) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & d(0) \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{19b}$$

and using (4) and (11c),

$$\mathbf{h}_c = \mathbf{f} \otimes \mathbf{g}. \tag{20}$$

Hence, the three bottom partitions of (18) can be expressed as

$$\begin{bmatrix} \mathbf{C}_1^T \otimes \mathbf{D}^T \\ \mathbf{C}^T \otimes \mathbf{D}_1^T \\ \mathbf{C}^T \otimes \mathbf{D}^T \end{bmatrix} \mathbf{h}_c = \mathbf{0}. \tag{21}$$

Equation (21) essentially implies that the matrices $\mathbf{C}_1 \otimes \mathbf{D}$, $\mathbf{C} \otimes \mathbf{D}_1$ and $\mathbf{C} \otimes \mathbf{D}$ are orthogonal to $\mathbf{h}_c$. Moreover, by construction, these matrices contain $k_1k_2 - m_1m_2$ linearly independent vectors of length k_1k_2. The remaining m_1m_2 linearly independent vectors, that complete the entire k_1k_2 dimensional vector space, are in $\mathbf{C}_1 \otimes \mathbf{D}_1$ as may be seen by rewriting the upper partition of equation (18) as follows,

$$[\mathbf{a} \otimes \mathbf{b}] = [\mathbf{F}_1 \otimes \mathbf{G}_1][\mathbf{c} \otimes \mathbf{d}] \tag{22a}$$

$$\text{or,} \quad [\mathbf{a} \otimes \mathbf{b}] = [\mathbf{C}_1^T \otimes \mathbf{D}_1^T][\mathbf{f} \otimes \mathbf{g}] \tag{22b}$$

$$= [\mathbf{C}_1^T \otimes \mathbf{D}_1^T]\mathbf{h}_c \tag{22c}$$

Therefore, the search for the coefficient vectors $\mathbf{c}$ and $\mathbf{d}$ is equivalent to the search for $k_1k_2 - m_1m_2$ linearly independent vectors which are orthogonal to the impulse response vector $\mathbf{h}_c$. In practice, we do not have $\mathbf{h}_c$ and if we replace $\mathbf{h}_c$ by the given space domain impulse response values $\mathbf{x}_c$ as defined in (6), the r.h.s. of (21) will not be equal to zero and there will be some equation error which we define as $\hat{\mathbf{e}}$, *i.e.*,

$$\begin{bmatrix} \mathbf{C}_1^T \otimes \mathbf{D}^T \\ \mathbf{C}^T \otimes \mathbf{D}_1^T \\ \mathbf{C}^T \otimes \mathbf{D}^T \end{bmatrix} \mathbf{x}_c = \hat{\mathbf{e}}. \tag{23}$$

But from (7), $\mathbf{x}_c = \mathbf{h}_c + \mathbf{e}$ and hence,

$$\begin{bmatrix} \mathbf{C}_1^T \otimes \mathbf{D}^T \\ \mathbf{C}^T \otimes \mathbf{D}_1^T \\ \mathbf{C}^T \otimes \mathbf{D}^T \end{bmatrix} [\mathbf{h}_c + \mathbf{e}] = \hat{\mathbf{e}}. \tag{24}$$

Substituting (21) in (24) we see that the *fitting error* $\mathbf{e}$ in (7) can be related to the equation error $\hat{\mathbf{e}}$ as,

$$\begin{bmatrix} \mathbf{C}_1^T \otimes \mathbf{D}^T \\ \mathbf{C}^T \otimes \mathbf{D}_1^T \\ \mathbf{C}^T \otimes \mathbf{D}^T \end{bmatrix} \mathbf{e} = \hat{\mathbf{e}}. \tag{25}$$

Following the EFM, the estimation of the numerator and denominator coefficients are now separated into two parts. If the denominator coefficients are known then the numerator coefficients may be estimated from (22c) by replacing $\mathbf{h}_c$ by $\mathbf{x}_c$. It may be noted here that a direct implementation of (22c) will give numerator coefficients $q(i,j)$'s which are the coefficients of the optimum irreducible numerator polynomial. This polynomial, in general, will not be separable. Hence, to obtain the separable form, we have to find (using SVD) the rank-1 approximation of the matrix $\mathbf{Q}$ formed with the elements $q(i,j)$. Except for the scale factor which is the largest singular value, the first column and row singular vectors will contain the coefficients of the separable numerator polynomials, $\mathbf{a}$ and $\mathbf{b}$, respectively.

For determining the optimal separable denominator polynomials, we have to make use of certain orthogonality conditions [8,9] in order to show that the minimization of $\|\mathbf{e}\|^2$ is exactly equivalent to optimizing the following criterion,

$$\min_{\mathbf{c},\mathbf{d}} (\mathbf{x}_c^T((\mathbf{I}_{k_1} \otimes \mathcal{P}_\mathbf{D}) + (\mathcal{P}_\mathbf{C} \otimes \mathbf{I}_{k_2}) - (\mathcal{P}_\mathbf{C} \otimes \mathcal{P}_\mathbf{D}))\mathbf{x}_c) \tag{26}$$

where, $\mathbf{I}_{k_1} \in \mathbb{R}^{k_1 \times k_1}$ and $\mathbf{I}_{k_2} \in \mathbb{R}^{k_2 \times k_2}$ are identity matrices and

$$\mathcal{P}_\mathbf{D} \triangleq \mathbf{D}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T \quad \text{and} \tag{27a}$$

$$\mathcal{P}_\mathbf{C} \triangleq \mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T \tag{27b}$$

are projection matrices.

For complex data and complex coefficients, an iterative algorithm for optimization of a criterion very similar to the one in (26) but with certain symmetry constraints was developed in [8] and [9]. Using derivations similar to the ones in [8] and [9] we can show that, similar to EFM, a quasi-linear relationship can be established between the error and the polynomial coefficients in **c** and **d** which enables us to optimize the criterion iteratively without resorting to any general optimization algorithms.

It may be pointed out here that in [10-12], only the first term in (26) was minimized separately with one set of denominator coefficients. Hence, the estimates of the denominator coefficients are suboptimal because they were not obtained by optimizing the true error criterion. Also, following [8,9] and unlike [10], the optimization of (26) can be carried out *w.r.t.* both sets of parameters **c** and **d** simultaneously. Details are omitted and will be published elsewhere. In the next Subsection, we generalize our result to the irreducible case.

IV. Optimal Design with Irreducible Numerator and Denominator

Rewriting (1) as,

$$H(z_1, z_2)P(z_1, z_2) \;=\; Q(z_1, z_2) \tag{28}$$

and equating the coefficients of like powers of $z_1^{-i}z_2^{-j}$, $\forall\ i, j$, we get the following relationship between the coefficients and the space-domain impulse response,

$$\begin{bmatrix} \mathbf{q} \\ \mathbf{0} \end{bmatrix} \;=\; \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \end{bmatrix} \mathbf{p}. \tag{29}$$

This relationship is equivalent to the one derived in (18) for the separable case though it should be emphasized that the coefficients and the space-domain response are not separable in the present case. Specifically, the numerator coefficients $q(i,j)$'s replace $a(i)b(j)$'s in the l.h.s. of (18) and $h(i,j)$'s and $p(i,j)$'s replace $f(i)g(j)$'s and $c(i)d(j)$'s, respectively, in the r.h.s of (18). Furthermore, $\mathbf{H}_1$ replaces $\mathbf{F}_1 \otimes \mathbf{G}_1$ in the uppermost partition of (18) and $\mathbf{H}_2$ replaces the rest of the three lower partitions of (18). Once again, the problem will be divided into two optimization problems. First, the denominator coefficients will be found by optimizing,

$$\min_{\mathbf{p}} \|\mathbf{p}^T \mathbf{X}_c (\mathbf{P}^T \mathbf{P})^{-1} \mathbf{X}_c \mathbf{p}\|^2, \tag{30}$$

where,

$$\mathbf{X}_c \;\underline{\underline{\Delta}}\; \left[\mathbf{X}_1^T\ \mathbf{X}_2^T\ \cdots\ \mathbf{X}_{k_2}^T\right]^T \tag{31}$$

and each $\mathbf{X}_i$ is formed using the elements in the i^{th} column of $\mathbf{X}$ as

$$\mathbf{X}_i(l,k) \;\underline{\underline{\Delta}}\; \mathbf{x}_i(m_1 m_2 + m_1 + m_2 + l - k + 1) \quad \text{for } i \;=\; 1, 2, \ldots, k_2. \tag{32}$$

Once **p** is found the numerator coefficients are found from the upper partition of (30).

V. Simulation Results

In this Subsection, we present two numerical examples to illustrate the effectiveness of the proposed algorithms for design of denominator separable filters.

Example 1: *Quarter-Plane Gaussian Filter*

Consider a "Gaussian filter" whose impulse response, defined over the first quadrant, is given by

$$H(i,j) = 0.256322\ exp\left[-0.103203\{(i-4)^2 + (j-4)^2\}\right],$$

where $(i,j) \in S_f$ and the support S_f is given by

$$S_f = \{(i,j) \mid 0 \leq i \leq N;\ 0 \leq j \leq M\}.$$

With $N = M = 10$ and using the technique proposed above, 2-D linear shift-invariant causal (m_1, m_1)-th order filter with the transfer function

$$H(z_1, z_2) = \frac{Q(z_1, z_2)}{C(z_1)D(z_2)}$$

where $Q(z_1, z_2) = \sum_{i=0}^{n_1} \sum_{j=0}^{n_1} q(i,j) z_1^{-i} z_2^{-j}$, $C(z_1) = 1 + c(1)z_1^{-1} + \cdots + c(m_1)z_1^{-m_1}$ and $D(z_2) = 1 + d(1)z_2^{-1} + \cdots + d(m_1)z_2^{-m_1}$, is designed to approximate the impulse response of the Gaussian filter. Note that the results presented below are for "strictly proper" filter, *i.e.* $n_1 = m_1 - 1$. We have compared the results obtained from our algorithm with those obtained using Hinamoto and Maekawa [12] for different order realizations.

Table 5.1: Example 1: Comparison of error norms

Order	Method in [12]	Proposed Method
2	7.0083E-01	6.5181E-01
3	3.6349E-01	2.1721E-01
4	1.7867E-01	2.5051E-02
5	1.0192E-01	4.5056E-04

Example 2: *Ideal Circular Low-Pass Filter*

The impulse response was generated as

$$H(i,j) = \frac{r J_1(r\sqrt{i^2 + j^2})}{2\pi\sqrt{i^2 + j^2}}$$

where $J_1(\cdot)$ denotes the Bessel function of the first kind with order one with $r = 3.0$. The support region is again $i, j = 1, \ldots, 11$. For this example, we designed a "strictly proper" as well as "proper" filter for various orders. Note that although the proper design is *sub-optimal*, it gives better results than the optimal strictly proper design. This may be due to the fact that we are able to match greater number of impulse response samples exactly. A comparison of the error norm is given in Table 5.2.

Table 5.2: Example 2: Error norms

Order	Strictly Proper	Proper
2	8.7947E-02	6.6811E-02
3	1.1528E-01	6.3363E-02
4	6.6971E-02	4.2060E-02
5	9.5587E-02	3.7600E-02

References

[1] J.L. Shanks, S. Treitel and J.H. Justice, "Stability and Synthesis of Two-Dimensional Recursive Filters" *IEEE Trans. Audio Electroacoust.*, Vol. AU-20, pp. 115-128, 1972.

[2] J. A. Cadzow, "Recursive Digital Filter Synthesis via Gradient Based Algorithms", *IEEE Trans. on Acoust., Speech and Signal Processing*, Vol. ASSP-24, pp. 349-355, 1976.

[3] G. A. Shaw and R. M. Mersereau, "Design, Stability and Performance of Two-Dimensional Recursive Digital Filters", *Tech. Report E21-B05-1*, Georgia Inst. of Technology School of Electrical Engg., 1979.

[4] D.E. Dudgeon and R.M. Mersereau, *Multidimensional Digital Signal Processing*, Englewood Cliffs, NJ: Prentice Hall, 1984.

[5] J.L.Shanks, "Recursion Filters for Digital Processing", *Geophysics*, Vol. 32, pp. 33-51, 1967.

[6] K. Steiglitz and L.E. McBride, "A Technique for Identification of Linear Systems", *IEEE Transactions on Automatic Control*, Vol. AC-10, pp. 461-464, 1965.

[7] A.G. Evans and R. Fischl, "Optimal Least Squares Time-Domain Synthesis of Recursive Digital Filters", *IEEE Transactions on Audio and Electro-Acoustics*, Vol. AU-21, pp. 61-65, 1973.

[8] A.K. Shaw, *Structured Matrix Problems in Signal Processing*, Ph.D. Dissertation, Univ. of Rhode Island, RI, 1987.

[9] A.K. Shaw and R. Kumaresan, "Some Structured Matrix Approximation Problems", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, NY, pp. 2324-2327, April, 1988.

[10] T. Hinamoto and S. Maekawa, "Spatial-Domain Design of a Class of Two-Dimensional Recursive Digital Filters", *IEEE Trans. on ASSP*, Vol.- ASSP-32, no.1, Feb.,1984.

[11] T. Hinamoto and F.W. Fairman, "Separable Denominator State Space Realization of Two Dimensional Filters using a Canonical Form", *IEEE Trans. Acoust. Speech, Sig. Proc.*, vol. ASSP-29, no. 4, pp. 846-853, 1981.

[12] T. Hinamoto and S. Maekawa, "Separable-Denominator 2-D Rational Approximation via 1-D Based Algorithm ", *IEEE Trans. Cir.Syst.*, Vol. CAS-32, pp. 989-999, Nov. 1985.

SECTION 2.4 : IDENTIFICATION OF DISCRETE TIME MULTIVARIABLE SYSTEMS FROM IMPULSE RESPONSE DATA

SUMMARY

In this Section we consider the problem of identification of transfer function matrices of discrete time multivariable systems. The proposed technique obtains an optimal approximation from the given (possibly noisy) measured *impulse response data*. It is assumed that the measured impulse response data corresponds to a system with a strictly proper transfer function matrix. Based on the proposed theoretical basis, an efficient computational algorithm is developed and illustrated by means of several examples.

1. INTRODUCTION

Mathematical models of linear systems can be broadly classified as non-parametric and parametric models. Non-parametric models include impulse responses, covariance functions, spectral density descriptions etc. These models tend to be infinite dimensional in nature. Parameterization leads to finite dimensional models. Some examples of parametric models are differential equations, difference equations, transfer functions, state space descriptions etc. In parametric modeling, having assigned a model structure to the system, the problem is to find best set of parameters to represent the system.

The problem of model identification of single-input, single-output (SISO) continuous as well as discrete-time systems is very well studied [1]-[10]. However, despite the importance of the problem of identification of multi-input, multi-output (MIMO) systems, relatively small proportion of the existing literature addresses it. This is due, in parts, to the

1) non-uniqueness in the parameterization of multivariable systems,
2) difficulty in determining a cost function that reflects appropriately the importance of various input output pairs and
3) limited success in extending the well established results from SISO system theory to multivariable systems.

In recent years, several authors have investigated the problem of parameterization (for the purpose of identification) and identification of MIMO systems. Choice of parameterization was discussed by Glover and Willems [11], Denham [12] and Gevers and Wertz [13], where it was shown that knowing the order of the system, a minimal set of parameters that uniquely define the system can be identified. The problem of MIMO system identification has been addressed by several researchers using several approaches: Among others, Moonen and Vandewalle [14] developed a quotient SVD framework for identifying state space models from the input-output error covariance matrix. Helmicki, Jacobson and Nett [15] and Gu and Khargonekar [16] have developed robustly convergent in H^∞ framework. Makila [17] uses Laguerre series for identification in H^∞ framework and Rao [18] uses Walsh functions for identification of multivariable systems. In the wide-sense stationary random process framework, Friedlander presents a modified Yule-Walker method for estimating the multi-channel ARMA parameters in [23]. It should be emphasized that the above references are only some of the most recent ones appearing in literature and those which address the multivariable identification problem. For the SISO systems, references [1]-[10] provide an excellent exposition for the solution of the identification problem and also contain extensive bibliography.

In this work, we study the problem of determining a parametric model (a discrete time transfer function matrix) from the given impulse response data. We will assume that the system that we wish to identify is

represented by a $(p \times m)$ rational function matrix:

$$\mathbf{G}(z) = \frac{1}{b(z)} \begin{bmatrix} a_{11}(z) & a_{12}(z) & \cdots & a_{1m}(z) \\ a_{21}(z) & a_{22}(z) & \cdots & a_{2m}(z) \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}(z) & a_{p2}(z) & \cdots & a_{pm}(z) \end{bmatrix} = \frac{\mathbf{A}(z)}{b(z)} \tag{1.1}$$

where $\deg(b(z)) = n$ and $\mathbf{A}(z)$ is a $(p \times m)$ polynomial matrix such that $\deg(a_{ij}(z)) = n-1$, $i(j) = 1, 2, \ldots, p(m)$. Further, it is assumed that the first N terms of the measured (possibly noisy) unit pulse response data of the system

$$\mathbf{H}(z) = \sum_{i=0}^{N-1} \mathbf{H}(i) z^{-i} \tag{1.2}$$

are available where $\mathbf{H}_i$ represents the matrix of impulse responses at the i-th time instant.

It is well known that even for a single input, single output system, when the unknown system contains both poles and zeros, the problem of identification of numerator and denominator polynomial coefficients is a highly non-linear optimization problem. Two of the techniques that have been used frequently in parameters identification of scalar plants in signal processing literature are those proposed by Steiglitz and McBride [19] and Evans and Ficshl in [20]. Evans-Fischl's approach minimizes the difference (in the least square sense) between the *measured* and the *desired* impulse response data, while Steiglitz and McBride approach uses linearized error criteria. In this respect, provided the degree of the numerator polynomial (m) is one less than the degree of the denominator polynomial (n), Evans-Fischl approach can be considered to be "optimal".

The primary purpose of this work is to generalize Evans-Fischl method (EFM), to the case when the number of inputs and output is greater than one. We propose a generalized error norm measure by giving equal weight to impulse response corresponding to each input/output pair. Based on this error norm, a single denominator polynomial with a pre-specified degree is computed. Knowing the coefficients of the denominator polynomial, the numerator polynomials are evaluated by solution of linear algebraic equations.

The layout of this Section is as follows: Since Evans-Ficshl technique is not very well known in the control systems literature, in Subsection 2, we briefly review this approach. In Subsection 3, the error criteria for multivariable system is defined and the error minimization technique is extended to multi-input multi-output systems. In Subsection 4, the identification results from extensive simulations on various order and varying degree of noise contamination are presented.

2. Scalar Systems

Assume that the given single input, single output plant is described by a strictly proper *stable* z-domain transfer function:

$$H(z) = \frac{a_0 + a_1 z^{-1} + \cdots + a_{n-1} z^{-(n-1)}}{1 + b_1 z^{-1} + \cdots + b_{n-1} z^{-(n-1)} + b_n z^{-n}}, \tag{2.1}$$

where the coefficient of z^0 term in denominator has been assumed to be unity without any loss of generality. Using long division, the above transfer function can be rewritten as the infinite series

$$H(z) = h_0 + h_1 z^{-1} + \cdots + h_n z^{-n} + h_{n+1} z^{-(n+1)} + \cdots \tag{2.2}$$

Define vectors $\mathbf{f}, \mathbf{h} \in \mathbb{R}^N$, where,

$$\begin{aligned} \mathbf{f} &= [f_0 \quad f_1 \quad \cdots \quad f_{N-1}]^T \\ \mathbf{h} &= [h_0 \quad h_1 \quad \cdots \quad h_{N-1}]^T, \end{aligned} \tag{2.3}$$

denote the N samples of the measured and the actual impulse response data, respectively. Then, the identification of parameters a_i and b_i can be stated as the following least squares minimization problem:

$$\min_{\mathbf{a},\mathbf{b}} \|\mathbf{e}\| \triangleq \min_{\mathbf{a},\mathbf{b}} \left[\sum_{i=0}^{N-1} e_i^2\right]^{1/2} \tag{2.4}$$

where,

$$\mathbf{e} \triangleq \mathbf{f} - \mathbf{h} \tag{2.5a}$$

$$\mathbf{a} \triangleq [a_0 \quad a_1 \quad \cdots \quad a_{n-1}]^T \tag{2.5b}$$

$$\mathbf{b} \triangleq [1 \quad b_1 \quad \cdots \quad b_n]^T. \tag{2.5c}$$

The transfer function coefficients are related to the impulse response samples in $H(z)$ as

$$\begin{bmatrix} \mathbf{a} \\ \cdots \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_1 \\ \cdots \\ \mathbf{H}_2 \end{bmatrix} \mathbf{b} \tag{2.6}$$

where, $\mathbf{a}$, $\mathbf{b}$ have been defined in (2.5) and

$$\mathbf{H}_1 \triangleq \begin{bmatrix} h_0 & 0 & \cdots & 0 & 0 \\ h_1 & h_0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ h_{n-1} & h_{n-2} & \cdots & h_0 & 0 \end{bmatrix} \in \mathbb{R}^{(n)\times(n+1)}, \tag{2.7a}$$

$$\mathbf{H}_2 \triangleq \begin{bmatrix} h_n & h_{n-1} & \cdots & h_0 \\ h_{n+1} & h_n & \cdots & h_1 \\ \vdots & \vdots & \ddots & \vdots \\ h_{N-1} & h_{N-2} & \cdots & h_{N-n-1} \end{bmatrix} \in \mathbb{R}^{(N-n)\times(n+1)}. \tag{2.7b}$$

If $\mathbf{b}$ and $\mathbf{H}_1$ are known, then $\mathbf{a}$ can be found by solving the system of linear algebraic equations $\mathbf{a} = \mathbf{H}_1\mathbf{b}$. However, in the present case, the exact $\mathbf{h}$ and therefore, the matrices $\mathbf{H}_1$ and $\mathbf{H}_2$ are not known. Therefore, we replace the elements of $\mathbf{H}_1$ and $\mathbf{H}_2$ by the corresponding matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ formed from the measured impulse response data $\mathbf{f}$. To obtain the initial estimate for $\mathbf{b}$, consider the lower half of (2.6) given by $\mathbf{H}_2\mathbf{b} = \mathbf{0}$. Replacing $\mathbf{H}_2$ by $\mathbf{F}_2$ and expanding the relation, we get

$$\begin{bmatrix} f_n & f_{n-1} & \cdots & f_1 & f_0 \\ f_{n+1} & f_n & \cdots & f_2 & f_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{N-1} & f_{N-2} & f_{N-3} & \cdots & f_{N-n-1} \end{bmatrix} \begin{bmatrix} 1 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} = \mathbf{d}(\mathbf{b}), \tag{2.8}$$

where $\mathbf{d}(\mathbf{b})$ is the equation error. The above equation can be rewritten as

$$\begin{bmatrix} f_{n-1} & f_{n-2} & \cdots & f_0 \\ f_n & f_{n-1} & \cdots & f_1 \\ \vdots & \vdots & \ddots & \vdots \\ f_{N-2} & f_{N-3} & \cdots & f_{N-n-1} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = - \begin{bmatrix} f_n \\ f_{n+1} \\ \vdots \\ f_{N-1} \end{bmatrix} + \mathbf{d}(\mathbf{b}). \tag{2.9}$$

Now, let

$$\mathbf{G} = \begin{bmatrix} f_{n-1} & f_{n-2} & \cdots & f_0 \\ f_n & f_{n-1} & \cdots & f_1 \\ \vdots & \vdots & \ddots & \vdots \\ f_{N-2} & f_{N-3} & \cdots & f_{N-n-1} \end{bmatrix} \quad \text{and} \quad \mathbf{g} = \begin{bmatrix} f_n \\ f_{n+1} \\ \vdots \\ f_{N-1} \end{bmatrix}, \tag{2.10}$$

the initial estimate for $\mathbf{b}$ can be obtained by minimizing $\|\mathbf{d}(\mathbf{b})\|$ with respect to $\hat{\mathbf{b}} = [b_1 \ b_2 \ \cdots \ b_n]^T$ and can be computed as

$$\mathbf{b}^{(0)} = \begin{bmatrix} 1 \\ \cdots\cdots \\ -\mathbf{G}^{\dagger}\mathbf{g} \end{bmatrix}, \tag{2.11}$$

where $\mathbf{G}^{\dagger}$ denotes the pseudo-inverse of $\mathbf{G}$. In general, the first approximation is fairly crude one because it only minimizes an equation error and does not minimize the actual fitting error norm of (2.4). Unlike the equation error, the fitting error will be shown to be non-linearly related to $\mathbf{b}$ and hence it has to be refined iteratively to obtain a better denominator polynomial to match the impulse response.

Note that in (2.8), if the exact impulse response $\mathbf{h}$ had been known, the equality will be satisfied. However, due to measurement noise, there will be some residual error $\mathbf{d}(\mathbf{b})$ as shown in (2.8). This equation error is now rewritten as:

$$\begin{aligned} \mathbf{d}(\mathbf{b}) &\triangleq \mathbf{F}_2\mathbf{b} \\ &= \begin{bmatrix} b_n & b_{n-1} & \cdots & 1 & & \\ & \ddots & \ddots & & \ddots & \\ & & \mathbf{b}_n & \mathbf{b}_{n-1} & \cdots & 1 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_{N-1} \end{bmatrix} \\ &\triangleq \mathbf{B}^T\mathbf{f}. \end{aligned} \tag{2.12}$$

Now, $\mathbf{f}$ can be represented in terms of $\mathbf{e}$ and $\mathbf{h}$ as $\mathbf{f} = \mathbf{h} + \mathbf{e}$ and (2.12) can be expressed as

$$\begin{aligned} \mathbf{d}(\mathbf{b}) &= \mathbf{B}^T[\mathbf{h} + \mathbf{e}] \\ &= \mathbf{B}^T\mathbf{e} \quad \text{because } \mathbf{H}_2\mathbf{b} = \mathbf{B}^T\mathbf{h} = \mathbf{0}. \end{aligned} \tag{2.13}$$

Rewriting the error to be minimized in terms of equation error and using the "projection theorem" [21], we get (some more explanation of the rationale behind this approach is given later for the multi-channel case),

$$\begin{aligned} \mathbf{e} &= \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{f} \\ &\triangleq \mathbf{W}\mathbf{B}^T\mathbf{f} \\ &= \mathbf{W}\mathbf{F}_2\mathbf{b} \\ &= \mathbf{W}[\mathbf{g} \quad \mathbf{G}]\mathbf{b} \\ &= \mathbf{W}\mathbf{g} + \mathbf{W}\mathbf{G}\hat{\mathbf{b}} \end{aligned} \tag{2.14}$$

where $\hat{\mathbf{b}} = [b_1 \ b_2 \ \cdots \ b_n]^T$. From, (2.15), it is clear that $\mathbf{W}\mathbf{G}\mathbf{b} = -\mathbf{W}\mathbf{g} + \mathbf{e}$ and a new expression for $\hat{\mathbf{b}}$ can be obtained by minimizing $\|\mathbf{e}\|$ as:

$$\begin{aligned} \hat{\mathbf{b}} &= -(\mathbf{W}\mathbf{G})^{\dagger}\mathbf{W}\mathbf{g} \\ &= -(\mathbf{G}^T\mathbf{W}^T\mathbf{W}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}^T\mathbf{W}\mathbf{g}. \end{aligned} \tag{2.15}$$

In summary, the estimate for $\mathbf{b}$ is iteratively refined using the following relation:

$$\mathbf{b}^{(i)} = \begin{bmatrix} 1 \\ \cdots\cdots\cdots \\ -[\mathbf{VG}]^{-1}[\mathbf{V}]\mathbf{g} \end{bmatrix} \tag{2.16}$$

where,

$$\mathbf{V} \triangleq [\mathbf{G}^T\mathbf{W}^T\mathbf{W}] \tag{2.17a}$$

$$\mathbf{W} \triangleq [\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}] \quad \text{and} \tag{2.17b}$$

$$\mathbf{B} \triangleq \begin{bmatrix} b_n & 0 & \cdots & 0 \\ b_{n-1} & b_n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & b_1 & \cdots & b_n \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & b_1 \\ 0 & 0 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{N-n-1 \times N-1}. \tag{2.17c}$$

It should be mentioned here that, at each iteration, the new improved estimates of $\mathbf{b}$ are used in forming the matrix $\mathbf{W}$. Since the above iterations minimize the exact fitting error of (2.4), at convergence, the optimal estimate of $\mathbf{b}$ is found. The iterations are performed in two phases. The scalar case being only a special case, these two phases are explained with more details in Subsection 3.

When the estimates of $\mathbf{b}$ converge, $\mathbf{a}$ can be computed directly as $\mathbf{a} = \hat{\mathbf{H}}_1\mathbf{b}$, where $\hat{\mathbf{H}}_1$ has the same form as $\mathbf{H}_1$, except the elements h_i are replaced by $f_i - e_i$, $i = 1, \ldots, N_1$ and e_i are the elements of the error vector $\mathbf{e} = [e_0 \ e_1 \ \cdots \ e_{N-1}]^T = \mathbf{W}\mathbf{B}^T\mathbf{f}$. Here $\mathbf{W}\mathbf{B}^T$ are formed using the optimized values of $\mathbf{b}$.

3. MULTIVARIABLE SYSTEMS

In this Subsection the EFM algorithm is generalized for the Multi-Input/Multi-Output (MIMO) case.

Assume that the given plant is described by the rational transfer function matrix $\mathbf{G}(z)$ in (1.1). Denoting each (i,j)-th element of this matrix as H_{ij}, $\mathbf{G}(z)$ may also be written as,

$$\mathbf{G}(z) = \begin{bmatrix} H_{11}(z) & H_{12}(z) & \cdots & H_{1m}(z) \\ H_{21}(z) & H_{22}(z) & \cdots & H_{2m}(z) \\ \vdots & \vdots & \ddots & \vdots \\ H_{p1}(z) & H_{p2}(z) & \cdots & H_{pm}(z) \end{bmatrix} = \frac{\mathbf{A}(z)}{b(z)} \tag{3.1}$$

where each $H_{ij}(z)$ is given by,

$$H_{ij}(z) = \frac{a_{ij}(0) + a_{ij}(1)z^{-1} + \cdots + a_{ij}(n-1)z^{-(n-1)}}{1 + b_{ij}(1)z^{-1} + \cdots + b_{ij}(n-1)z^{-(n-1)} + b_{ij}(n)z^{-n}}. \tag{3.2}$$

Now, similar to (2.2), $H_{ij}(z)$ can also be written as,

$$H_{ij}(z) = h_{ij}(0) + h_{ij}(1)z^{-1} + \cdots + h_{ij}(n)z^{-n} + \cdots \tag{3.3}$$

Let $\mathbf{f}_{ij}$, $\mathbf{h}_{ij} \in \mathbb{R}^N$, where,

$$\begin{aligned} \mathbf{f}_{ij} &= [f_{ij}(0) \quad f_{ij}(1) \quad \cdots \quad f_{ij}(N-1)]^T \\ \mathbf{h}_{ij} &= [h_{ij}(0) \quad h_{ij}(1) \quad \cdots \quad h_{ij}(N-1)]^T, \end{aligned} \tag{3.4}$$

denote the N samples of the measured and the desired impulse response data, respectively, corresponding to the ij-th polynomial element of the $\mathbf{G}(z)$ matrix. It should be mentioned here that the number of measured samples, N, has been assumed to be equal for each case without any loss of generality and that the algorithm can be modified easily if the number of measurements are unequal. The only restriction is that for each case, there must be at least $2n$ measurements, n denoting the order. Next, defining the error vector for the ij-th case as,

$$\mathbf{e}_{ij} \triangleq \mathbf{f}_{ij} - \mathbf{h}_{ij}, \tag{3.5}$$

the whole error matrix may be written as,

$$\mathbf{E} \triangleq \begin{bmatrix} \mathbf{e}_{11} & \mathbf{e}_{12} & \cdots & \mathbf{e}_{1m} \\ \mathbf{e}_{21} & \mathbf{e}_{22} & \cdots & \mathbf{e}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{e}_{p1} & \mathbf{e}_{p2} & \cdots & \mathbf{e}_{pm} \end{bmatrix} \tag{3.6a}$$

$$\triangleq \begin{bmatrix} | & | & \cdots & | \\ \mathbf{e}_1 & \mathbf{e}_2 & \ddots & \mathbf{e}_m \\ | & | & \cdots & | \end{bmatrix} \tag{3.6b}$$

Then, the least-squares minimization problem can be stated as

$$\min_{\{\mathbf{a}_{ij}\},\mathbf{b}} \|\mathbf{E}\|_F \triangleq \min_{\{\mathbf{a}_{ij}\},\mathbf{b}} \left[\sum_{i=1}^{p} \sum_{j=1}^{m} \sum_{k=0}^{N-1} e_{ij}(k)^2 \right]^{1/2} \tag{3.7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm (or the matrix ℓ_2 norm) and

$$\mathbf{a}_{ij} = [a_{ij}(0) \quad a_{ij}(1) \quad \cdots \quad a_{ij}(n-1)]^T \tag{3.8a}$$

$$\mathbf{b} = [1 \quad b(1) \quad \cdots \quad b(n)]^T . \tag{3.8b}$$

For ease of formulation, a large vector of errors is created from $\mathbf{E}$ next. Define,

$$\mathbf{e}_l \triangleq Vec\ [\mathbf{E}] = \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_m \end{bmatrix}, \tag{3.9a}$$

where $Vec[\cdot]$ is the operation of stacking the columns of a matrix to form a large vector. Note that e is a $pmN \times 1$ vector which can be related to the measured and the desired impulse responses as,

$$\mathbf{e}_l = \mathbf{f}_l - \mathbf{h}_l, \tag{3.9b}$$

where

$$\mathbf{f}_l \triangleq \begin{bmatrix} \mathbf{f}_{11} \\ \mathbf{f}_{21} \\ \vdots \\ \mathbf{f}_{p1} \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{f}_{1m} \\ \mathbf{f}_{2m} \\ \vdots \\ \mathbf{f}_{pm} \end{bmatrix} \quad \text{and} \quad \mathbf{h}_l \triangleq \begin{bmatrix} \mathbf{h}_{11} \\ \mathbf{h}_{21} \\ \vdots \\ \mathbf{h}_{p1} \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{h}_{1m} \\ \mathbf{h}_{2m} \\ \vdots \\ \mathbf{h}_{pm} \end{bmatrix}. \tag{3.9c}$$

With this definition, the least-squares minimization criterion of (3.7) can be restated as

$$\min_{\{\mathbf{a}_{ij}\},\mathbf{b}} \|\mathbf{e}_l\|_2 \triangleq \min_{\{\mathbf{a}_{ij}\},\mathbf{b}} \left[\sum_{i=1}^{pmN} e_l^2(i)\right]^{1/2} \tag{3.10}$$

where $\|\cdot\|_2$ denotes the vector ℓ_2 norm. Now, corresponding to each ij-th element in $\mathbf{G}(z)$ matrix, we can form similar to (2.6),

$$\begin{bmatrix} \mathbf{a}_{ij} \\ \cdots \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{ij}^1 \\ \cdots \\ \mathbf{H}_{ij}^2 \end{bmatrix} \mathbf{b} \tag{3.11}$$

where, $\mathbf{a}_{ij}$, $\mathbf{b}$ have been defined in (3.8) and

$$\mathbf{H}_{ij}^1 \triangleq \begin{bmatrix} h_{ij}(0) & 0 & \cdots & 0 & 0 \\ h_{ij}(1) & h_{ij}(0) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ h_{ij}(n-1) & h_{ij}(n-2) & \cdots & h_{ij}(0) & 0 \end{bmatrix} \in \mathbb{R}^{(n)\times(n+1)}, \tag{3.12a}$$

$$\mathbf{H}_{ij}^2 \triangleq \begin{bmatrix} h_{ij}(n) & h_{ij}(n-1) & \cdots & h_{ij}(0) \\ h_{ij}(n+1) & h_{ij}(n) & \cdots & h_{ij}(1) \\ \vdots & \vdots & \ddots & \vdots \\ h_{ij}(N-1) & h_{ij}(N-2) & \cdots & h_{ij}(N-n-1) \end{bmatrix} \in \mathbb{R}^{(N-n)\times(n+1)}. \tag{3.12b}$$

Now, stacking the upper partitions of (3.11) for all i, j we get,

$$\mathbf{a}_l \triangleq \begin{bmatrix} \mathbf{a}_{11} \\ \mathbf{a}_{21} \\ \vdots \\ \mathbf{a}_{p1} \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{a}_{1m} \\ \mathbf{a}_{2m} \\ \vdots \\ \mathbf{a}_{pm} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{11}^{(1)} \\ \mathbf{H}_{21}^{(1)} \\ \vdots \\ \mathbf{H}_{p1}^{(1)} \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{H}_{1m}^{(1)} \\ \mathbf{H}_{2m}^{(1)} \\ \vdots \\ \mathbf{H}_{pm}^{(1)} \end{bmatrix} \mathbf{b} \triangleq \mathbf{H}_l^{(1)}\mathbf{b}. \tag{3.13a}$$

Similarly, the lower partitions of (3.11) may also be stacked as,

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \cdots \\ \vdots \\ \cdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{11}^{(2)} \\ \mathbf{H}_{21}^{(2)} \\ \vdots \\ \mathbf{H}_{p1}^{(2)} \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{H}_{1m}^{(2)} \\ \mathbf{H}_{2m}^{(2)} \\ \vdots \\ \mathbf{H}_{pm}^{(2)} \end{bmatrix} \mathbf{b} \triangleq \mathbf{H}_l^{(2)}\mathbf{b} = \begin{bmatrix} \mathbf{B}^T\mathbf{h}_{11} \\ \mathbf{B}^T\mathbf{h}_{21} \\ \vdots \\ \mathbf{B}^T\mathbf{h}_{p1} \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{B}^T\mathbf{h}_{1m} \\ \mathbf{B}^T\mathbf{h}_{2m} \\ \vdots \\ \mathbf{B}^T\mathbf{h}_{pm} \end{bmatrix} = (\mathbf{I}_{pm} \otimes \mathbf{B}^T)\mathbf{h}_l \tag{3.13b}$$

where, $\otimes$ denotes the matrix Kronecker product, $\mathbf{B}$ is defined in (2.17c) and $\mathbf{I}_{pm}$ denotes an $pm \times pm$ identity matrix. If the elements in $\mathbf{H}_l^{(1)}$ and $\mathbf{b}$ are known, $\mathbf{a}_l$ can be found uniquely from (3.13a). Replacing the $\mathbf{H}_{ij}$'s with the corresponding $\mathbf{F}_{ij}$'s, an equation error is formed,

$$\mathbf{d}_l(\mathbf{b}) \underline{\underline{\Delta}} \mathbf{F}_l^{(2)}\mathbf{b} \underline{\underline{\Delta}} \begin{bmatrix} \mathbf{F}_{11}^{(2)} \\ \mathbf{F}_{21}^{(2)} \\ \vdots \\ \mathbf{F}_{p1}^{(2)} \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{F}_{1m}^{(2)} \\ \mathbf{F}_{2m}^{(2)} \\ \vdots \\ \mathbf{F}_{pm}^{(2)} \end{bmatrix} \mathbf{b}. \tag{3.14}$$

Following the steps analogous to (2.8)-(2.11), one can again find an initial estimate of $\mathbf{b}$ as follows,

$$\mathbf{b}^{(0)} = \begin{bmatrix} 1 \\ \cdots\cdots \\ -\mathbf{G}_l^{\dagger}\mathbf{g}_l \end{bmatrix}, \tag{3.15}$$

where, $\mathbf{g}_l$ contains the first column of $\mathbf{F}_l^{(2)}$ and $\mathbf{G}_l$ contains the rest of the columns. But in order to find the optimum estimate of $\mathbf{b}$ we still have to optimize the criterion in (3.10). To proceed in that direction the equation error in (3.14) is rewritten in a more useful form as,

$$\begin{aligned} \mathbf{d}_l(\mathbf{b}) &= \begin{bmatrix} \mathbf{B}^T\mathbf{f}_{11} \\ \mathbf{B}^T\mathbf{f}_{21} \\ \vdots \\ \mathbf{B}^T\mathbf{f}_{pm} \end{bmatrix} \\ &= (\mathbf{I}_{pm} \otimes \mathbf{B}^T)\mathbf{f}_l. \end{aligned} \tag{3.16}$$

Replacing $\mathbf{f}_l$ by $\mathbf{e}_l + \mathbf{h}_l$ and using (3.13b) we get,

$$\mathbf{d}_l(\mathbf{b}) = (\mathbf{I}_{pm} \otimes \mathbf{B}^T)\mathbf{e}_l \tag{3.17}$$

But in order to facilitate the minimization of the fitting error norm of (3.10), we have to find an inverse relationship between $\mathbf{e}_l$ and $\mathbf{d}_l(\mathbf{b})$. According to orthogonality principle, the error $\mathbf{e}_l$ for a given $\mathbf{b}$ and corresponding to the optimum $\mathbf{a}_l$ must be orthogonal to the desired response vector $\mathbf{h}_l$. Otherwise there would remain some information contained in the non-zero projection of $\mathbf{e}_l$ onto $\mathbf{h}_l$. The complete orthogonal basis space of this error can be found from equation (3.13b) which clearly demonstrates that the $pm(N-n)$ columns of $(\mathbf{I}_{pm} \otimes \mathbf{B})$ are orthogonal to $\mathbf{h}_l$. Hence the error $\mathbf{e}_l(\mathbf{a}_l^*)$ corresponding to optimal $\mathbf{a}_l$ may be formed as a linear combination of all its orthogonal basis vectors as follows,

$$\mathbf{e}_l(\mathbf{a}_l^*) \underline{\underline{\Delta}} (\mathbf{I}_{pm} \otimes \mathbf{B})\mathbf{c} \tag{3.18}$$

where $\mathbf{c}$, a vector of unknown constants, needs to be determined. By plugging in (3.18) in (3.17) we obtain

$$\mathbf{d}_l(\mathbf{b}) = (\mathbf{I}_{pm} \otimes \mathbf{B}^T)(\mathbf{I}_{pm} \otimes \mathbf{B})\mathbf{c} \tag{3.19a}$$

$$= (\mathbf{I}_{pm} \otimes \mathbf{B}^T\mathbf{B})\mathbf{c}. \tag{3.19b}$$

Hence,

$$\begin{aligned} \mathbf{c} &= (\mathbf{I}_{pm} \otimes (\mathbf{B}^T\mathbf{B})^{-1})\mathbf{d}_l(\mathbf{b}) \\ &= (\mathbf{I}_{pm} \otimes (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T)\mathbf{f}_l \qquad \text{[using (3.16)]}. \end{aligned} \tag{3.20}$$

Plugging this back in (3.18) and following the similar steps as in (2.14),

$$\mathbf{e}_l(\mathbf{a}_l^*) = (\mathbf{I}_{pm} \otimes \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T)\mathbf{f}_l \tag{3.21a}$$

$$= (\mathbf{I}_{pm} \otimes \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1})(\mathbf{I}_{pm} \otimes \mathbf{B}^T)\mathbf{f}_l \tag{3.21b}$$

$$\underline{\underline{\Delta}}\ \mathbf{W}_l(\mathbf{I}_{pm} \otimes \mathbf{B}^T)\mathbf{f}_l \tag{3.21c}$$

$$= \mathbf{W}_l\mathbf{F}_l^{(2)}\mathbf{b}; \qquad \text{[using (3.14)]} \tag{3.21d}$$

$$\underline{\underline{\Delta}}\ \mathbf{W}_l[\mathbf{g}_l \quad \mathbf{G}_l]\mathbf{b} \tag{3.21e}$$

$$= \mathbf{W}_l\mathbf{g}_l + \mathbf{W}_l\mathbf{G}_l\hat{\mathbf{b}}. \tag{3.21f}$$

For an optimum $\mathbf{a}$, this is the fitting error that we need to minimize, *i.e.*, the criterion in (3.10) is exactly equivalent to,

$$\min_{\mathbf{b}} \|\mathbf{e}_l(\mathbf{a}_l^*)\|_2. \tag{3.22}$$

At minimum this will produce $\mathbf{e}(\mathbf{a}_l^*, \mathbf{b}^*)$. Note that in the expression of $\mathbf{e}_l(\mathbf{a}_l^*)$, the matrix $\mathbf{W}$ does have dependence on $\mathbf{b}$ and $\mathbf{b}$ contains the elements with respect to which the criterion needs to be minimized. Following EFM, the minimization of (3.22) is performed in two steps of iterations. In the first phase of iterations, the matrix $\mathbf{W}$ is constructed from the estimate of $\mathbf{b}$ found at the previous iteration. The new update of $\mathbf{b}$ is then obtained from,

$$\begin{aligned} \hat{\mathbf{b}}^{(i+1)} &= -(\mathbf{W}_l^{(i)}\mathbf{G}_l)^\dagger\mathbf{W}_l^{(i)}\mathbf{g}_l \\ &= -\left(\mathbf{G}_l^T\mathbf{W}_l^{T^{(i)}}\mathbf{W}_l^{(i)}\mathbf{G}_l\right)^{-1}\mathbf{G}_l^T\mathbf{W}_l^{T^{(i)}}\mathbf{W}_l^{(i)}\mathbf{g}_l. \end{aligned} \tag{3.23}$$

In summary, the estimate for $\mathbf{b}$ is iteratively refined using the following relation:

$$\mathbf{b}^{(i+1)} = \begin{bmatrix} 1 \\ \cdots\cdots\cdots\cdots \\ -[\mathbf{V}_l^{(i)}\mathbf{G}_l]^{-1}[\mathbf{V}_l^{(i)}]\mathbf{g}_l \end{bmatrix} \tag{3.24}$$

where,

$$\mathbf{V}_l^{(i)} \underline{\underline{\Delta}}\ \mathbf{G}_l^T\mathbf{W}^{T^{(i)}}\mathbf{W}^{(i)} \tag{3.25a}$$

$$= \mathbf{G}_l^T(\mathbf{B}^{T^{(i)}}\mathbf{B}^{(i)})^{-1} \tag{3.25b}$$

It should be mentioned here that, at any iteration step $(i+1)$, the new estimates of $\mathbf{b}^{(i)}$ are used in forming the matrix $\mathbf{W}^{(i)}$. Note that the initial estimate $\mathbf{b}^{(0)}$ comes from the equation error minimization step of (3.15).

The first phase alone may not converge to the absolute optimum of $\mathbf{b}$ that minimizes $\mathbf{e}(\mathbf{a}_l^*)$ completely though our experience with many examples does indicate that the first phase comes quite close to the optimum. In some cases, especially when the deviations of the measured responses from the desired ones are relatively large, a second phase of EFM needs to be invoked. In the second phase of iterations, the variation of $\mathbf{W}$ w.r.t. $\mathbf{b}$ is also taken into account. The details of phase 2 for the scalar case may be found in [20]. The extension for the multi-channel case is similar to the development of the extension for the first phase given above.

At convergence of the second phase, the optimum value $\mathbf{b}^*$ is found. Plugging that in (3.21) the optimal error vector $\mathbf{e}(\mathbf{a}_l^*, \mathbf{b}^*)$ is computed. The optimal impulse response $\hat{\mathbf{h}}_l$ is then found from (3.9b) as,

$$\hat{\mathbf{h}}_l = \mathbf{f}_l - \mathbf{e}(\mathbf{a}_l^*, \mathbf{b}^*). \tag{3.26}$$

Finally, the optimal $\mathbf{a}_l$ is computed from (3.13a) as

$$\mathbf{a}_l^* = \hat{\mathbf{H}}_l^{(1)} \mathbf{b}^*, \tag{3.27}$$

where $\hat{\mathbf{H}}_l^{(1)}$ has the same form as $\mathbf{H}_l^{(1)}$ except that the elements $h_{ij}(k)$ are replaced by the corresponding $\hat{\mathbf{h}}_{ij}(k)$. We should mention here that the separable optimization of **a** and **b**, as given here, falls within a special class of non-linear optimization problems which have been studied extensively by numerical analysts [22]. It has been shown in [22] that if some of the unknown variables are linearly related to the error and the other variables are non-linearly related and if the variables do separate as the case studied here, the two step optimization do produce the optimum for both sets of variables.

We should mention here that the separable optimization of a and b, as presented in this paper, falls within a special class of non-linear optimization problems which have been studied extensively by numerical analysts [22]. It was shown by Golub and Pereyera in [22] that if some of the unknown variables are linearly related to the error and the other variables are non-linearly related and if the variables do separate as the case studied here, then the two level optimization of the kind described in the preceding sections produces the *optimum* values for both sets of variables.

4. Simulation Results

A (2×2) transfer function matrix was used for simulations. Table 4.1 contains the coefficients of the denominator polynomial $b(z)$ and Tables 4.2(a) and (b) contain the coefficients of the numerator polynomials.

Table 4.1: Coefficients of denominator of $\mathbf{G}(z)$

Coeff	Denominotor
z^6	$1.000000000000000\epsilon + 00$
z^5	$-2.239200000000000\epsilon + 00$
z^4	$1.682120780000000\epsilon + 00$
z^3	$-4.675245365940000\epsilon - 01$
z^2	$5.995215508524930\epsilon - 02$
z^1	$-3.069148211131338\epsilon - 02$
z^0	$7.921660299005685\epsilon - 03$

Table 4.2(a): Coefficients of numerators of $H(z)$

Coeff	$a_{11}(z)$	$a_{12}(z)$
z^5	$2.310000000000000\epsilon - 02$	$4.120000000000000\epsilon - 01$
z^4	$-3.829980000000000\epsilon - 02$	$-6.465516000000000\epsilon - 01$
z^3	$2.825354370300000\epsilon - 02$	$3.900814928800000\epsilon - 01$
z^2	$-1.163969877511980\epsilon - 02$	$-1.124197444290240\epsilon - 01$
z^1	$2.305936357609712\epsilon - 03$	$1.534472131069005\epsilon - 02$
z^0	$-1.599293795106799\epsilon - 04$	$-7.808982566335611\epsilon - 04$

TABLE 4.2(B): Coefficients of numerators of $H(z)$

Coeff	$a_{21}(z)$	$a_{22}(z)$
z^5	$-2.120000000000000e-01$	$2.318000000000000e-01$
z^4	$2.986020000000000e-01$	$-2.805161954172887e-01$
z^3	$-1.568783612400000e-01$	$1.031386276967342e-01$
z^2	$5.140274245157200e-02$	$-3.466242652862272e-03$
z^1	$-8.602449814804975e-03$	$-4.646062737792217e-03$
z^0	$4.959796217794915e-04$	$6.208784880946757e-04$

The original system is of 6-th order and has 2 inputs and 2 outputs. Using the algorithm developed in SECTION 4, we generated transfer function matrices, such that each element of the estimated transfer function matrix had orders 5, 4 and 3. A comparison of the impulse responses of the lower order approximation with the actual one is expressed as SNR in second column of TABLE 4.3. Further, in FIGURES 1(a), 2(a), 3(a), 4(a) and 5(a) we have plotted the actual unit pulse response, and absolute errors corresponding to approximation of order 6, 5, 4 and 3 respectively. The low maagnitudes of the errors clearly indicate the effectiveness of the proposed technique.

The impulse response of the transfer function matrix was then corrupted by random noise such that the SNR was 20.5 dB. The noisy impulse response was used for estimating the parameters of the transfer function matrix. The SNR for the estimated system is tabulated in the third column of TABLE 4.3 below. Note that in computing the various $a_{ij}(z)$'s the algorithm uses the measured impulse response data. Therefore, it tries to match the noisy impulse response data. To make a fair evaluation of the performance of the algorithm, the SNR for the noisy case is computed by using the trailing elements of the estimated and the unit pulse response of the original transfer function matrix. In FIGURES 1(b), 2(b), 3(b), 4(b) and 5(b), we have plotted the noisy pulse response data and respectively the 6-th, 5-th, 4-th and the 3-rd order approximations.

TABLE 4.3: Simulation Results

Order	SNR (dB) Noiseless	SNR (dB) Noisy
6	96.6577	30.1450
5	78.5530	32.9529
4	45.9135	33.4319
3	31.5113	30.7571

Extensive simulations, with various SNR values show close approximation to noiseless impulse response. The results obtained using the proposed method consistently show better performance compared to the existing methods.

5. Concluding Remarks

In this paper we addressed the problem of identification of discrete time transfer function matrices from the noisy unit pulse response data. The proposed method is a generalization of an existing technique that estimates the parameters of a discrete time scalar transfer function. The simulation results presented here and extensive experience with the proposed scheme clearly indicate that it can be reliably used for estimating the parameters of discrete time transfer function.

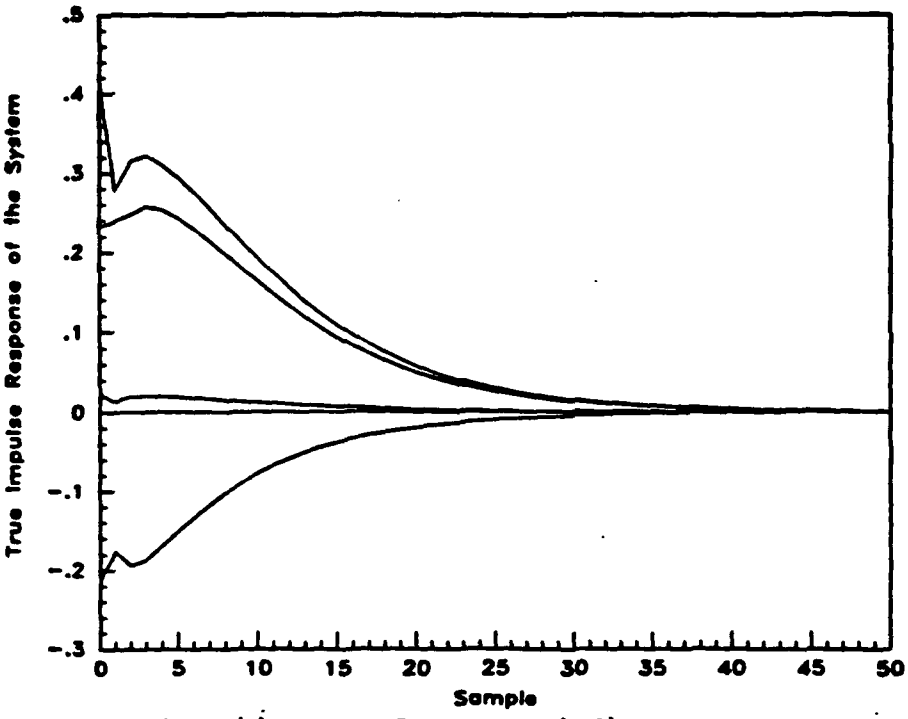

Fig. 1(a): Impulse Response of (2x2) System

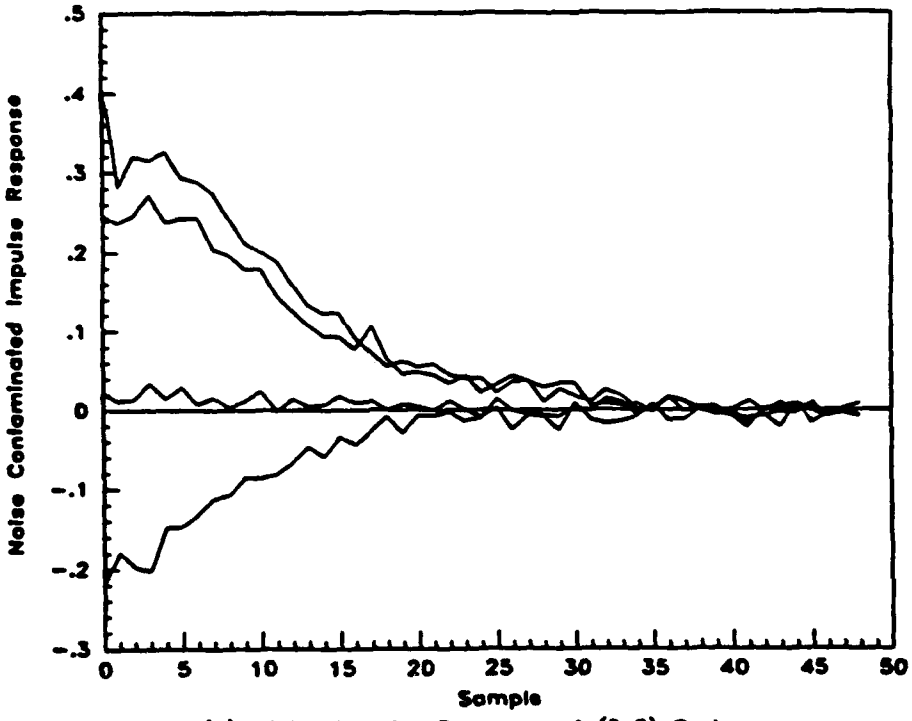

Fig. 1(b): Noisy Impulse Response of (2x2) System

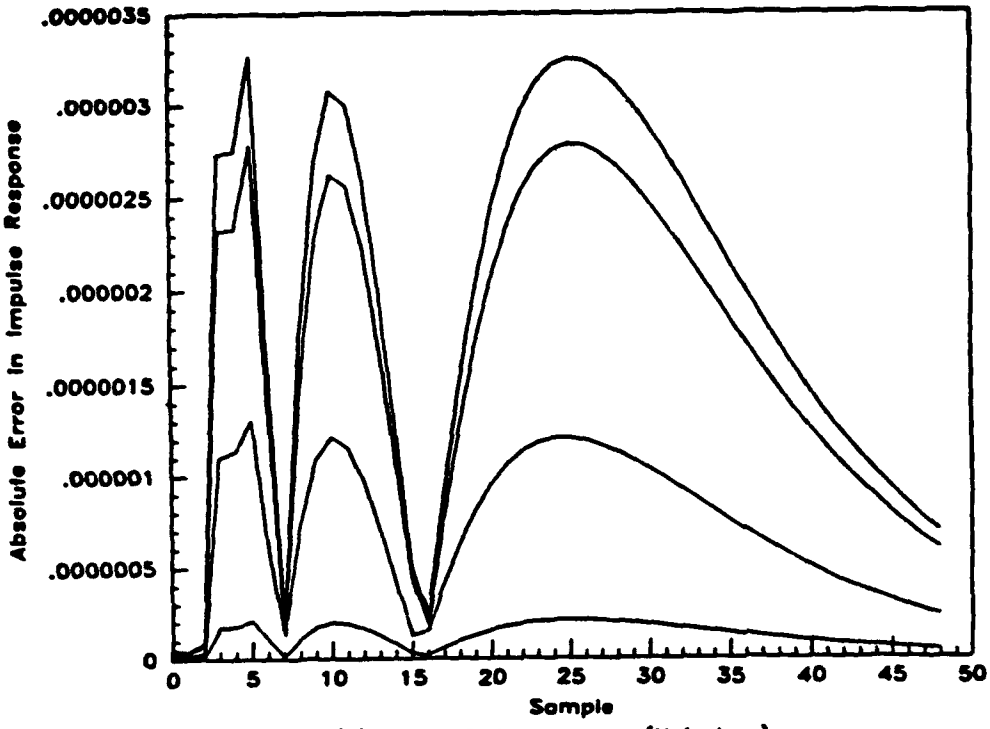

Fig. 2(a): 6-th Order Approx. (Noiseless)

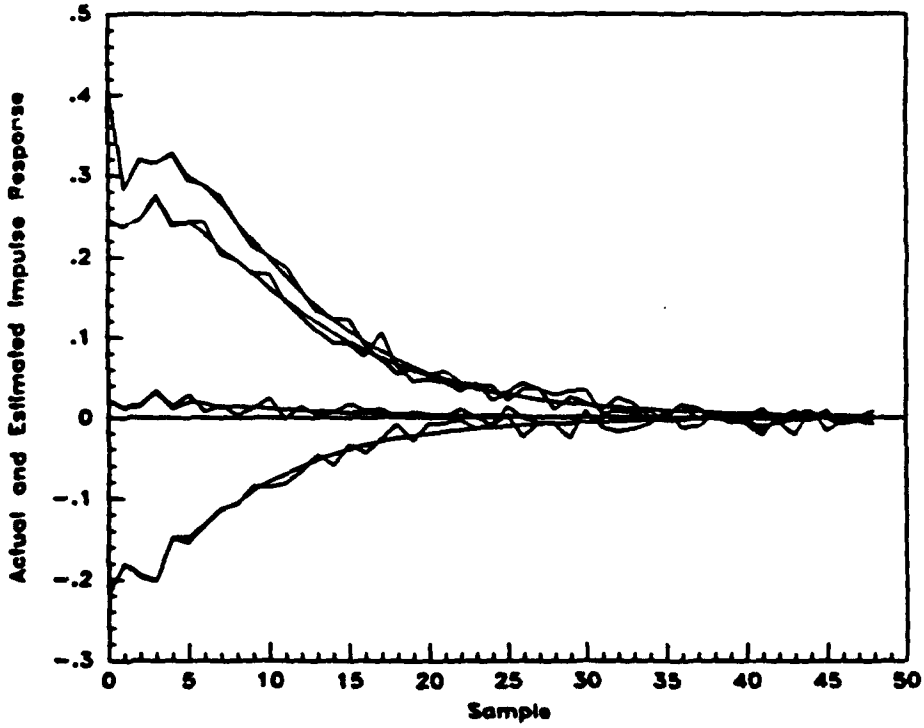

Fig. 2(b): 6-th Order Approx. (Noisy)

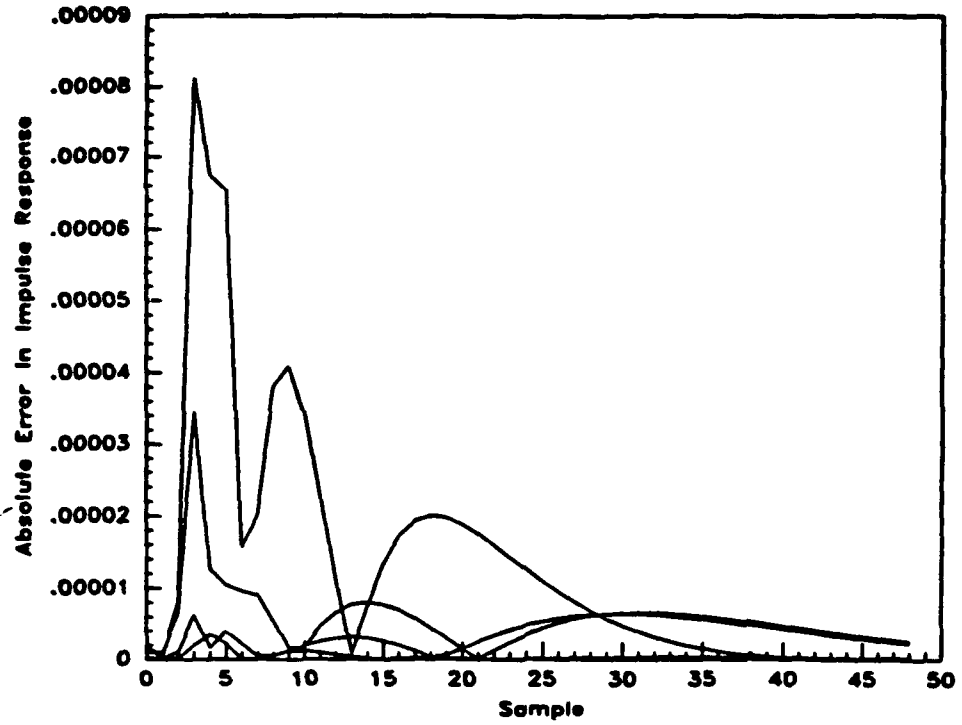

Fig. 3(a): 5-th Order Approx. (Noiseless)

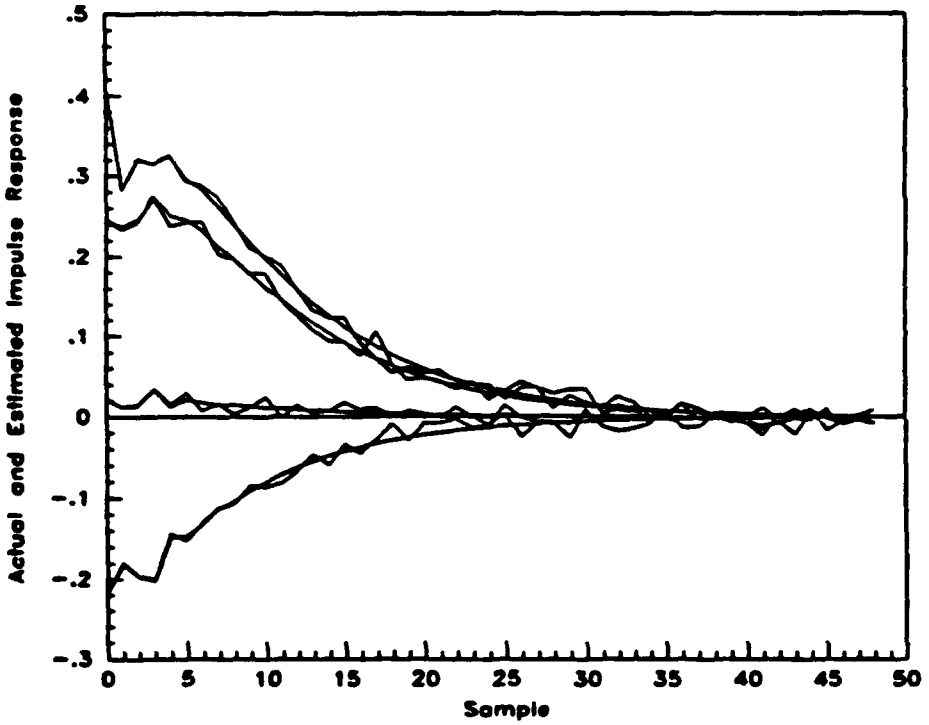

Fig. 3(b): 5-th Order Approx. (Noisy)

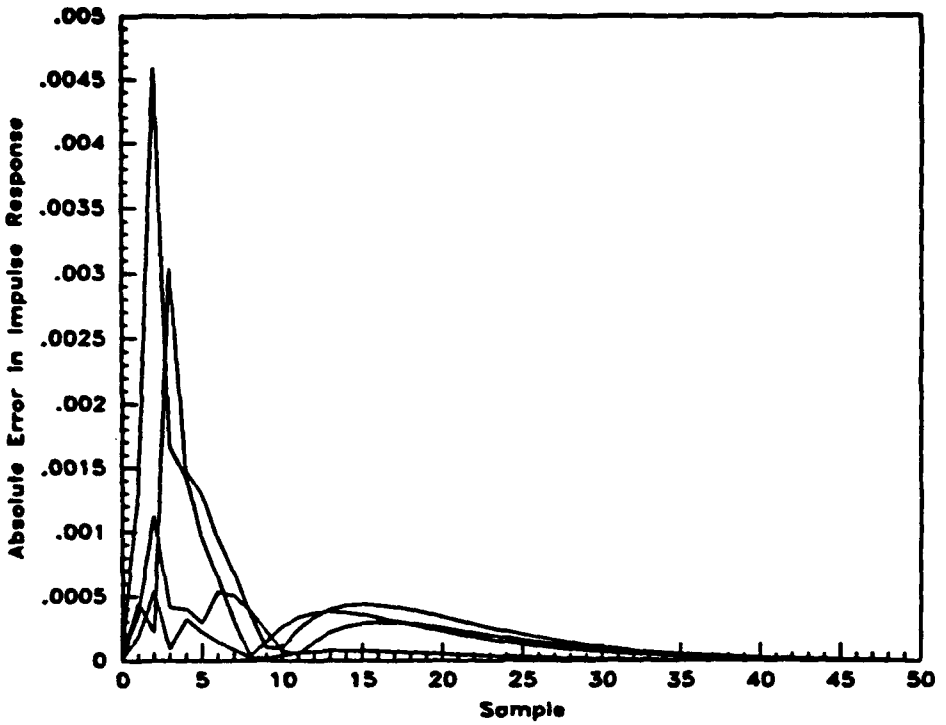

Fig. 4(a): 4-th Order Approx. (Noiseless)

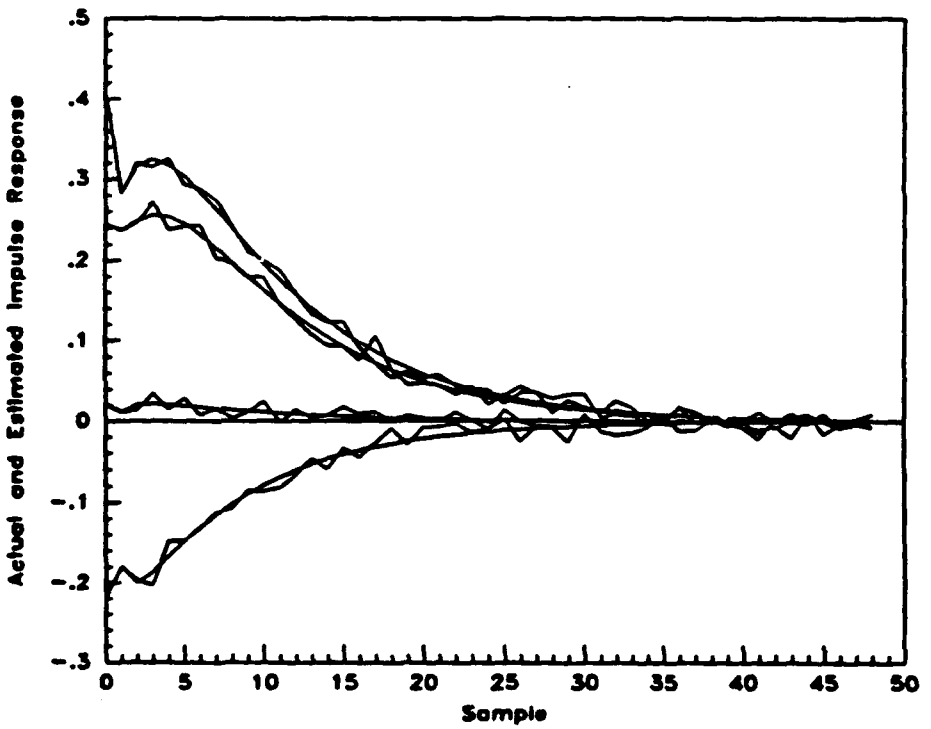

Fig. 4(b): 4-th Order Approx. (Noisy)

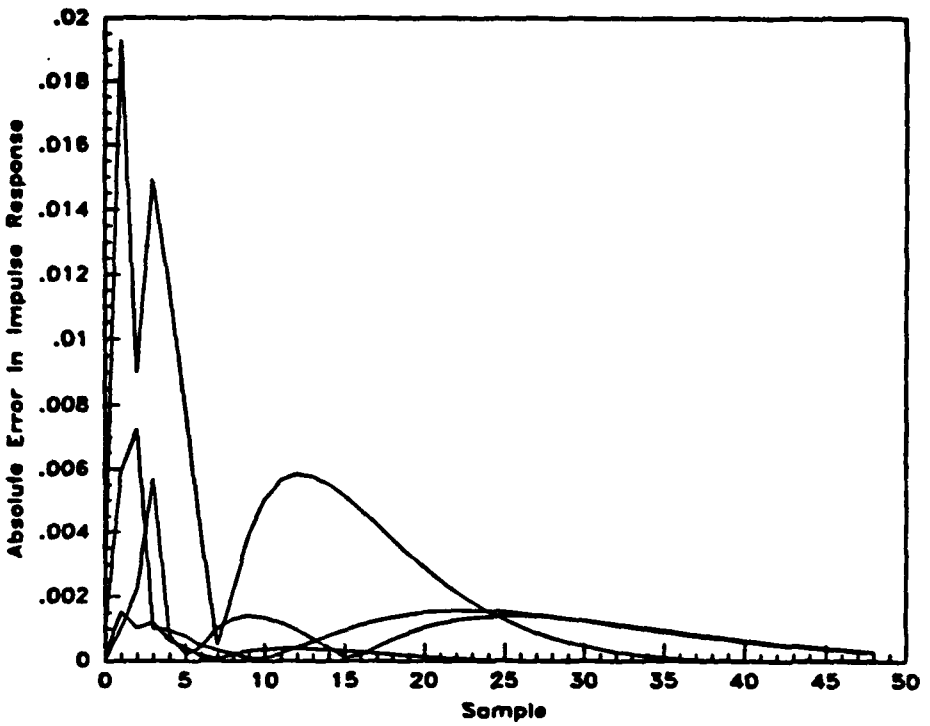

Fig. 5(a): 3-rd Order Approx. (Noiseless)

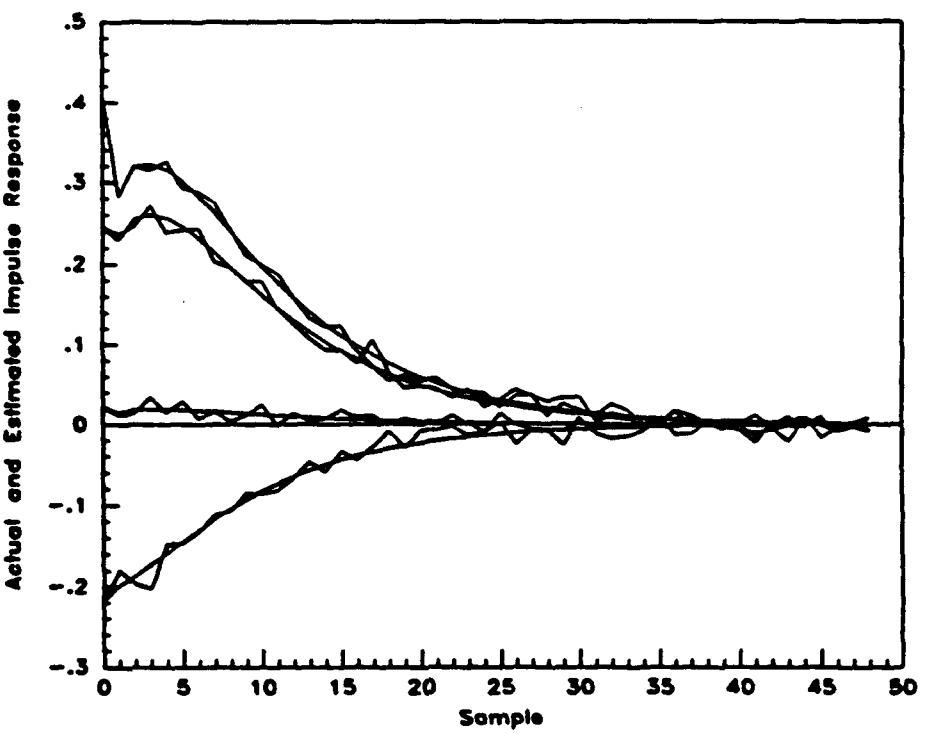

Fig. 5(b): 3-rd Order Approx. (Noisy)

5. REFERENCES

[1] R. Isermann, Ed., *Automatica*, Special issue on system identification, vol. 17, 1971.

[2] A.P. Sage and J.L. Melsa, *System Identification*, Academic Press, NY, 1971.

[3] K.J. Åström and P. Eykhoff, "System identification: a survey", *Automatica*, vol. 7, pp. 123-167, 1971.

[4] R.K. Mehra and D.G. Lainiotis, *System Identification-Advances and Case Studies*, Academic Press, NY, 1976.

[5] L.B. Jackson, *Digital Filters and Signal Processing*, Kluwer, Boston, 1986.

[6] J.P. Norton, *An Introduction to Identification*, Academic Press, NY, 1986.

[7] L. Ljung, *System Identification: Theory for the Users*, Prentice Hall, NJ, 1987.

[8] T. Söderström and P. Stoica, *System Identification*, Prentice Hall, NJ, 1987.

[9] S.M. Kay, *Modern Spectral Estimation: Theory and Applications*, Prentice Hall, Englewood Cliffs, NJ, 1988.

[10] K.J. Åström and B. Wittenmark, *Computer Controlled Systems: Theory and Design*, Prentice Hall, Englewood Cliffs, NJ, 1990.

[11] K. Glover and J.C. Willems, "Parameterizations of linear dynamical systems: Canonical forms and identifiability", *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 640-646, 1974.

[12] M.J. Denham, "Canonical forms for identification of multivariable linear systems", *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 646-656, 1974.

[13] M. Gevers and V. Wertz, "Uniquely identifiable state-space and ARMA parameterizations for multivariable linear systems", *Automatica*, vol. 20, pp. 333-347, 1984.

[14] M. Moonen and J. Vandewalle, "QSVD approach to on- and off-line state-space identification", *Int. J. Contr.*, vol. 51, pp. 1133-1146, 1990.

[15] A.J. Helmicki, C.A. Jacobson and C.N. Nett, "Identification in H^∞: A robust convergent nonlinear algorithm", *Proc. 1990 Amer. Contr. Conf.*, pp. 386-391, Pittsburgh, PA.

[16] G. Gu and P. Khargonekar, "Linear and non-linear algorithms for identification in H^∞", preprints

[17] P.M. Makila, "Approximation and identification of continuous-time systems", to appear in *Int. J. Contr.*

[18] G.P. Rao, *Piecewise Constant Orthogonal Functions and their Applications to Systems and Control*, Springer Verlag, NY, 1983. 1963.

[19] K. Steiglitz and L.E. McBride, "A Technique for Identification of Linear Systems", *IEEE Transactions on Automatic Control*, Vol. AC-10, pp. 461-464, 1965.

[20] A.G. Evans and R. Fischl, "Optimal Least Squares Time-Domain Synthesis of Recursive Digital Filters", *IEEE Transactions on Audio and Electro-Acoustics*, Vol. AU-21, pp. 61-65, 1973.

[21] D.G. Luenberger, *Optimization by Vector Space Method*, New York: Wiley, 1969.

[22] G. H. Golub and V. Pereyra [1973], "The Differentiation of Pseudoinverses and Nonlinear Problems Whose Variables Separate," *SIAM Journal on Numerical Analysis*, vol. 10, no. 2, pp. 413-432, Apr..

[23] B. Friedlander and B. Porat [1986], "Multichannel ARMA Spectral Estimation by the Modified Yule-Walker Method," *Signal Processing*, vol. 10, pp. 49-59.

APPENDIX

Computational Algorithm : Phase II

In this appendix the second phase of the iterative algorithm is described in detail. In this phase, the derivative of the matrix $\mathbf{W}_l$ *w.r.t.* $\hat{\mathbf{b}}$ is taken into account. The complete error expression is rewritten below,

$$\|\mathbf{e}_l(\mathbf{a}_l^*, \mathbf{b})\|_2^2 = \mathbf{e}_l^T(\mathbf{a}_l^*, \mathbf{b})\mathbf{e}_l(\mathbf{a}_l^*, \mathbf{b}). \tag{A.1}$$

By setting the derivative of this squared norm to zero, we obtain the updated $\hat{\mathbf{b}}^{(i+1)}$ at the $(i+1)$-th iteration as,

$$\hat{\mathbf{b}}^{(i+1)} = -[\mathbf{U}_l^{(i)}\mathbf{G}_l]^{-1}[\mathbf{U}_l^{(i)}]\mathbf{g}_l \tag{A.2}$$

where (suppressing the superscript $^{(i)}$),

$$\mathbf{U}_l \triangleq \mathbf{L}_l^T\mathbf{W}_l + \mathbf{G}_l\mathbf{W}_l^T\mathbf{W}_l, \tag{A.2a}$$

$$\mathbf{L}_l \triangleq \left[\frac{\partial \mathbf{W}_l}{\partial b(1)}\mathbf{d}_l(\mathbf{b}) \;\middle|\; \cdots \middle| \frac{\partial \mathbf{W}_l}{\partial b(n)}\mathbf{d}_l(\mathbf{b})\right], \tag{A.2b}$$

$$\frac{\partial \mathbf{W}_l}{\partial b(k)} \triangleq \left(I_{pm} \otimes \frac{\partial \mathbf{W}}{\partial b(k)}\right), \tag{A.2c}$$

$$\frac{\partial \mathbf{W}}{\partial b(k)} \triangleq \frac{\partial}{\partial b(k)}[\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}] = \frac{\partial \mathbf{B}}{\partial b(k)}(\mathbf{B}^T\mathbf{B})^{-1} - \mathbf{W}\left[\left[\frac{\partial \mathbf{B}^T}{\partial b(k)}\right]\mathbf{B} + \mathbf{B}^T\left[\frac{\partial \mathbf{B}}{\partial b(k)}\right]\right](\mathbf{B}^T\mathbf{B})^{-1} \quad \text{and} \tag{A.2d}$$

$\frac{\partial \mathbf{B}}{\partial b(k)}$ has the same form as the $\mathbf{B}$ matrix defined in (2.17) but filled with all zeros except at the locations where $b(k)$ appear. For example,

$$\frac{\partial \mathbf{B}}{\partial b(n)} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{A.2e}$$

Once $\hat{\mathbf{b}}^{(i+1)}$ is found, $\mathbf{b}^{(i+1)}$ can be found from,

$$\mathbf{b}^{(i+1)} = \begin{bmatrix} 1 \\ \cdots\cdots \\ \hat{\mathbf{b}}^{(i+1)} \end{bmatrix} \tag{A.3a}$$

$$= \begin{bmatrix} 1 \\ \cdots\cdots\cdots\cdots\cdots \\ -[\mathbf{U}_l^{(i)}\mathbf{G}_l]^{-1}[\mathbf{U}_l^{(i)}]\mathbf{g}_l \end{bmatrix}. \tag{A.3b}$$

This minimization phase continues until $\mathbf{b}^{i+1} \simeq \mathbf{b}^i$ is reached and this optimum $\mathbf{b}^*$ vector corresponds to a minimum of the error surface of $\|\mathbf{e}_l(\mathbf{a}_l^*, \mathbf{b})\|_2^2$.

CHAPTER 3

SPECTRUM ESTIMATION AND RELATED TOPICS

SECTION 3.1 : A NOVEL CYCLIC ALGORITHM FOR MAXIMUM LIKELIHOOD FREQUENCY ESTIMATION

SUMMARY

An algorithm for estimation of frequencies of narrow-band sources from noisy observation data is presented in this Section. For Gaussianly distributed noise, the algorithm produces maximum likelihood estimates, otherwise least-squares estimates are obtained. The proposed algorithm is iterative and at each step of iteration, the optimization is *w.r.t.* a single frequency only and hence, simple hardware/software (using FFT, *e.g.*) is sufficient for implementation. The performance of the algorithm has been compared with the theoretical Cramer-Rao bounds.

I. Introduction :

Estimation of frequencies from data composed of multiple narrowband signals in noise is one of the oldest as well as a current research problem that is of great interest in several branches of science. In the recent years several techniques that produce optimal estimates have been developed. Unfortunately, the optimal techniques are based on computation intensive nonlinear optimization procedures. Hence the optimal techniques, though theoretically sound, have seen limited practical usage. In fact, most of the well-known and established frequency estimation algorithms are actually suboptimal. The suboptimal algorithms are popular because they can be implemented relatively inexpensively and, except at low SNR, they perform equally effectively [Tufts and Kumaresan, 1982]. But if one is interested in real-time computation, especially at very high sampling rate needed for high frequency applications, both the optimal as well as the suboptimal techniques would require rather expensive special-purpose hardware/software. The motivation of the this work was to investigate the possibility of devising an algorithm that would rely on off-the-shelf hardware/software for implementation but would still be optimal in the Maximum Likelihood (ML) sense.

It is well known that if the observation data is composed of a single sinusoid in gaussianly distributed noise, the peak of the periodogram corresponds to the maximum likelihood (ML) estimate of the unknown frequency [Palmer, 1974; Rife and Boorstyn, 1974, 1976]. The hardware or software implementation of the periodogram is based on the Fast Fourier Transform (FFT) [Cooley and Tukey, 1965] which is a highly efficient but simple algorithm. Because of this simplicity, the periodogram indeed is the main workhorse for most practical applications even when more than one sinusoids are present. In case of multiple sinusoids, the effectiveness and applicability of periodogram is greatly diminished unless the unknown frequencies are well separated. The periodogram peaks in such cases, in general, do not correspond to the ML estimates. In fact, if the separation between two adjacent frequencies is less than the FFT bin width, a plot of the periodogram would only show one merged peak instead of two distinct ones.

Overcoming the resolution limit of the periodogram has been a major focus of research over the past decade. Periodogram is basically a brute-force method which does not make any explicit use of the exponential nature underlying the multiple sinusoids data. In contrast to that, the modern high-resolution techniques exploit the known information about the exponential character of the observed data and assume an appropriate model, either implicitly or explicitly. The problem then is converted to a multidimensional search over the parameter space

of the chosen model. In some cases [Parthasarathy and Tufts, 1985; Rife and Boorstyn, 1976], the unknown frequencies may themselves constitute the parameter search space. But these direct methods are based on the maximization of optimal criterion which require non-linear optimization.

One of the major objectives in developing suboptimal techniques has been to come up with linear solution for this essentially non-linear problem. The origin of development in this direction may be found in the algorithm due to Prony [1795]. Complex exponentials may be considered to be the roots of a Linear Predictor (LP) polynomial and the estimation of the LP coefficients (or equivalently, the parameters of an Auto-Regressive [AR] model) is a linear problem [Makhoul, 1975]. Realizing this unique property of complex exponentials, enormous research effort has concentrated on this particular idea [see Jackson et al, 1978; Tufts and Kumaresan, 1982; Lang and McClellan, 1980; Ulrych and Bishop, 1975; among others]. The roots of the estimated LP polynomial are the estimates of frequencies. It should also be noted that the parallel development of the maximum entropy method [Burg, 1967], which is based on a completely different theoretical viewpoint, essentially produces exactly same results as AR modeling. These methods have been shown to be significantly more effective when the corrupting influence of noise in the observed data or in the correlation matrix is reduced by the incorporation of Singular Value Decomposition (SVD) or Eigen-Decomposition (ED) as the case may be [Tufts and Kumaresan, 1982; Kay and Shaw [1988], Kung et. al. [1983] and others]. The ED of the correlation matrix and the SVD of the data matrix composed of the multiple frequencies in noise data possess certain signal/noise-subspace orthogonality properties which were also effectively exploited by many researchers for frequency estimation [Pisarenko, 1972; Owsley, 1978; Schmidt, 1979; Bienvenue and Kopp, 1979; Reddi, 1979; Kumaresan and Tufts, 1983, Kung et. al. [1983] and others]. More recently, a structured matrix approximation based method has been proposed which essentially reparameterizes the maximum likelihood criterion to depend on LP-type coefficients [Kumaresan and Shaw, 1985, 1988; Kumaresan, Scharf and Shaw, 1986; Shaw, 1987]. This reparameterization leads to an iterative method where the problem is inherently made linear at every iteration step. As one would expect, this ML algorithm gives accurate frequency estimates even at low SNR. References to various other methods of frequency estimation may be found in the papers cited here.

The main goal of the present work is to integrate the positive aspects of both periodogram as well as the model based estimation techniques. In this work, we propose an iterative algorithm which achieves this objective by essentially splitting the multidimensional non-linear optimization problem of MLE into several one-dimensional searches. The exact model of the multiple complex exponential data is invoked and the exact ML criterion is optimized *w.r.t.* a single frequency at every step while keeping the others fixed at previously estimated values. It may be emphasized here that the computational simplicity of the proposed approach comes from the fact that every iteration requires optimization with respect to only one frequency. This can be easily performed by finding the peak of the periodogram. The proposed method is iterative. It requires crude prior estimates or regions of interest of the frequencies. The initial estimates may again be obtained from the periodogram peaks or any other method. In fact, our simulations indicate the effectiveness of the algorithm does not diminish much, even if the initial estimates are chosen in a completely random manner.

This Section is arranged as follows: The problem is formulated in Subsection II and the proposed algorithm is described in Subsection III. Simulation results are given in Subsection IV and finally some concluding remarks have been included in Subsection V.

II. Formulation of the Problem :

Let $x(n)$, $n = 0, 1, \ldots, N-1$, be an observation data record of N consecutive samples of the multiple complex

exponential signal model which is defined as,

$$\tilde{x}(n) \triangleq \sum_{k=1}^{p} a_k e^{j(\omega_k n + \phi_k)} \quad n = 0, 1, \ldots, N-1 \quad \text{where,} \tag{II.1}$$

a_k : unknown real amplitude of the k^{th} sinusoid,
ϕ_k : unknown phase of the k^{th} sinusoid,
ω_k : unknown angular frequency of the k^{th} sinusoid and
p : assumed number of sinusoids.

The observation samples are expressed as,

$$x(n) = \tilde{x}(n) + z(n), \tag{II.2}$$

where, $z(n)$ represents observation noise and/or modeling error. In vector form, the observed samples $x(n)$, the model samples $\tilde{x}(n)$ and the noise samples $z(n)$, for $n = 0, 1, \ldots, N-1$, are related as,

$$\mathbf{x} \triangleq \tilde{\mathbf{x}} + \mathbf{z} \quad \text{where,} \tag{II.3}$$

$$\mathbf{x} \triangleq [x(0)\ x(1)\ \ldots,\ x(N-1)]^t \tag{II.4a}$$
$$\tilde{\mathbf{x}} \triangleq [\tilde{x}(0)\ \tilde{x}(1)\ \ldots,\ \tilde{x}(N-1)]^t \quad \text{and} \tag{II.4b}$$
$$\mathbf{z} \triangleq [z(0)\ z(1)\ \ldots,\ z(N-1)]^t, \tag{II.4c}$$

where, 't' denotes matrix or vector transpose. The multiple complex sinusoids model vector $\tilde{\mathbf{x}}$ is equivalently described by the following matrix-vector decomposition,

$$\tilde{\mathbf{x}} = \mathbf{T}\mathbf{a} \quad \text{where,} \tag{II.5}$$

$$\mathbf{T} \triangleq \begin{pmatrix} 1 & 1 & \ldots & 1 \\ e^{j\omega_1} & e^{j\omega_2} & \ldots & e^{j\omega_p} \\ \vdots & \vdots & \ddots & \vdots \\ e^{j\omega_1(N-1)} & e^{j\omega_2(N-1)} & \ldots & e^{j\omega_p(N-1)} \end{pmatrix} \text{ and } \mathbf{a} \triangleq \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_p \end{pmatrix} \tag{II.6}$$

where, $A_k \triangleq a_k e^{j\phi_k}$, for $k = 1, 2, \ldots, p$, respectively, are the complex amplitudes. The problem under consideration here is to choose or estimate the best model parameters $A_1, A_2, \ldots, A_p$, and $\omega_1, \omega_2, \ldots, \omega_p$ such that the modeling error norm,

$$\|\mathbf{e}\|^2 \triangleq E(\mathbf{T}, \mathbf{a}) \triangleq \|\mathbf{x} - \mathbf{T}\mathbf{a}\|_2^2 \tag{II.7}$$

is minimized.

Next, a brief derivation is given to show that the least-squares criterion in (II.7) is indeed the one required for MLE when the noise samples $z(n)$ are white and gaussianly distributed. More details may be found in [Rife and Boorstyn, 1976]. If the observed samples $\mathbf{x}$ in (II.4a) are composed of multiple sinusoids in Gaussianly distributed, zero-mean and complex white noise, the probability density function (PDF) of $\mathbf{x}$ is given by,

$$\mathcal{P}(\mathbf{x} - \tilde{\mathbf{x}}) = \frac{1}{\pi^N det(\mathbf{R}_{zz})} e^{-(\mathbf{x}-\tilde{\mathbf{x}})^H \mathbf{R}_{zz}^{-1} (\mathbf{x}-\tilde{\mathbf{x}})} \tag{II.8}$$

where $\tilde{\mathbf{x}}$ is defined in (II.4b) and $\mathbf{R}_{zz}$ is the $N \times N$ autocorrelation matrix of the noise. Since the noise is assumed to be white,

$$\mathbf{R}_{zz} = \sigma_z^2 \mathbf{I}. \tag{II.9}$$

To obtain the MLE of the unknown parameters one needs to optimize the following criterion,

$$\max_{\{A_k\},\{\omega_k\}} \mathcal{P}(\mathbf{x} - \tilde{\mathbf{x}}) = \max_{\{A_k\},\{\omega_k\}} \frac{1}{\pi^N \sigma_x^{2N}} e^{-\frac{1}{\sigma_x^2}(\mathbf{x}-\tilde{\mathbf{x}})^H(\mathbf{x}-\tilde{\mathbf{x}})}. \quad (II.10)$$

This criterion is exactly equivalent to

$$\min_{\{A_k\},\{\omega_k\}} (\mathbf{x} - \tilde{\mathbf{x}})^H(\mathbf{x} - \tilde{\mathbf{x}}). \quad (II.11)$$

Now, writing the model vector $\tilde{\mathbf{x}}$ explicitly from (II.5) and defining the fitting error as,

$$\mathbf{e}(\mathbf{T}, \mathbf{a}) \triangleq \mathbf{x} - \mathbf{T}\mathbf{a}, \quad (II.12)$$

the criterion becomes

$$\min_{\{A_k\},\{\omega_k\}} E(\mathbf{T}, \mathbf{a}) \triangleq \min_{\{A_k\},\{\omega_k\}} \mathbf{e}^H(\mathbf{T}, \mathbf{a})\mathbf{e}(\mathbf{T}, \mathbf{a}). \quad (II.13)$$

This criterion is exactly same as the least squares fitting criterion in (II.7). Note that the error $\mathbf{e}(\mathbf{T}, \mathbf{a})$ in (II.13) is linearly related to the parameters in $\mathbf{a}$ while it is nonlinearly related to the frequencies in $\mathbf{T}$. Hence the minimization of $E(\mathbf{T}, \mathbf{a})$ is a non-linear multidimensional optimization problem.

Considerable work have been reported on the direct optimization of the criterion in (II.11). Notable among them are the work by Golub and Pereyra [1975], Parthasarathy and Tufts [1984] and Rife and Boorstyn [1976]. These approaches are mainly based on Newton-Raphson or Gauss-Newton type algorithms. Performance of the general non-linear multidimensional optimization algorithms depend primarily on the choice of initial estimates. The proposed algorithm is described next.

III. The Maximum-Likelihood Algorithm :

In order to motivate the proposed approach let us consider an ideal case first. To facilitate the splitting of the multi-dimensional optimization criterion in (II.13) into several 1-D optimization problems, we rewrite the Vandermonde matrix $\mathbf{T}$ in (II.6) in the following form,

$$\mathbf{T} \triangleq [\mathbf{t}_1\ \mathbf{t}_2\ \ldots \mathbf{t}_k \ldots \mathbf{t}_p]^t \quad \text{where,} \quad (III.1)$$

$$\mathbf{t}_k \triangleq [1\ e^{j\omega_k} e^{j2\omega_k} \ldots e^{j(N-1)\omega_k}]^t. \quad (III.2)$$

Next, a Vandermonde matrix $\mathbf{T}_k$ is formed with $(p-1)$ of the p frequencies and excluding the $\mathbf{t}_k$ vector, *i.e.*,

$$\mathbf{T}_k \triangleq [\mathbf{t}_1\ \mathbf{t}_2\ \ldots \mathbf{t}_{k-1}\ \mathbf{t}_{k+1} \ldots \mathbf{t}_p]^t. \quad (III.3)$$

Similarly, define the corresponding amplitude vector as,

$$\mathbf{a}_k \triangleq [a_1\ a_2\ \ldots a_{k-1}\ a_{k+1} \ldots a_p]^t. \quad (III.4)$$

Using the above notations, the model vector $\tilde{\mathbf{x}}$ of (II.5) is rewritten as,

$$\tilde{\mathbf{x}} = \mathbf{T}\mathbf{a} = \mathbf{T}_k\mathbf{a}_k + \mathbf{t}_k A_k. \quad (III.5)$$

Plugging this in (II.12) we get,

$$\mathbf{e}(\mathbf{T}, \mathbf{a}) = \mathbf{x} - \mathbf{T}_k\mathbf{a}_k - \mathbf{t}_k A_k. \quad (III.6)$$

If the frequency and the amplitude terms in $\mathbf{T}_k$ and $\mathbf{a}_k$ are known exactly, the error can be rewritten as,

$$\mathbf{e}(\mathbf{t}_k, A_k) = \mathbf{x}_k^m - \mathbf{t}_k A_k, \tag{III.7}$$

where, $\mathbf{x}_k^m$ is a modified data vector defined as,

$$\mathbf{x}_k^m \triangleq \mathbf{x} - \mathbf{T}_k \mathbf{a}_k. \tag{III.8}$$

Clearly, the minimization of the norm of the error in (III.7), *i.e.*,

$$\min_{\omega_k, A_k} \|\mathbf{x}_k^m - \mathbf{t}_k A_k\|_2^2 \tag{III.9}$$

is a one-dimensional optimization problem. In fact, the maximum-likelihood estimate of the unknown frequency ω_k in $\mathbf{t}_k$ can be found from the peak location of periodogram plot of the modified 'data' vector $\mathbf{x}_k^m$. Furthermore, the minimization in (III.9) can be carried out for each $k = 1, 2, \ldots, p$ and the estimates of all the p unknown frequencies can be found by performing p FFTs!

The approach outlined above seems perfect except that it presumes the ideal case of exact knowledge of the frequencies which, in fact, we sought to evaluate in the first place. In practical situations, the $p-1$ frequencies in $\mathbf{T}_k$ and the corresponding amplitudes in $\mathbf{a}_k$ which are needed to form the $\mathbf{x}_k^m$ in (III.8) will not be known. Instead, we propose to replace $\mathbf{T}_k$ and $\mathbf{a}_k$ by the corresponding estimates. Then the optimization can not be accomplished with only one set of p FFTs as anticipated. In that case, the optimization procedure would have to be done iteratively. The iterative algorithm is outlined next.

Let us first assume that approximate initial estimates of $(p-1)$ of the frequencies and corresponding amplitudes are available (from periodogram or linear prediction or any other method) and that the k^{th} of the p frequencies is unknown or needs to be updated. Now, separating the known and unknown parts of $\mathbf{T}$, we can write the observation vector $\mathbf{x}$ in (II.3) as,

$$\begin{aligned} \mathbf{x} &\triangleq \hat{\mathbf{T}}_k \hat{\mathbf{a}}_k + \mathbf{t}_k A_k + \mathbf{z} \\ &\triangleq \hat{\mathbf{x}}_k + \mathbf{t}_k A_k + \mathbf{z} \end{aligned} \tag{III.10}$$

where $\hat{\mathbf{T}}_k$ and $\hat{\mathbf{a}}_k$ have the same forms as in (III.3) and (III.4) except that the exact values are replaced by the corresponding estimated values and $\hat{\mathbf{x}}_k \triangleq \hat{\mathbf{T}}_k \hat{\mathbf{a}}_k$. Now defining $\mathbf{x}_k \triangleq \mathbf{x} - \hat{\mathbf{x}}_k$, (III.10) can be rewritten as,

$$\mathbf{x}_k \triangleq \mathbf{t}_k A_k + \mathbf{z}. \tag{III.11}$$

We now treat the vector $\mathbf{x}_k$ as the 'data' vector and correspondingly rewrite the error criterion for the k-th frequency in (III.9) as,

$$\begin{aligned} \min_{\omega_k, A_k} \|\mathbf{e}_k\|^2 &\triangleq \min_{\omega_k, A_k} E(\mathbf{t}_k, A_k) \\ &\triangleq \min_{\omega_k, A_k} \|\mathbf{x}_k - \mathbf{t}_k A_k\|_2^2. \end{aligned} \tag{III.12}$$

Once again, the optimization problem in (III.12) is with respect to the parameters associated with a single frequency only. Once the k^{th} frequency is estimated, the corresponding amplitude may be obtained by the following pseudo-inverse solution,

$$\hat{A}_k = \frac{1}{N} \hat{\mathbf{t}}_k^{\mathbf{H}} \mathbf{x}_k. \tag{III.13}$$

The updated estimates of ω_k and A_k can then be included to form $\hat{\mathbf{T}}_{k+1}$ and $\hat{\mathbf{A}}_{k+1}$ similar to (III.3) and (III.4). The same steps may then be followed for updating the $k+1^{th}$ frequency by optimization of the following criterion,

$$\min_{\omega_{k+1}, A_{k+1}} \|\mathbf{x}_{k+1} - \mathbf{t}_{k+1} A_{k+1}\|_2^2. \qquad (III.14)$$

In this manner at each updating step one of the frequency estimates is re-evaluated while keeping the other $p-1$ frequencies fixed at previously estimated values. At every iteration level the updating starts with $k = 1$ and goes up to $k = p$, exhausting all the frequency components. The iterations are continued till no further changes are observed in the frequency estimates. Note that at every step of optimization of (III.12), the ML criterion in (II.7) is minimized, albeit with respect to one frequency at a time. Hence, at convergence the ML estimates are obtained.

The attractiveness of the proposed algorithm comes from its sheer simplicity in implementation. The algorithm may also be used as an adaptive frequency tracker. In a radar or sonar environment, the spatial frequencies which are related to the angular positions of far-field sources may change as the sources move [Kumaresan and Shaw, 1987], the proposed algorithm will be able to adapt to such changing environment. In the next two paragraphs we consider the important issues of model order selection and the choice of the initial estimates.

Model Order Selection : In the above discussion of the algorithm, we have assumed that the model order p is exactly known. In practice, one may not have that knowledge and the model order needs to be estimated. Two of the most commonly used order selection criteria are the Akaike Information Criterion (AIC) [Akaike, 1974] and the Minimum Description Length (MDL) criterion [Rissanen, 1978, 1983]. Both these criterions are perfectly suited for the proposed approach because both AIC and MDL criterions are computed using the logarithm of the minimized error norm in (II.7) along with the number of parameters under optimization. MDL also requires the observation data length N.

Choice of Initial Estimates of frequencies : Any estimation technique such as linear prediction or coarse periodogram peaks may be used to get the initial estimates. In our simulations we chose the periodogram peak locations as the initial estimates of frequencies. If there is a single merged peak indicating the possibility of more than one frequency in that region, one may choose the peak locations as the possible frequencies and the other/s may be chosen within the peak lobe width according to the number of sinusoids present. But the algorithm seemed to be highly robust in the sense that in our simulations the performance did not deteriorate even for random choice of initial estimates.

IV : Simulation Results

The algorithm described above has been tested on simulated data. The following formula is used to generate the data,

$$x(n) = a_1 e^{j\omega_1} + a_2 e^{j\omega_2 n} + z(n) \qquad (IV.1)$$
$$n = 0,\ 1,\ \ldots, 24$$

where $\omega_1 = 2\pi f_1, \omega_2 = 2\pi f_2$, f_1 and f_2 being 0.23 and 0.26, respectively, $a_1 = a_2 = 1$ and $z(n)$ is a computer generated white complex gaussianly distributed noise sequence with variance $2\sigma^2$. σ^2 is the variance of the real and the imaginary parts of $z(n)$. SNR is defined as 10 $log_{10}(|\ a_i\ |^2\ /2\sigma^2)$. One hundred sets of samples with different noise epochs were used.

For the observation records described above the error criterion given in (II.3) was minimized by following the algorithm described in Subsection III. The periodogram peak location with a 64-point FFT was used to find

the coarse initial estimate to form $\hat{\mathbf{T}}_1$ and from that $\hat{A}_1$. In the figure below, $10 \log \frac{1}{MSE}$ calculated with the bias and variance for 100 trials at different SNR values are plotted. Cramer-Rao bounds (CRB), which give the lower bound on the variance of f_1 (or f_2) for the corresponding SNR values, are also plotted. The results indicate that the bias in the estimates is negligible and the Mean Squared Errors remain close to the CRB up to about 0dB SNR. The iterative procedure always converged at every trial in 5 to 15 iterations. The present algorithm seems to push the SNR threshold even lower than reported results. More studies and comparisons with other algorithms need too done. In order to test the robustness and sensitivity of the algorithm to initial estimates, we also ran simulations with initial frequency estimates chosen completely randomly from a uniform noise generator. It was found that the algorithm performed almost as well as in the previous experiment. This robustness aspect of the algorithm should be studied more in future.

V : Conclusion

A new algorithm for Maximum-Likelihood frequency estimation is presented. Unlike all known ML methods for frequency estimation, the proposed algorithm does not require any multidimensional optimization. The multidimensional problem is split into several 1-D optimization problems. For computation, the algorithm only requires the FFT algorithm which is extensively used in signal processing. Our simulations indicate that the algorithm pushed the threshold down to 0dB SNR. The iterative algorithm also showed remarkable robustness to the choice of initial estimates. We intend to extend the algorithm to the more general case of frequency-wavenumber estimation and also for Toeplitz matrix approximation.

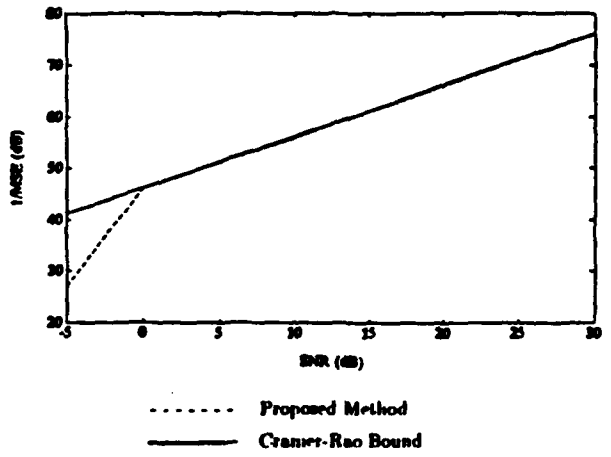

Fig. Comparison of the performance of the proposed method with the Theoretical Cramer-Rao bound.

Bibliography

[1] H. Akaike [1974], "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, vol. 34, pp. 716-723.

[2] G. Bienvenue and L. Kopp [1979], "Principle de la goniometric od passive adaptive," *Proceedings 7'eme Colloque GRESTI*, Nice, France, pp. 106/1-106/10.

[3] Y. Bresler and A. Mackovski [1985], "Exact Maximum Likelihood Estimation of Superimposed Exponential Signals in Noise," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Tampa, Florida, pp.1824-1827.

[4] J. P. Burg [1967], "Maximum Entropy Spectral Analysis," presented at the *37th Annual International SEG Meeting*, Oklahoma City, OK.

[5] J. W. Cooley and J. W. Tukey [1965], "An Algorithm for the Machine Calculation of Fourier Series," *Math. Comput.*, vol. 19, pp. 297-301.

[6] G. H. Golub and V. Pereyra [1973], "The Differentiation of Pseudoinverses and Nonlinear Problems Whose Variables Separate," *SIAM Journal on Numerical Analysis*, vol. 10, no. 2, pp. 413-432, Apr..

[7] S. Haykin et al, Editors [1985], *Array Signal Processing*, Prentice-Hall.

[8] L. B. Jackson et al. [1979], "Frequency Estimation by Linear Prediction," in the *Proceedings of the IEEE International Conference of Acoustics, Speech and Signal Processing-1979*, Washington, DC, pp. 352-356, Apr..

[9] S. M. Kay and A. K. Shaw [1988], "Frequency Estimation by Principal Component Autoregressive Spectral Estimator Without Eigendecomposition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol-36, no. 1, pp. 95-101.

[10] R. Kumaresan [1982(b)], *Estimating the Parameters of Exponentially Damped and Undamped Sinusoidal Signals*, Ph. D. Dissertation, University of Rhode Island.

[11] R. Kumaresan and D. W. Tufts [1983], "Estimating the Angles of Arrival of Multiple Planewaves," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-19, no .1, pp. 134-139, Jan..

[12] R. Kumaresan and A. K. Shaw [1985], "High Resolution Bearing Estimation Without Eigendecomposition," *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, Florida, April.

[13] R. Kumaresan, L. L. Scharf and A. K. Shaw [1986], "An Algorithm for Pole-Zero Modeling and Spectral Estimation," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol.ASSP-34, pp. 637-640, June.

[14] R. Kumaresan and A. K. Shaw [1986], "An Exact Least Squares Fitting Technique for Two-Dimensional Frequency Wavenumber Estimation," *Proceedings of the IEEE*, vol. 74, no. 4, pp. 606-607, April.

[15] R. Kumaresan and A. K. Shaw [1988], "Superresolution by Structured Matrix Approximation," *IEEE Transactions on Antennas and Propagation*, vol. 36, no. 1, pp. 34-44.

[16] S. Y. Kung, K. S. Arun and D. V. Bhaskar Rao [1983], "State-Space and Singular-Value Decomposition-Based Approximation Methods for the Harmonic Retrieval Problem," *Journal of the Optical Society of America*, vol. 73, pp. 1799-1811, Dec..

[17] S. W. Lang and J. M. McClellan [1980], "Frequency Estimation with Maximum Entropy Spectral Estimator," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol.ASSP-28, pp. 716-724.

[18] J. Makhoul [1975], "Linear Prediction : A Tutorial Review," *Proceedings of the IEEE*, vol. 63, pp. 561-580,

April.

[19] N. L. Owsley [1978], "Data Adaptive Orthonormalization," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Tulsa, OK, pp. 109-112.

[20] L. C. Palmer [1974], "Coarse Frequency Estimation Using the Discrete Fourier Transform," *IEEE Transactions on Information Theory*, Vol. IT-20, pp. 104-109, Jan..

[21] S. Parthasarathy and D. W. Tufts [1985], "Maximum-Likelihood Estimation of Parameters of Exponentially Damped Sinusoids," *IEEE Proceedings*, vol. 73, no. 10, pp. 1528-1530, Jan..

[22] V. F. Pisarenko [1973], "The Retrieval of Harmonics from Covariance Functions," *Geophysical Journal of the Royal Astronomical Society*, Vol. 33, pp. 347-366.

[23] R. Prony [1795], "Essai Experimental et Analytique etc.," L'Polytechnique, Paris, 1 Cahier 2, pp. 24-76.

[24] S. S. Reddi [1979], "Multiple Source Location- A Digital Approach," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-15, no.1, pp. 95-105.

[25] D. C. Rife and R. R. Boorstyn [1974], "Single Tone Parameter Estimation from Discrete-Time Observations," *IEEE Transactions on Information Theory*, Vol. IT-20, pp. 591-598, Sept..

[26] D. C. Rife and R. R. Boorstyn [1976], "Multiple Tone Parameter Estimation from Discrete Time Observations," *Bell Systems Technical Journal*, vol. 55, pp. 1389-1410.

[27] J. Rissanen [1978], "Modeling by Shortest Data Description," *Automatica*, vol. 14, pp. 465-471.

[28] J. Rissanen [1983], "A Universal Prior for the Integers and Estimation by Minimum Description Length," *Annals of Statistics*, vol. 11, pp. 417-431.

[29] R. O. Schmidt [1979], "Multiple Emitter Location and Signal Parameter Estimation," *Proceedings of RADC Spectral Estimation Workshop*, pp. 243-258, Rome, New York.

[30] A. K. Shaw [1987], "Structured Matrix Approximation Problems in Signal Processing," Ph. D. Dissertation, Dept. of Elect. Engg., University of Rhode Island.

[31] A. K. Shaw and R. Kumaresan [1986], "Frequency-Wavenumber Estimation by Structured Matrix Approximation," *Proceedings of the 3rd. IEEE-ASSP Workshop on Spectral Estimation*, Boston, pp. 81-84, Nov..

[32] D. W. Tufts and R. Kumaresan [1982], "Frequency Estimation of Multiple Sinusoids : Making Linear Prediction Perform Like Maximum Likelihood," *Proceedings of the IEEE*, vol. 70, pp. 975-989. Sept..

[33] T. J. Ulrych and T. N. Bishop [1975], "Maximum Entropy Spectral Analysis and Autoregressive Decomposition," *Rev. Geophysics and Space Physics*, vol. 13, pp. 183-200, Feb..

[34] M. Wax and T. Kailath [1983], "Optimum Localization of Multiple Sources by Passive Arrays," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.ASSP-31, no.5, pp.1210-1218, Oct..

SECTION 3.2 : A PARAMETER ADAPTIVE SIMULATED ANNEALING SCHEME FOR FREQUENCY ESTIMATION

SUMMARY

A Simulated Annealing scheme based on parameter adaptive cooling schedule is proposed. In the existing annealing schemes, the temperature parameter is predetermined for every iteration step and is independent of the unknown parameter values. In the proposed scheme, the cooling temperature is made proportional to the deviation of each individual parameter at the earlier iteration step. The other key difference in the proposed scheme is that it never accepts a higher energy level and remains at the present lower energy position. Instead, the Boltzmann Distribution is used to accept a larger cooling temperature *i.e*, a broader parameter search space. The algorithm is then applied to the well known nonlinear optimization problem of frequency/angles of arrival estimation of multiple sources. Simulation results indicate that the proposed scheme converges to the minimum energy level in fewer iteration steps when compared to an existing fast annealing algorithm.

Introduction

Multi-dimensional and non-linear optimization problems occur in engineering, economics, geophysics, as well as in almost all fields of science. A plethora of literature on standard optimization algorithms are available in the literature but in this decade a new and powerful optimization algorithm called *Simulated Annealing* has emerged and has found ever increasing attention in many applications. Although simulated annealing was originally proposed by Metropolis in 1953, only recently it has found successful applications in constrained and unconstrained optimization problems [Kirkpatrick et al, 1983; Gelfand and Mitter, 1985; El Gamal et. al, 1987, Sharman, 1988]. Also very recently, Szu [1986] has developed a fast annealing scheme which has improved convergence rates. It is well known that the performance of almost all other existing nonlinear optimization techniques are highly sensitive to the initial estimates and most algorithms cannot come out of a local optimum if it happens to reach one during the iterative process. Simulated annealing is a form of stochastic optimization and two of its most important and exciting features are that it can *escape* local stationary points of a cost function and that, 'theoretically', it is guaranteed to reach the *global optimum*.

Although the presently available annealing algorithms are quite attractive it is generally felt that even the fast annealing technique by Szu requires considerable computations. This is mainly due to the large number of iteration steps involved. A further reduction in the convergence rate will have far reaching advantages in many applications. The major goal of the proposed research is to explore a possible annealing scheme which, according to preliminary studies, seems to provide faster convergence than existing techniques.

Simulated Annealing and the Globally Optimal Design

The idea of simulated annealing may be understood in the following manner. Assume that we have a box containing an unknown 3-dimensional terrain with valleys and peaks at unknown locations and we are interested in finding the deepest point in the terrain. A simple method would be to take a ball and drop it inside the box containing the terrain because the ball will eventually roll down and reach the deepest point in the valley in which it was dropped. Of course, if the ball is dropped in another location which may be another valley, it would roll down to the deepest point of that particular valley. Unfortunately, none of these two deep points of the individual valleys may be the deepest point of the whole terrain which is what we were seeking in the first place. This is exactly how the conventional gradient based optimization algorithms work. The terrain surface

may represent the multimodal surface of a cost criterion $C(\theta)$ one is trying to minimize where, θ is a vector containing the underlying unknown parameters. To reach a minimum, one supplies an arbitrary initial estimate. The conventional optimization algorithms, which are mainly gradient based, would take one to the closest local minimum by going in the direction of decreasing gradient (*i.e*, decreasing cost) until zero gradient is reached. But again, if the cost function has multiple local minima, one may not have reached the global minimum of $C(\theta)$. The whole thing really boils down to the choice of the initial estimate. A straight-forward solution for this problem would be to compute the cost criterion at all possible values of the underlying parameters and choose the set which corresponds to the lowest cost. Unfortunately, this simple solution will be prohibitively expensive even for a small number of parameters with infinite possibilities for each. But consider the ball-in-the-terrain-box scenario again and assume that the ball is resting at one of the local minimum. Now, if one can somehow agitate the terrain box and do it vigorously enough such that there is a positive probability that the ball will be dislocated from its resting position and will jump to a position with higher cost but in another valley of the terrain. Then one can start the search process all over again. In this manner, with a large enough number of attempts, one would ultimately reach the valley containing the global optimum. The problem then is to seek the global optimum and remain there. This, indeed, is the fundamental premise of simulated annealing, *i.e*, it can *escape* a local minimum of a multimodal cost criterion and can ultimately reach the global optimum.

From the description given above, it is obvious that one has to be very careful in how much one should agitate the box when the ball is resting at one of the deeper points. If one shakes the box too much, the ball may even jump from the globally deepest point and if one shakes the box too slow, the ball may remain trapped forever inside a valley with a local optimum. Both of these would be undesirable and hence, one needs a proper strategy. In true annealing, a heated solid is allowed to cool to its minimum energy state. The molecules use random motion to search for new positions of lower energy. The likelihood of reaching the least energy state depends on how fast the solid is cooled. To apply this analogy to the cost minimization problem, it is necessary to meet the following four objectives [Szu, 1986], namely,

(i) A cost criterion, $C(\theta)$ which we seek to optimize, where $\theta \underline{\underline{\Delta}} [\theta^1 \ \theta^2 \ldots \theta^p]^t$ is the vector containing the p unknown parameters. "^{t}" denotes the matrix or vector transpose operation.

(ii) A rule for generating new candidate parameter estimates : The fast annealing scheme employs a Cauchy generating density, which for N parameters in θ may be expressed as,

$$d_\Theta(\theta) = \frac{c}{[|\theta|^2 + c^2]^{\frac{(N+1)}{2}}}, \tag{1}$$

where, the parameter c is the temperature parameter $T_c(t)$ which is found directly from the cooling schedule introduced next.

(iii) A gradual cooling schedule $T_c(t)$: It has been shown by [Szu, 1986] that the necessary and sufficient condition for convergence of the fast annealing algorithm to the global optimum requires the cooling schedule to be no faster than the following inverse time law :

$$\frac{T_c(t)}{T_0} = \frac{1}{1 + t}, \tag{2}$$

where, $T_0 \neq 0$ is a sufficiently high initial temperature.

(iv) A hill-climbing (cost-increasing) acceptance probability : This is given by the following Boltzmann distribution ,

$$p(\text{accepting higher cost}) \underline{\underline{\Delta}} p_a = \frac{1}{1 + e^{\frac{\Delta C}{k_B T_c(t)}}} \tag{3}$$

where, k_B is the Boltzmann constant and

$$\Delta C \triangleq C_i(\theta) - C_{i-1}(\theta). \tag{4}$$

From these objectives, it is clear that the requirements $(ii) - (iv)$ are of generic nature and do not change for different problems, whereas for (i), the cost function will have to be decided specifically for each optimization problem under consideration. Next, we briefly motivate a possible new annealing scheme which seems to provide faster convergence than existing techniques.

Motivation for the Proposed Annealing Scheme

The main motivation of the development of the simulated annealing algorithms has been to achieve convergence at the global optimum of a cost criterion which is nonlinearly related to the underlying unknown parameters. The existing algorithms by Metropolis *et al* and Szu do achieve this goal but even the faster scheme, due to Szu, may have very slow convergence for many applications. This drawback of FSA may be due to the rigid cooling schedule (2) used by the annealing scheme. Experimentations with the FSA indicate that the convergence rate depends on the choice of the initial temperature T_0. For example, if the parameter values in θ at the optimum point happen to be relatively small compared to T_0, the cooling may not have any major effect until the temperature becomes low and comparable to the parameter values. We feel that it may be helpful to make the cooling schedule depend adaptively to the parameter values. Also, the cooling schedule of FSA does not provide any possibility for reheating which, we think, may be advantageous in the search for the optimum. Especially, *reheating* may be useful when the temperature has become very low but the algorithm has not reached close to the optimum point. We also feel that constant temperature search may also facilitate the annealing procedure. Our preliminary experiments also indicate a potential problem with FSA in that, if the annealing iterations happen to reach the globally optimum point at a relatively high temperature, there is a large probability that it may leave the optimum point. This is due to the built in cost-increasing acceptance probability inherent in the FSA scheme. This seems to be a necessary evil because it is the cost-increasing acceptance probability which really provides FSA the mechanism to escape from local optimum points. But the problem is that the algorithm has no way of knowing if it is escaping a local or the global optimum point. Theoretically, of course, the algorithm will eventually come back to the globally optimum point but that may be after a very large number of iterations. We feel, one way to avoid this may be to stay fixed at the point of the lowest available cost until another point with lower cost is found. The mechanism to escape from a local optimum point can be incorporated in the algorithm by constant and high temperature search and also by expanding the search space according to the acceptance probability.

Keeping in mind the drawbacks of the existing simulated annealing schemes and the possible solutions as outlined above, we propose here a new simulated annealing scheme so as to reduce the convergence rate. The temperature parameter $T_c(k)$ plays a pivotal role in our annealing scheme. At every iteration the Cauchy parameter generating density in (1) uses $T_c(k)$ as the parameter c to generate the random deviations in parameter values. But c is the semi-interquartile of the Cauchy density function which means that there is a 50% percent probability that the random deviations generated by the Cauchy density will fall within $\pm c$. It seems logical to expect that, instead of rigidly pre-specifying the cooling schedule, one may be able to hasten the search process by adaptively selecting $T_c(k)$ according to the parameter values. Also, the cost function is optimized *w.r.t.* several parameters and at the point of lowest energy, these parameters may have values with different orders of magnitude. In such cases, it may be advantageous to make the semi-interquartile (*i.e*, the temperature $T_c(k)$) for each parameter dependent on the change in individual parameters. With these observations in mind we propose to study a possible faster adaptive simulated annealing scheme outlined in the next Subsection.

A Faster Simulated Annealing Scheme Based on Adaptive Cooling Schedule

The major steps of the proposed annealing algorithm are summarized below :

1 : Initialize the parameter set θ with arbitrary values and compute the cost function $C(\theta)$. Initialize the temperatures, *i.e*, the semi-interquartile parameters T_0^i for each parameter θ^i at sufficiently large values. Initialize iteration no. $k = 0$.

2 : Set $k = k + 1$. Stochastically generate a new set of parameters following the Cauchy state-transition probability defined in (1) but with different semi-interquartile values. Initialize the parameter number $i = 0$.

3 : Set $i = i + 1$. Compute the cost function $C^i(\theta)$ using the new value of the parameter θ^i and the values of the other parameters set at the previous iteration step. Then calculate the difference in cost

$$\Delta C^i \triangleq C_k^i(\theta) - C_k^{i-1}(\theta) \tag{5}$$

Note that if $i = 1$, the last term will be $C_{k-1}^p(\theta)$.

4(a) : If ΔC^i is negative, *i.e*, if the new cost is lower than the one in the previous iteration step, accept the new parameter set and the new lower cost and set the temperature at $T_c^i(k) = \Delta\theta^i \triangleq |\theta^i(k) - \theta^i(k-1)|$. If $i = p$ go to step 2, else go to step 3.

4(b) : If ΔC^i is positive, the cost is fixed at the previous (lower) value, *i.e*, $C_k^{i-1}(\theta)$ but accept a new temperature (semi-interquartile) $T_c^i(k) = m\Delta\theta^i$ with a probability determined by the Boltzmann distribution given by (3) using the temperature $T_c^i(k)$ and the cost change in (5). $m \geq 2$ is a constant that expands the search space. If $i = p$ go to step 2 else, go to step 3. If all the ΔC^i's remain same for a "long" time then STOP.

The new adaptive simulated algorithm as outlined above avoids many of the problems associated with the existing simulated annealing schemes. For example, the temperature parameter is made dependent on the difference in the parameter values and hence they are of the order of the parameters themselves. The minimum energy level parameters are kept unchanged so that there will not be any possibility of escape from the optimum minimum. Also an escape mechanism from local minimum is provided by expanding the search space according to the Boltzmann's probability when the cost is higher than that in the previous step.

To demonstrate the effectiveness of the proposed algorithm, we applied it to the problem of locating the optimum point of the cost function $C(\theta) = \theta^4 - 16\theta^2 + 5\theta$ [Szu, 1986]. The results of the simulations were quite encouraging and are discussed later. The results clearly indicate that the proposed algorithm converges faster to the optimum point than the existing technique due to Szu. Next we apply the algorithm to one of the well known optimization problems in signal processing, namely, frequency or angles of arrival estimation. Earlier, Sharman [1988] used Szu's technique to the frequency estimation problem. Here we compare the results of the two algorithms with a number of simulation studies. In the next Subsection the frequency estimation problem and the appropriate cost criterion are formulated.

The Frequency Estimation Problem

Estimation of frequencies from data composed of multiple narrowband signals in noise is one of the oldest problems studied in several branches of science. To define the problem, let $x(n)$, $n = 0, 1, \ldots, N-1$, be a data record of N consecutive samples. The multiple complex exponential signal model is defined as

$$\tilde{x}(n) \triangleq \sum_{k=1}^{p} a_k e^{j(\omega_k n + \phi_k)} \quad n = 0, 1, \ldots, N-1 \tag{6}$$

where

a_k : unknown amplitude of the k^{th} sinusoid,
ϕ_k : unknown phase of the k^{th} sinusoid,
ω_k : unknown angular frequency of the k^{th} sinusoid and
p : assumed number of sinusoids.

Note that for the angle of arrival estimation problem N denotes the number of sensors. The sensors are assumed to be separated by half the wavelengths of the incident waves and hence $\omega_k = \pi \sin\theta_k$, where θ_k denotes the angle of the kth source relative to the array normal.

The observation samples are expressed as

$$x(n) = \tilde{x}(n) + z(n) \tag{7}$$

where, $z(n)$ represents observation noise and/or modeling error. In vector form, the observed samples $x(n)$, the model samples $\tilde{x}(n)$ and the modeling error samples $z(n)$, for $n = 0, 1, \ldots, N-1$, are related as

$$\mathbf{x} \underline{\underline{\Delta}} \tilde{\mathbf{x}} + \mathbf{z} \tag{8}$$

where,

$$\mathbf{x} \underline{\underline{\Delta}} [x(0)\ x(1)\ \ldots,\ x(N-1)]^t \tag{9a}$$
$$\tilde{\mathbf{x}} \underline{\underline{\Delta}} [\tilde{x}(0)\ \tilde{x}(1)\ \ldots,\ \tilde{x}(N-1)]^t \tag{9b}$$
$$\mathbf{z} \underline{\underline{\Delta}} [z(0)\ z(1)\ \ldots,\ z(N-1)]^t. \tag{9c}$$

The multiple complex sinusoids model vector $\tilde{\mathbf{x}}$ is equivalently described by the following matrix-vector decomposition,

$$\tilde{\mathbf{x}} = \mathbf{Ta} \tag{10}$$

where

$$\mathbf{T} \underline{\underline{\Delta}} \begin{pmatrix} 1 & 1 & \ldots & 1 \\ e^{j\omega_1} & e^{j\omega_2} & \ldots & e^{j\omega_p} \\ \vdots & \vdots & \ddots & \vdots \\ e^{j\omega_1(N-1)} & e^{j\omega_2(N-1)} & \ldots & e^{j\omega_p(N-1)} \end{pmatrix} \text{ and } \mathbf{a} \underline{\underline{\Delta}} \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_p \end{pmatrix} \tag{11}$$

where, $A_k \underline{\underline{\Delta}} a_k e^{j\phi_k}$, for $k = 1, 2, \ldots, p$, respectively, are the complex amplitudes. The problem under consideration here is how to choose or estimate the best model parameters $A_1, A_2, \ldots, A_p$, and $\omega_1, \omega_2, \ldots, \omega_p$ such that the following modeling error norm

$$\|\mathbf{e}\|^2 \underline{\underline{\Delta}} E(\mathbf{T}, \mathbf{a}) \underline{\underline{\Delta}} \|\mathbf{x} - \mathbf{Ta}\|_2^2 \tag{12}$$

is minimized.

One may use this error criterion as the energy function and minimize it with respect to the unknown amplitudes, phases and the frequencies. In most cases, one is mainly interested in the unknown frequencies (or the angles of arrival). Also, the unknown vector a is linearly related to the error vector e and can be eliminated from the error criterion. It can be shown [Shaw, 1987, Kumaresan et al, 1986] that the frequencies can be directly obtained by minimizing the following alternate error criterion :

$$E(\mathbf{T}) \underline{\underline{\Delta}} \|(\mathbf{I} - \mathbf{P_T})\mathbf{x}\|_2^2, \tag{13}$$

where, $\mathbf{P_T}$ is the projection matrix and is defined as follows :

$$\mathbf{P_T} \triangleq \mathbf{T}(\mathbf{T}^H\mathbf{T})^{-1}\mathbf{T}^H \tag{14}$$

where 'H' denotes complex conjugated matrix transpose operation. The optimization problem described by equation (13) is a highly non-linear multi-dimensional optimization problem. Regarding the presently available methods, the simpler of the existing algorithms are usually suboptimal and the optimal techniques, being dependent on the initial conditions, are not guaranteed to attain the global optimum and are usually computationally expensive. We strongly feel that simulated annealing offers an exciting alternative to solve this optimization problem. Extensive simulation studies have been done with both the annealing schemes outlined above using the following cost criterion :

$$C(\theta) \triangleq \|(\mathbf{I} - \mathbf{P_T})\mathbf{x}\|_2^2 \tag{15}$$

where $\theta \triangleq [\omega_1\ \omega_2\ \ldots\ \omega_p]^t$. The simulation results are summarized in the next Subsection.

Simulation Results

The proposed simulated annealing algorithm has been evaluated by extensive simulations on a variety of data sets and we report the results of the three experiments here. In the last two experiments the cost function of the form given in (15) is minimized and gaussianly distributed white noise are used.

Experiment 1:

In this experiment we considered a simple example [Szu, 1986] to demonstrate how the proposed technique outperforms the existing annealing methods. In this example the cost function $C(x) = x^4 - 16x^2 + 5x$ has one local minima and a global minima. Several trials of annealing runs were made by both the methods and the results are tabulated in Table I. In both the methods the iterations were continued till the global minima is reached. In the proposed technique, once the global minima is reached it continues to stay there for ever whereas in Sharman's [4] method it would remain in the global minimum only when the temperature becomes sufficiently low.

Experiment 2:

This experiment involves estimating the frequencies of two narrowband signals in noise. The data record consists of 25 consecutive samples generated from two sources of closely spaced frequencies of 0.5 and 0.52 Hertz. The SNR's of both the sources were 10 dB's. The annealing technique is applied to a single set of data and the results are compared with that of Sharman [4].

Figures 1a and 1b show a typical annealing run by the proposed method and that of Sharman's [4]. Both the runs were started with identical initial conditions. The temperature was initialized to 1000. The graphs show the trajectories of the source frequencies as the annealing proceeds. The values of the cost function after 3000 iterations were 2.305 in the proposed method and 2.445 in Sharman's [4] method.

Experiment 3:

In this experiment there are three point sources illuminating an array of 8 sensors at directions of 10, 25 and 35 degrees from the array normal. The signal strengths over the additive white noise were 10, 15 and 12 dB's. This experiment was taken from Sharman[4]. The temperature was initialized to 1000. Figures 2a and 2b show the 3000 annealing iterations by the proposed technique and Sharman's [4] method, respectively, with identical initial conditions. In all our trials of annealing iteration the global minima was reached within few hundreds of iteration in the proposed method whereas in Sharman's [4] method the estimates were varying erratically even after few thousands of iterations.

The key difference between the proposed method and the existing method is that the cost function

monotonously decreases as the iteration proceeds whereas in Sharman's [4] method the cost function varies erratically as long as the temperature is high.

Table I

θ	initial temperature	Number of iterations		
		Sharman's method	Proposed method	
			m=1	m=5
10	100	4684	78	102
-50	1000	69758	94	2286
-5000	500	27283	1591	1420
5	50	1893	23	23
2.74	200	11714	444	427
-350	50	1936	320	292
-65	2000	96858	120	120
55	2350	99438	103	103
50	10000	99998	378	371
-200	5555	99946	102	92

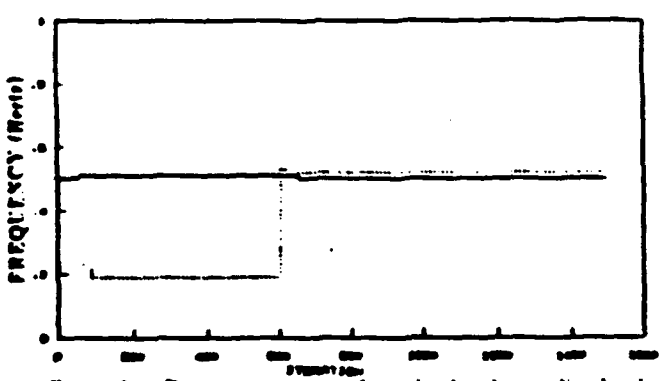

Figure 1a . Frequency estimates from simulated annealing by the proposed technique 2 point sources of frequencies 0.5 and 0.52 Hertz SNR = 10, initial temperature = 1000 The graph shows 1500 iterations

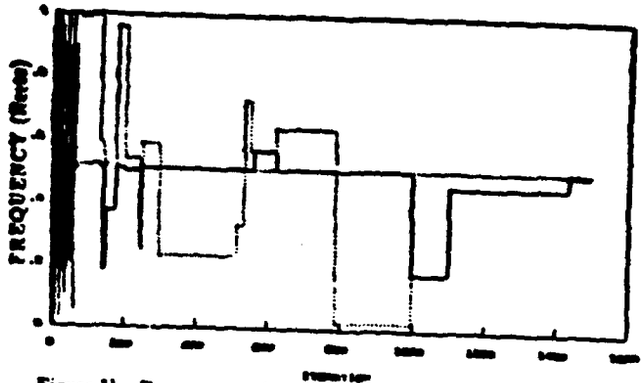

Figure 1b : Frequency estimates from simulated annealing by Sharman's [4] method. 2 point sources of frequencies 0.5 and 0.52 Hertz. SNR = 10, initial temperature = 1000 The graph shows 1500 iterations

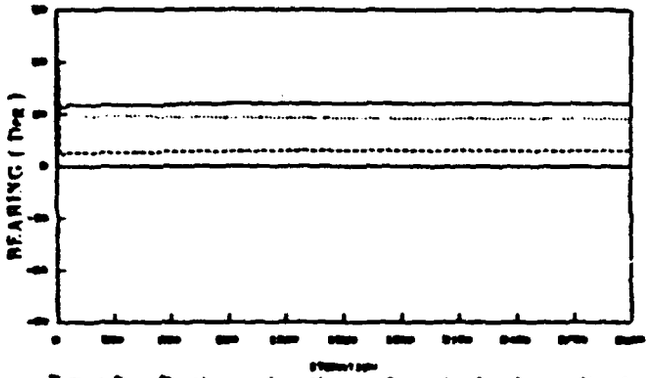

Figure 2a : Bearing angle estimates from simulated annealing by the proposed technique 3 uncorrelated point sources at bearings of 10 25 and 35 degrees. SNR's of 10, 15 and 12 dB's, initial temperature = 1000 The graph shows 3000 iterations.

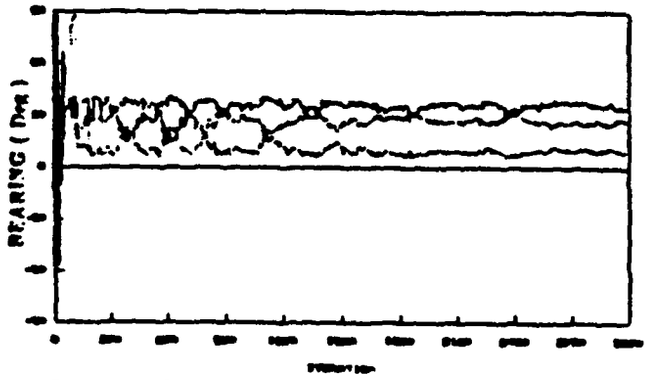

Figure 2b : Bearing angle estimates from simulated annealing by Sharman's [4] method 3 uncorrelated point sources at bearings of 10 25 and 35 degrees. SNR's of 10, 15 and 12 dB's, initial temperature = 1000 The graph shows 3000 iterations

References

[1] S. Kirkpatrick *et al*, "Optimization by Simulated Annealing", *Science*, Vol. 220, pp. 671-680, 1983.

[2] S.B. Gelfand and S.K. Mitter, "Analysis of Simulated Annealing for Optimization", *in Proceedings of the 24th IEEE Conference on Decision & Control*, pp. 779 - 786, Dec., 1985.

[3] A. El Gamal *et al*, "Using Simulated Annealing to Design Good Codes", *IEEE Transactions on Information Theory*, Vol. IT-33, pp. 116-123, 1987.

[4] K.C. Sharman, "Maximum Likelihood Parameter Estimation by Simulated Annealing", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, NY, pp. 2741-2744, April, 1988.

[5] H. Szu, "Fast Simulated Annealing", *in Proceedings of AIP Conference 151, Neural Networks for Computing*, J. S. Denker (Editor), Snowbird, Utah, 1986.

[6] M. Metropolis *et al*, "Equations of State Calculations by Fast Computing Machines", *The Journal of Chemical Physics*, Vol. 21, pp. 1087-1092, 1953.

[7] A. K. Shaw, "Some Structured Matrix Approximation Problems in Signal Processing," Ph.D. Dissertation, University of Rhode Island, 1987.

[8] R. Kumaresan, L. L. Scharf and A. K. Shaw, "An Algorithm for Pole-Zero Modeling and Spectral Analysis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol.-ASSP-34, no. 3, pp. 637-640, June, 1986.

SECTION 3.3 : ONE-STEP ESTIMATION OF ANGLES OF ARRIVALS OF WIDEBAND SIGNALS

SUMMARY

A high resolution algorithm for estimating the angles of arrivals of multiple wideband sources is studied for this part of the work. The algorithm is effective for a dense and equally spaced array structure where a bilinear transformation is utilized in the frequency domain for combining the signal subspaces at different frequencies for coherent processing. When compared with existing coherent approaches, the algorithm is non-iterative in the sense that *all* the arrival angles can be estimated in only one step of the algorithm. Existing algorithms can only estimate the angles of a cluster of sources in a particular direction. The proposed algorithm, unlike the existing ones, does not need the knowledge of the initial estimates of the arrival angles. The work reported here is a variation of some earlier work by the author [32]. Instead of using generalized eigendecomposition or matrix-pencil method, here we pre- or post-multiply the signal-subspace matrix with the noise matrix. This enables us to use regular eigendecomposition routine to estimate the source angles. It is also shown that it may be numerically more stable if the coherent combination is not focused in the center frequency the numerical value of which could be very large. The new focusing matrix given here allows to focus independent of the center frequency. The performance of the algorithm is presented using simulated data.

I. Introduction : Estimation of locations and characteristics of radiating sources using data collected at the output of an array of sensors is a frequently researched problem in signal processing [31]. This problem has many ramifications and different kinds of categorization are possible depending on the assumptions and the knowledge or otherwise of the source and noise characteristics, array geometry, etc. One broad categorization is based on the radiating sources having narrowband or broadband spectra. A plethora of publications are available in literature and in the next two paragraphs some of the approaches are briefly outlined.

Ia. Narrowband Sources : Estimation of frequencies of sinusoidal signals in noise and estimation of angles of arrivals of planewaves with narrowband signals are related problems. Various existing techniques are primarily based on Fourier methods [1,2], eigendecomposition and singular value decomposition based methods [3-10, among others], and maximum likelihood based methods [11-13]. Some of these techniques have also been extended to the 2-D problem of simultaneous estimation of frequencies and wavenumbers [14-17].

Ib. Broadband Sources : The broadband problem is radically different from the narrowband problem in many respects. As for example, unlike the narrowband case, a delay in time cannot be as simply accounted for in the phase of the exponentials in the time domain representation. This is because a broadband signal is composed of a continuum of frequencies where each constituent frequency goes through a phase change different from any other. Within any nonzero bandwidth of the signal there may be an infinity of frequencies all of which experience changes in their respective phases. Hence the broadband signal cannot be modeled simply by a sum of a finite number of exponentials as is the case for narrowband signals. For these same reasons the data matrix, correlation matrix and the Hankel matrix formed with the time domain broadband data do not possess the nice properties which the narrowband case enjoys even when absolutely no background noise is present. In fact, all these matrices, if formed with noiseless time domain broadband data, will be of full rank, no matter how large the dimensions of the matrices are.

It should be clear from the discussions in the last paragraph that it will be futile to attempt to solve the angles of arrival problem for the broadband case by directly implementing the high resolution methods in [1-13] on the time domain broadband data. Coker and Ferrara [1982] have provided a fairly thorough discussion

on the problems of applying the high resolution methods developed for narrowband problems directly to the broadband problem. Wax et al [1984] have formulated both the narrowband and the broadband problems and have emphasized the inconvenience of processing broadband data in the time domain.

The problems depicted above change substantially in the spectral domain representation of wideband array data [Wax et al 1984]. It will be shown in Subsection II that the spectral density matrix at a frequency bin within the nonzero bandwidth of the wideband signal possesses many interesting properties which are quite similar to the narrowband case. Namely, the rank of the spectral density matrix is exactly equal to the number of sources when no noise is present. Also, the column and the row spaces of the spectral density matrices have familiar Vandermonde type structures which can be compared to the structure of the correlation matrix for narrowband case. Hence the structured matrix approximation approach [17] could be applied to the spectral density matrix at a frequency bin to estimate the arrival angles. Techniques based on processing the spectral density matrix of a single bin have been proposed in literature [Owsley and Swope, 1981]. This approach does not utilize all the information available in the spectral density matrices at different frequency bins within the nonzero bandwidth of the signal and is more suited in the case of multipath propagation of a narrowband source. Regarding the spectral density matrices at different bins, although the true ranks of all the spectral density matrices are same and equal to the number of distinct sources, the column/row spaces vary for different frequency bins. It is rather expensive to process at separate bins separately and then combine the results to obtain the angle estimates.

In this part of our research, a spectral domain based method is presented which coherently combines all the spectral density matrices and the angle estimates are obtained from the coherently combined spectral density matrix. Compared to an existing coherent method [27, 28], the present approach is non-iterative in the sense that all the incident angles are estimated at a single step. Also, unlike the existing coherent method, the initial estimates of the arrival angles are not required while coherently combining the spectral density matrices. The method is based on generalized eigen-decomposition and utilizes dense array approximation and a bilinear transformation for coherently combining the spectral density matrices of all the frequency bins. Some preliminary work on the coherent method described here was outlined in [32]. Here we show that by pre- or post-multiply the signal-subspace matrix with the combined noise matrix we can essentially avoid the computation of the generalized eigendecomposition or matrix-pencil. This enables us to use regular eigendecomposition routine to estimate the source angles. It is also shown that it may be numerically more stable if the coherent combination is not focused in the center frequency the numerical value of which could be very large. The new focusing matrix given here allows to focus independent of the center frequency.

Some other techniques for localization and source spectra estimation of wideband sources have been reported by several researchers in the last two decades. See the tutorial papers in the Special Issue on Time Delay Estimation [Carter, 1981]. The conventional approach for the case of a single source is to use a form of generalized correlator [Knapp and Carter, 1976] to estimate the Time-Difference-Of-Arrival (TDOA) of the signal at the sensors. Maximum likelihood based methods [Bangs and Schultheiss, 1973; Hahn and Tretter, 1973; Wax and Kailath, 1983] for single and multiple sources require the knowledge of source and noise spectra and are computationally expensive. Parameter estimation based methods [Morf et al, 1979; Porat and Friedlander, 1983, Nehorai et al, 1983; Su and Morf, 1983] assume Auto-Regressive Moving Average (ARMA) models for the received signals and the estimated ARMA parameters are utilized for TDOA estimation. Computational complexity of these methods is high and the performance of these approaches depend on the correctness of the assumed model for the unknown wideband signals.

Extending existing ideas in the narrowband problem, Wax et al [1984] proposed an eigen-decomposition based approach for wideband source localization. In their approach, the eigenvectors of the estimated spectral density matrix at each narrowband bin of the signal bandwidth were incoherently combined to estimate the TDOAs.

Recently, Wang and Kaveh [1984, 1985] presented a coherent signal subspace based approach which avoids the rather expensive eigen-decomposition of spectral density matrices at each frequency bin. In their approach, initial estimates of the angles of arrival are used to transform the signal eigenspaces at different frequency bins to generate a single coherent signal subspace and then a generalized eigen-decomposition is used for obtaining more accurate estimates. The algorithm iteratively estimates well separated angles by focusing at different angles at each time.

In this work a simple bilinear transformation matrix and the approximation resulting from dense and equally spaced array structure assumption are utilized to combine the individual narrowband spectral density matrices for coherent processing. In a related problem, Henderson [1985] used a bilinear transformation and dense array approximation for rank reduction of Hankel/Block-Hankel type data matrices. Henderson considered the angle estimation problem of multiple sources when each source is emitting multiple narrowband spectra. His formulation and approach is completely in the time domain and is based on Singular Value Decomposition of Prony type Data/Hankel matrices. Application of his time domain method to the broadband data may encounter the problems outlined in the introduction of this Section and also in [Coker and Ferrara, 1982]. Also, no proof was given to ensure invertibility of the bilinear transformation used and the effect of the transformation to the noise subspace is unclear since no assumption was made on the noise characteristic. It is also not obvious if the time domain approach is applicable for correlated sources.

The coherent method described in this Section is based on spectral domain representation. The method is non-iterative and does not require preprocessing for obtaining initial estimates of the angles of arrival and all the angles are estimated from a single step of coherent subspace computation. The performance of the proposed method is characterized by several simulation experiments.

This Section is arranged as follows. In Subsection II, the coherent problem is formulated and a new coherent algorithm is given. In Subsection III some observations on our use of Structured matrix approximation approach is given. Finally in Subsection IV, tehe results of our simulations are shown.

II. Problem Formulation: The observed signal is assumed to be composed of p plane waves with an overlapping bandwidth of B Hz. They are sampled simultaneously at the output of a linear array of M ($> p$) equally spaced sensors. The signal received at the ith sensor is expressed as

$$r_i(t) = \sum_{k=1}^{d} s_k(t - (i-1)\frac{D}{c}\sin v_k) + n_i(t) \qquad (1)$$
$$-\frac{T}{2} \le t \le \frac{T}{2} \quad 1 \le i \le M$$

where $s_k(\cdot)$ is the signal radiated by the kth source, D is the separation between the sensors, c is the propagation velocity of the signal wavefront, v_k is the angle that the kth wavefront makes with the line of array and $n_i(\cdot)$ is the additive noise at the ith sensor.

Representing both sides by respective Fourier coefficients,

$$R_i(\omega_l) = \sum_{k=1}^{d} e^{-j\omega_l(i-1)\frac{D}{c}\sin v_k} S_k(\omega_l) + N_i(\omega_l) \qquad (2)$$

with $\omega_l = \frac{2\pi}{T}l, l = l_1, \ldots, l_1 + n_f$, where ω_{l_1} and $\omega_{l_1+n_f}$ are the lowest and highest frequencies in B. In matrix notation,

$$\mathbf{R}(\omega_l) = \mathbf{A}(\omega_l)\mathbf{S}(\omega_l) + \mathbf{N}(\omega_l) \qquad (3)$$

where

$$\mathbf{R}(\omega_l) = [R_1(\omega_l) \dots R_M(\omega_l)]^t \tag{4a}$$

$$\mathbf{N}(\omega_l) = [N_1(\omega_l) \dots N_M(\omega_l)]^t \tag{4b}$$

$$\mathbf{S}(\omega_l) = [S_1(\omega_l) \dots S_p(\omega_l)]^t \tag{4c}$$

and $\mathbf{A}(\omega_l)$ is an $M \times p$ direction-frequency matrix

$$\mathbf{A}(\omega_l) = \begin{pmatrix} 1 & \dots & 1 \\ e^{-j\omega_l\tau_1} & \dots & e^{-j\omega_l\tau_p} \\ \vdots & \ddots & \vdots \\ e^{-j\omega_l(M-1)\tau_1} & \dots & e^{-j\omega_l(M-1)\tau_p} \end{pmatrix} \tag{4d}$$

with $\tau_i = \frac{D}{c}\sin\ v_i$ being the TDOA of the ith source. The covariance matrix of the Fourier coefficient vector $\mathbf{R}(\omega_l)$ will approach the spectral density matrix if the observation time is large enough compared to the correlation time of the processes [30]. With this assumption,

$$\mathbf{K}(\omega_l) = \mathbf{A}(\omega_l)\mathbf{P}_s(\omega_l)\mathbf{A}^H(\omega_l) + \sigma_n^2\mathbf{P}_n(\omega_l) \tag{5}$$

where, $\mathbf{K}(\omega_l)$, $\mathbf{P}_s(\omega_l)$ and $\mathbf{P}_n(\omega_l)$ are the spectral density matrices of the processes $r_i(\cdot)$, $s_k(\cdot)$ and $n_i(\cdot)$, respectively. $\mathbf{A}^H(\omega_l)$ stands for the transpose conjugate of $\mathbf{A}(\omega_l)$. The noise process is assumed to be independent of the sources and the noise spectral density matrix is assumed to be known except for a multiplicative constant σ_n^2. With the above model at hand, the problem is to estimate the τ_i's from the estimated covariance matrices $\hat{\mathbf{K}}(\omega_l)$ of the received signal plus noise. Estimates of the angles of arrivals v_i's can then be computed using the relationship in (4e).

IIa. Previous Methods : The two major approaches which exploit the properties of the eigenspaces of $\mathbf{K}(\omega_l)$ to estimate the arrival angles (or TDOAs) are briefly described below.

In [26], eigendecompositions of $\hat{\mathbf{K}}(\omega_l)$s are performed in all the frequency bins in B and globally orthogonal direction vectors are obtained by computing and plotting any of the following two measures,

$$J_1(v) = \frac{1}{\frac{1}{n_f+1}\sum_{l=l_1}^{l_1+n_f}\frac{1}{M-p}\sum_{k=p+1}^{M}\|\mathbf{a}_v^H(\omega_l)\hat{\mathbf{v}}_k(\omega_l)\|^2} \tag{6}$$

or

$$J_2(v) = \frac{1}{\prod_{l=l_1}^{l_1+n_f}\left(\frac{1}{M-p}\sum_{k=p+1}^{M}\|\mathbf{a}_v^H(\omega_l)\hat{\mathbf{v}}_k(\omega_l)\|^2\right)^{\frac{1}{n_f+1}}} \tag{7}$$

where $\mathbf{a}_v(\omega_l)$'s are direction-frequency vectors defined as, $\mathbf{a}_v(\omega_l) \triangleq [1\ e^{-j\omega_l\frac{D}{c}\sin v}\ \dots\ e^{-j\omega_l\frac{D}{c}(M-1)\sin v}]^t$ and $\hat{\mathbf{v}}_k(\omega_l)$'s are the eigenvectors corresponding to the smaller $M-p$ eigenvalues of $\hat{\mathbf{K}}(\omega_l)$. Note that the distance measures in (6) and (7) require eigendecomposition of $\hat{\mathbf{K}}(\omega_l)$ for all $l = l_1,\ l_1+1,\ \dots,\ l_1+n_f$. This obviously is a computationally expensive procedure. Instead in [27] and [28], a transformation matrix was employed to reduce this burden. Using initial estimates of the possible angles of arrivals, transformation matrices $\mathbf{T}_{\hat{v}}(\omega_l)$ were formed such that direction-frequency matrices of all the frequency bins are transformed to the center frequency ω_c in B, i.e,

$$\mathbf{T}_{\hat{v}}(\omega_l)\mathbf{A}(\omega_l) = \mathbf{A}(\omega_c), \quad l = l_1, l_1+1, \dots, l_1+n_f \tag{8}$$

Then using the transforming matrices $\mathbf{T}_{\hat{v}}(\omega_l)$, $l = l_1, l_1+1, \ldots, l_1+n_f$, all the spectral density estimates are combined in the following manner,

$$\begin{aligned}\mathbf{G} &\triangleq \sum_{l=l_1}^{l_1+n_f} \mathbf{T}_{\hat{v}}(\omega_l)\hat{\mathbf{K}}(\omega_l)\mathbf{T}_{\hat{v}}^H(\omega_l) \\ &= \mathbf{A}(\omega_c)\mathbf{G}_s\mathbf{A}^H(\omega_c) + \sigma_n^2\mathbf{G}_n \end{aligned} \tag{9}$$

where

$$\mathbf{G}_s = \sum_{j=l}^{l+n_f} \mathbf{P}(\omega_l), \quad \mathbf{G}_n = \sum_{l=l_1}^{l_1+n_f} \mathbf{T}_{\hat{v}}(\omega_l)\mathbf{P}_n(\omega_l)\mathbf{T}_{\hat{v}}^H(\omega_l) \tag{10}$$

Next, a coherent signal subspace theorem [28] for the matrix pencil $(\mathbf{G}, \mathbf{G}_n)$ is utilized to estimate the angles in the direction/s of $\hat{\mathbf{v}}$ by computing the maxima of the following measure,

$$J(v) = \frac{1}{\sum_{k=p+1}^{M} \|\mathbf{a}_v^H(\omega_c)\hat{\mathbf{e}}_k(\omega_c)\|^2} \tag{11}$$

where $\hat{\mathbf{e}}_{p+1}(\omega_c), \hat{\mathbf{e}}_{p+2}(\omega_c), \ldots, \hat{\mathbf{e}}_M(\omega_c)$ are the generalized eigenvectors of the matrix pencil $(\mathbf{G}, \mathbf{G}_n)$ corresponding to the smallest $M-p$ eigenvalues. If the arrival angles are well separated, ***different*** transformation matrices needs to be utilized to 'focus' to particular directions ***using initial estimates of the angles*** [27]. The coherent transformation described here ***does not require*** any initial estimates of the arrival angles.

IIb. Proposed Method : The coherent signal subspace processing technique described above requires the knowledge of the initial estimates of the angles of arrivals to form the transformation matrix $\mathbf{T}_{\hat{v}}(\omega_l)$. To avoid this requirement, a different approach that utilizes a bilinear transformation and dense array approximation to form the transformation matrices is presented here.

Synthesizing a Bilinear Transformation Matrix : Let $\mathbf{B}$ be an $M \times M$ matrix constructed from the coefficients of the *M-1 th* order z-polynomials $p_k(z) = (1+z)^{M-k}(1-z)^{k-1}, k = 1, 2, \ldots, M$. The elements of the kth row of the matrix $\mathbf{B}$ are the coefficients of $p_k(z)$ taken in ascending order of z. The coefficients of $p_k(z)$ can be found by convolving the coefficients of the polynomials $(1+z)^{M-k}$ and $(1-z)^{k-1}$. For example, a 4×4 $\mathbf{B}$ matrix will have the form,

$$\mathbf{B}_{M=4} = \begin{pmatrix} 1 & 3 & 3 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -3 & 3 & -1 \end{pmatrix} \tag{12}$$

Proposition : The $M \times M$ matrix $\mathbf{B}$, defined above, is nonsingular and hence, its rows are linearly independent.

Proof : Let $\mathbf{Z}$ be a real $M \times M$ Vandermonde matrix of the form

$$\mathbf{Z} \triangleq \begin{pmatrix} 1 & 1 & \ldots & 1 \\ z_1 & z_2 & \ldots & z_M \\ z_1^2 & z_2^2 & \ldots & z_M^2 \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{M-1} & z_2^{M-1} & \ldots & z_M^{M-1} \end{pmatrix} \tag{13}$$

where $z_i \neq z_j \neq -1 \ \forall i,j$. So $\mathbf{Z}$ is nonsingular and its non-zero determinant is given by $det\ \mathbf{Z} = \prod_{M \geq i > j \geq 1}(z_i - z_j)$. Since the kth row of $\mathbf{B}$ are the coefficients of *M-1 th* order z-polynomials $p_k(z) = (1+z)^{M-k}(1-z)^{k-1}, k = 1, 2, \ldots, M$ and $\mathbf{Z}$ has Vandermonde structure, it can be easily shown that,

$$\mathbf{BZ} = \begin{pmatrix} (1+z_1)^{M-1} & \cdots & (1+z_M)^{M-1} \\ (1+z_1)^{M-2}(1-z_1) & \cdots & (1+z_M)^{M-2}(1-z_M) \\ \vdots & \ddots & \vdots \\ (1-z_1)^{M-1} & \cdots & (1-z_M)^{M-1} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \frac{1-z_1}{1+z_1} & \frac{1-z_2}{1+z_2} & \cdots & \frac{1-z_M}{1+z_M} \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{1-z_1}{1+z_1}\right)^{M-1} & \left(\frac{1-z_2}{1+z_2}\right)^{M-1} & \cdots & \left(\frac{1-z_M}{1+z_M}\right)^{M-1} \end{pmatrix} \begin{pmatrix} (1+z_1)^{M-1} & & \\ & \ddots & \\ & & (1+z_M)^{M-1} \end{pmatrix} \tag{14}$$

so that,

$$\begin{aligned} det\ [\mathbf{BZ}] &= \prod_{M \geq i > j \geq 1} \frac{2(z_i - z_j)}{(1+z_i)(1+z_j)} \prod_{i=1}^{M} (1+z_i)^{(M-1)} \\ &= \prod_{M \geq i > j \geq 1} 2(z_i - z_j) \\ &= det\mathbf{B}\ det\mathbf{Z} \end{aligned} \tag{15}$$

so $det\ \mathbf{B} = \prod_{M \geq i > j \geq 1} 2 = 2^{\frac{M(M-1)}{2}} \neq 0$. Hence, $\mathbf{B}$ is nonsingular and its rows are linearly independent. $\Diamond$

A similar bilinear transformation matrix was given in [29] though from the manner in which it was constructed, it was rather difficult to appreciate its properties.

Now, since $\mathbf{A}(\omega_l)$ in (4*a*) has a Vandermonde structure, following the same steps as in (15) it is easy to see that,

$$\mathbf{BA}(\omega_l) = \begin{pmatrix} 1 & \cdots & 1 \\ \frac{1-e^{-j\omega_l \tau_1}}{1+e^{-j\omega_l \tau_1}} & \cdots & \frac{1-e^{-j\omega_l \tau_p}}{1+e^{-j\omega_l \tau_p}} \\ \vdots & \ddots & \vdots \\ \left(\frac{1-e^{-j\omega_l \tau_1}}{1+e^{-j\omega_l \tau_1}}\right)^{M-1} & \cdots & \left(\frac{1-e^{-j\omega_l \tau_p}}{1+e^{-j\omega_l \tau_p}}\right)^{M-1} \end{pmatrix} \begin{pmatrix} \left(1+e^{-j\omega_l \tau_1}\right)^{M-1} & & \\ & \ddots & \\ & & \left(1+e^{-j\omega_l \tau_p}\right)^{M-1} \end{pmatrix} \tag{16}$$

$$= \begin{pmatrix} 1 & \cdots & 1 \\ j\tan(\frac{\omega_l \tau_1}{2}) & \cdots & j\tan(\frac{\omega_l \tau_p}{2}) \\ \vdots & \ddots & \vdots \\ \left(j\tan(\frac{\omega_l \tau_1}{2})\right)^{M-1} & \cdots & \left(j\tan(\frac{\omega_l \tau_p}{2})\right)^{M-1} \end{pmatrix} \mathbf{E}(\omega_l) \tag{17}$$

where, $\mathbf{E}(\omega_l)$ denotes the diagonal matrix in (16). If the sensor to sensor separation D is small compared to all wavelengths in B, then $\tan(\frac{\omega_l \tau_i}{2}) \simeq \frac{\omega_l \tau_i}{2}, \forall i,j$. Using this approximation and premultiplying $\mathbf{BA}(\omega_l)$ by an $M \times M$

diagonal matrix $\mathbf{D}(\frac{1}{\omega_l})$ whose $(m,m)th$ term is given by $\left(\frac{2}{j\omega_l}\right)^{m-1}$ we obtain,

$$\mathbf{D}(\frac{1}{\omega_l})\mathbf{B}\mathbf{A}(\omega_l) \simeq \begin{pmatrix} 1 & \dots & 1 \\ \tau_1 & \dots & \tau_p \\ \vdots & \ddots & \vdots \\ \tau_1^{M-1} & \dots & \tau_p^{M-1} \end{pmatrix} \mathbf{E}(\omega_l)$$

$$\underline{\triangle}\ \mathbf{A}'\mathbf{E}(\omega_l) \tag{18}$$

The matrix $\mathbf{A}'$ whose columns are the transformed direction-frequency vectors has no dependence on the frequencies at all! It may be pointed out here that in [32], we had proposed using a center-frequency dependent coherent transformation matrix, $\mathbf{D}(\frac{\omega_c}{\omega_l})$ whose $(m,m)th$ term was given by $\left(\frac{2\omega_c}{j\omega_l}\right)^{m-1}$, where $\omega_c = 2\pi f_c$, and f_c is the midband frequency in B. Also note that $\mathbf{A}'$ has a Vandermonde structure and its columns are linearly independent as long as $\tau_i \neq \tau_j$ for $i \neq j$.

A Center-Frequency-Independent Transformation Matrix and Coherent Processing : The new transformation matrix which we define as $\mathbf{T}'(\omega_l)\ \underline{\triangle}\ \mathbf{D}(\frac{1}{\omega_l})\mathbf{B}$, does not depend on the arrival angles or the center frequency and since $\mathbf{B}$ is nonsingular and $\mathbf{D}(\frac{1}{\omega_l})$ is diagonal with nonzero diagonal elements, $\mathbf{T}'(\omega_l)$ is also nonsingular. Using these transformation matrices $\mathbf{T}'(\omega_l)$, $l = l_1, l_1+1\dots, l_1+n_f$, all the spectral density estimates can now be combined in the following manner,

$$\begin{aligned}\mathbf{G}' &\underline{\triangle} \sum_{l=l_1}^{l_1+n_f} \mathbf{T}'(\omega_l)\hat{\mathbf{K}}(\omega_l)\mathbf{T}'^H(\omega_l) \\ &= \mathbf{A}'\mathbf{G}_s'\mathbf{A}'^H + \sigma_n^2\mathbf{G}_n'\end{aligned} \tag{19}$$

where

$$\mathbf{G}_s' = \sum_{l=l_1}^{l_1+n_f} \mathbf{E}(\omega_l)\mathbf{P}_s(\omega_l)\mathbf{E}^H(\omega_l) \tag{20}$$

and

$$\mathbf{G}_n' = \sum_{l=l_1}^{l_1+n_f} \mathbf{T}'^H(\omega_l)\mathbf{P}_n(\omega_l)\mathbf{T}'^H(\omega_l) \tag{21}$$

Next, the coherent signal subspace theorem for the matrix pencil $(\mathbf{G}', \mathbf{G}_n')$ is utilized to estimate all the angles of arrivals by computing the maxima of the following measure,

$$J'(v) = \frac{1}{\sum_{k=p+1}^{M} \|\mathbf{a}_v'^H \hat{\mathbf{e}}_k'\|^2} \tag{22}$$

where $\hat{\mathbf{e}}_{p+1}', \hat{\mathbf{e}}_{p+2}', \dots, \hat{\mathbf{e}}_M'$ are the generalized eigenvectors of the matrix pencil $(\mathbf{G}', \mathbf{G}_n')$ corresponding to the smallest $M-p$ eigenvalues and $\mathbf{a}_v'$'s are the new direction-frequency vectors defined as, $\mathbf{a}_v'\ \underline{\triangle}\ [1\ \ \frac{D}{c}\sin v\ \ \dots\ \ (\frac{D}{c}\sin v)^{M-1}]^t$.

From (19) it is also obvious that since $\mathbf{G}_n'$ is known to be positive definite [28], pre-multiplying $\mathbf{G}'$ by $\mathbf{G}_n'$, we get,

$$\mathbf{G}_{PRE}\underline{\triangle}\mathbf{G}_n'^{-1}\mathbf{A}'\mathbf{G}_s'\mathbf{A}'^H + \sigma_n^2\mathbf{I}. \tag{23}$$

Similarly, post-multiplying $\mathbf{G}'$ by $\mathbf{G}_n'$ and taking transpose conjugated, we get,

$$\mathbf{G}_{POST}\underline{\triangle}\mathbf{G}_n'^{-1}\mathbf{A}'\mathbf{G}_s'^H\mathbf{A}'^H + \sigma_n^2\mathbf{I}. \tag{24}$$

It can be shown that both $\mathbf{G}_{PRE}$ and $\mathbf{G}_{POST}$ possess exactly same set of eigenvectors and hence regular eigendecomposition routine on either of these matrices would yield similar maxima as in (22).

III. Discussion on Structured Approach : If we look at the matrix structures of $\mathbf{G}'$, ${\mathbf{G}'_{PRE}}^H$ or ${\mathbf{G}'_{POST}}^H$ in (19), (23) and (24), respectively, the signal part of each one of these matrices have a Vandermonde Matrix $\mathbf{A}'$ as a matrix factor appearing in the front. Such Vandermonde structures had been exploited successfully in [17] for estimation of frequencies and wavenumbers using structured matrix approximation approach. But there is a key difference. The Vandermonde matrix appearing in [17] had complex exponentials as their elements. Hence complex-conjugate-symmetry had to be imposed on the estimated polynomial coefficients so that the roots lie on the unit circle. But in the present case the elements of $\mathbf{A}'$ are all real. The constraints to enforce the roots to lie on the real axis are all non-linear in nature [33, 34]. Incorporation of these constraints in the optimization criteria appearing in [17] would lead to the use of general non-linear optimization criteria such as Fletcher-Powell or Gauss-Newton-Marquardt methods. Instead we had attempted to use the structured approach directly without the conjugate symmetry constraints on the polynomial coefficients. The initial results were encouraging and at high SNR the approach gave excellent results. But as the SNR was lowered, the roots, without any constraints, started to become complex in order to minimize the unconstrained criteria. These results were not as good as the results we obtained using eigendecomposition method as shown below.

IV. Simulation Results : The same simulation examples as presented in [27] and [28] were used to evaluate the performance of the proposed method. In all the simulations, a linear array of $M = 16$ equally spaced sensors was used. The spacing between two consecutive Φ sensors is $D = \frac{c}{rf_c}$, where $f_c = \frac{1}{2\pi} =$ the midband frequency in B and r is the ratio of the wavelength at the midband frequency f_c and the inter-element spacing D. The source signals are temporarily stationary bandpass white Gaussian processes with zero mean. The noise processes at each sensor are stationary, statistically independent, identical white Gaussian bandpass processes with zero mean and are independent of the source processes. The sources and the noise processes have the same bandwidth of $B = 100$. The same sampling specifications and data segmentation as described in [27] and [28] were used. The received signal plus noise processes were sampled at each sensor at 80Hz and then divided into 64 segments of 64 samples each and then each segment was transformed into frequency domain by unwindowed FFT to obtain $n_f + 1 = 33$ narrowband components. The covariance matrix estimate at the jth frequency was estimated as

$$\hat{\mathbf{K}}(\omega_l) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{X}_n(\omega_l)\mathbf{X}_n^H(\omega_l),\ l = l_1, l_1+1, \ldots, l_1+n_f \tag{25}$$

where, the vector $\mathbf{X}_n(\omega_l)$ is the lth component of the Fourier transformed data for the nth segment. The signal to noise ratio is defined as the ratio of the signal power of one source and the power of the the noise at the output of a single sensor.

Figure 1 shows the results for the case of two uncorrelated sources at $v_1 = 9.0°$ and $v_2 = 12.0°$ for SNR = 10 dB and r = 4. The Figure shows overlapped plots of 5 independent runs. The source angles are clearly well resolved at this low SNR. In Figure 2 the results of the case of one signal being well separated from two closely spaced sources is shown. Three independent sources with $\sin v_1 = 0.15$, $\sin v_2 = 0.2$ and $\sin v_3 = 0.4$ were used with SNR = 10dB and r = 3. The results of the five runs show that all the three angles are well estimated by one step of the modified coherent signal subspace processing method described above, whereas in [27] at least three iterations were required to ensure the angular positions of the three sources. Also the focusing transformation matrix inherently required previous estimate of the arrival angles. The results for the case of two completely correlated sources [23] are shown in Fig. 3 for SNR = 10dB and r = 4. The results show that the angles were resolved in all cases though there seems to be some variability in the estimate of the source at $v_2 = 12.0°$.

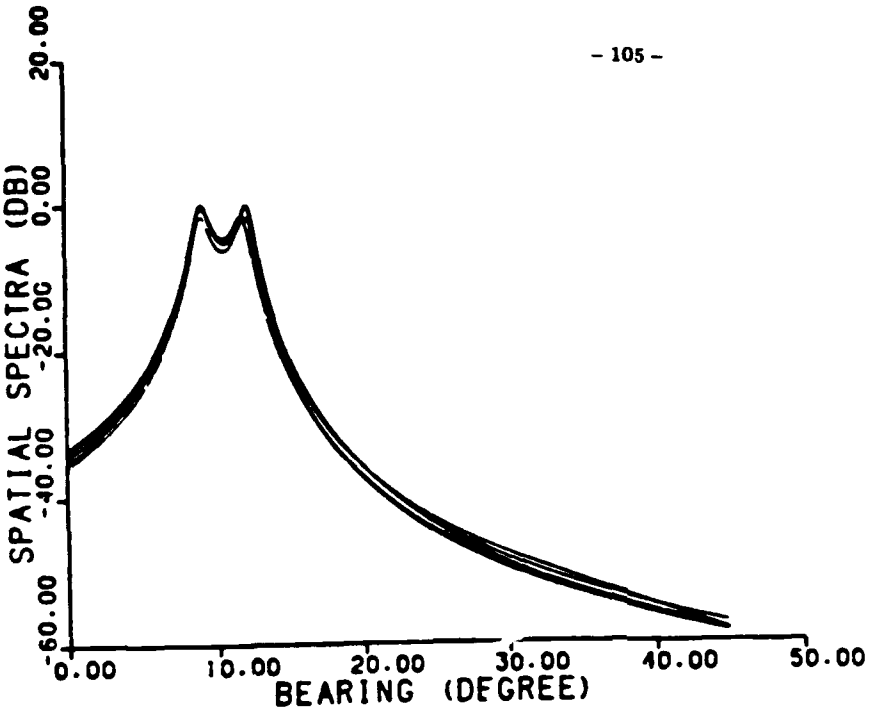

Fig. 1 : The spatial spectrum of five runs when two uncorrelated wideband sources are at $v_1 = 9°$ and $v_2 = 12°$. $r = 4$ and SNR = 10dB.

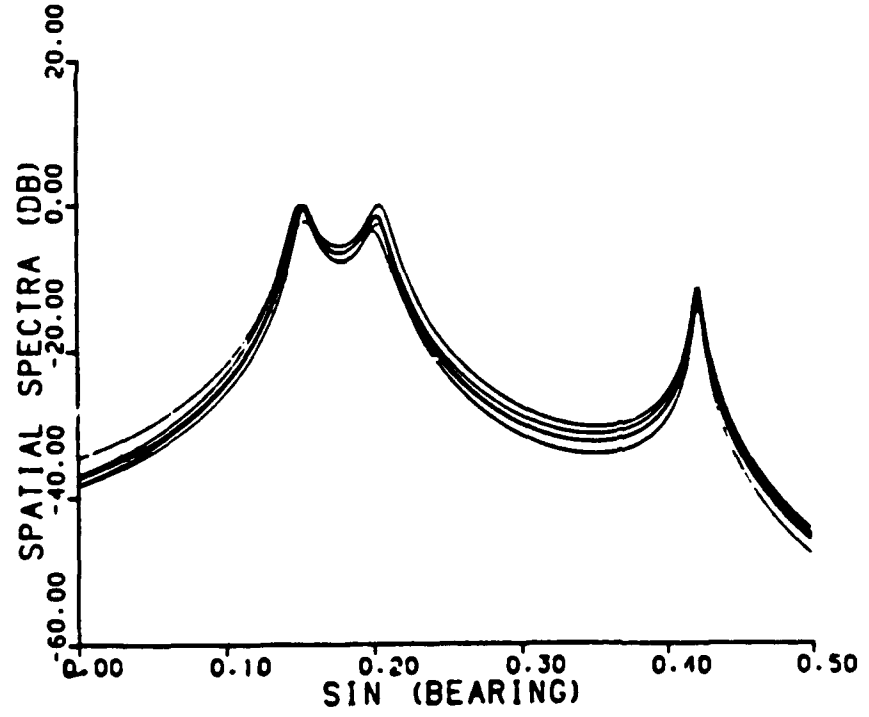

Fig. 2 : The spatial spectrum of five runs when three uncorrelated wideband sources are at $\sin v_1 = 0.15$, $\sin v_2 = 0.2$ and $\sin v_3 = 0.4$. $r = 3$ and SNR = 10dB. Note that all the angles are estimated at a single step.

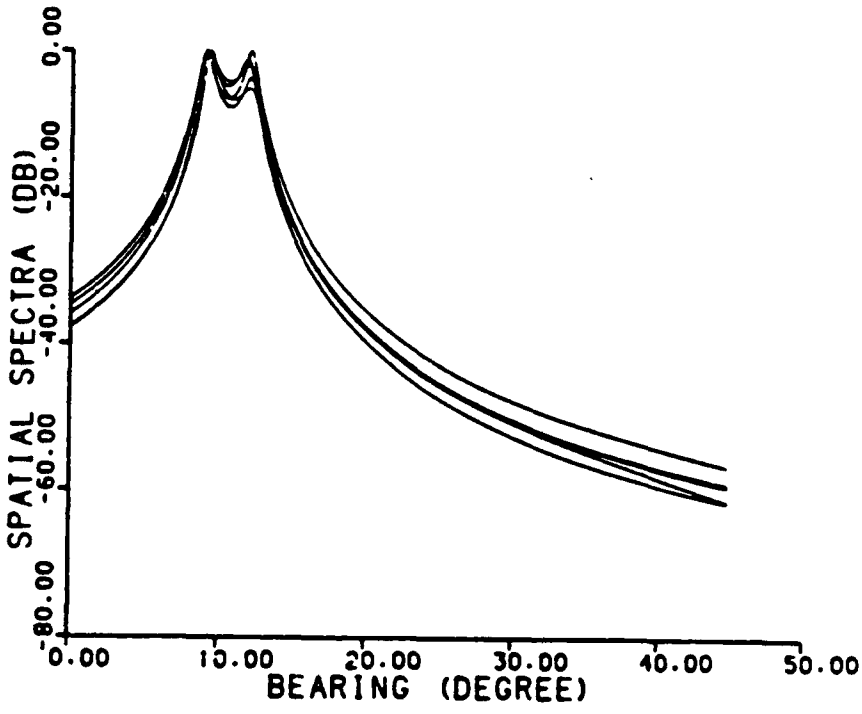

Fig. 3 : **The spatial spectrum of five runs when two correlated wideband sources are at $v_1 = 9°$ and $v_2 = 12°$. $r = 4$ and SNR = 10dB.**

References

[1] L. C. Palmer [1974], "Coarse Frequency Estimation Using the Discrete Fourier Transform," *IEEE Transactions on Information Theory*, Vol. IT-20, pp. 104-109, Jan..

[2] D. C. Rife and R. R. Boorstyn [1974], "Single Tone Parameter Estimation from Discrete-Time Observations," *IEEE Transactions on Information Theory*, Vol. IT-20, pp. 591-598, Sept..

[3] V. F. Pisarenko [1973], "The Retrieval of Harmonics from Covariance Functions," *Geophysical Journal of the Royal Astronomical Society*, Vol. 33, pp. 347-366.

[4] N. L. Owsley, "Adaptive Data Orthogonalization", *Proc. of ICASSP 78*, pp. 109-112, 1978.

[5] R. O. Schmidt, "A Signal Subspace Approach ...", Ph. D. Dissertation, Dept. of Elect. Engg., Stanford University, 1981.

[6] S. S. Reddi [1979], "Multiple Source Location- A Digital Approach," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-15, no.1, pp. 95-105.

[7] R. Kumaresan [1982(b)], *Estimating the Parameters of Exponentially Damped and Undamped Sinusoidal Signals*, Ph. D. Dissertation, University of Rhode Island.

[8] R. Kumaresan and D. W. Tufts [1983], "Estimating the Angles of Arrival of Multiple Planewaves," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-19, no .1, pp. 134-139, Jan..

[9] D. H. Johnson [1982], "The Application of Spectral Methods to Bearing Estimation Problems," *Proceedings of the IEEE*, Vol. 70, pp. 1018-1028, Sept.. Sept. 1982.

[10] S. M. Kay and A. K. Shaw [1988], "Frequency Estimation by Principal Component Autoregressive Spectral Estimator Without Eigendecomposition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol-36, no. 1, pp. 95-101.

[11] R. Kumaresan, L. L. Scharf and A. K. Shaw, "An Algorithm for Pole-Zero Modeling and Spectral Estimation," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol.ASSP-34, pp. 637-640, June, 1986.

[12] R. Kumaresan and A.K. Shaw, "Superresolution by Structured Matrix Approximation", *IEEE Transactions on Antennas and Propagation*, Vol. AP-36, pp. 34-44, 1988.

[13] S. M. Kay [1984], "Accurate Frequency Estimation at Low Signal-to-Noise Ratio," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 540-547, June.

[14] L. B. Jackson and H. C. Chien [1979], "Frequency and Bearing Estimation by Two-Dimensional Linear Prediction," in the *Proceedings of the IEEE International Conference of Acoustics, Speech and Signal Processing-1979*, Washington, DC, pp. 665-668, Apr..

[15] R. Kumaresan and D. W. Tufts [1981(b)], "A Two-Dimensional Technique for Frequency-Wavenumber Estimation," *Proceedings of the IEEE*, vol. 69, no. 11, pp. 1515-1517, Nov..

[16] J. H. McClellan [1986], "Two-Dimensional Spectrum Analysis in Sonic Logging," *IEEE ASSP Magagine*, vol. 3, no. 3, pp. 12-18, July.

[17] A. K. Shaw and R. Kumaresan, "Frequency-Wavenumber Estimation by Structured Matrix Approximation," *Proceedings of the 3rd. IEEE-ASSP Workshop on Spectral Estimation*, Boston, pp. 81-84, Nov. 1986.

[18] G. C. Carter, [Editor], "Special Issue on Time-Delay Estimation", *IEEE Trans. on ASSP*, vol. 29, no. 3, 1981.

[19] C. H. Knapp and G. C. Carter [1976], "The Generalized Correlation Method for Estimation of Time Delay,"

IEEE Transactions Acoustics, Speech and Signal Processing, vol.24, no.4, pp.320-327.

[20] W. J. Bangs and P. Schultheiss [1973], "Space-Time Processing for Optimal Parameter Estimation," in *Signal Processing*, J. W. R. Griffiths et al, Editors, New York, Academic Press, pp. 577-590.

[21] W. R. Hahn and S. A. Tretter [1973], "Optimum Processing for Delay Vector Estimation in Passive Signal Array," *IEEE Transactions Information Theory*, vol.19, no.5, pp.608-614.

[22] M. Wax and T. Kailath [1983], "Optimum Localization of Multiple Sources by Passive Arrays," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.ASSP-31, no.5, pp.1210-1218, Oct..

[23] M. Morf. et al [1981], "Investigation of New Algorithms for Locating and Identifying Spatially Distributed Sources and Receivers," DARPA, Technical Report M355-1.

[24] B. Porat and B. Friedlander [1983], "Estimation of Spatial and Spectral Parameters of Multiple Sources," *IEEE Transactions on Information Theory*, vol.IT-29, no. 6, pp. 412-425, May.

[25] A. Nehorai, G. Su and M. Morf [1983], "Estimation of Time Difference of Arrival for Multiple ARMA Sources by Pole Decomposition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.ASSP-31, pp.1478.-1491, Dec..

[26] M. Wax, T. J. Shan and T. Kailath [1984], "Spatio-Temporal Spectral Analysis by Eigenstructure Methods," *IEEE Transactions on Acoustics, Speech and Signal Processing.* vol. ASSP-32, no. 4, Aug..

[27] H. Wang and M. Kaveh [1984], "Estimation of Angles of Arrival for Wideband Sources," in *Proceeedings of the IEEE Internatinal Conference on Acoustics, Speech and Signal Processing-1984*, San Diego, California, pp. 7.5.1-7.5.4, Mar. 19-21.

[28] H. Wang and M. Kaveh [1985], "Coherent Signal Subspace Processing for the Detection and Estimation of Angles of Arrival of Multiple Wideband Sources," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol.ASSP-33, no.4, pp.823-831, Aug..

[29] T. L. Henderson [1985], "Rank Reduction for Broadband Waves Incident on a Linear Receiving Aperture," *19th Asilomar Conference on Circuits, Systems and Computers*, Nov..

[30] A. D. Whalen, *Detection of Signals in Noise*, Academic Press, 1971.

[31] S. Haykin et al, Editors, *Array Signal Processing*, Prentice-Hall, 1985.

[32] A. K. Shaw and R. Kumaresan, "Estimation of Angles of Arrivals of Broad-band Sources," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2296-2299, Dallas, Apr. 1987.

[33] M. Marden [1966], *Geometry of Polynomials*, Math. Surveys, no. 3, American Mathematical Society, Providence, RI, pp.

[34] V. Ramachandran et al. [1984], "Direct Design of Recursive Digital Filters Based on a New Stability Test," *Journal of the Franklin Institute*, vol. 318, no. 6, pp. 407-414, Dec..

SECTION 3.4 : ORDER RECURSIVE PARAMETRIC BISPECTRUM ESTIMATION

SUMMARY

Order recursive computation of AR parameters from cumulants is addressed. If the Cumulant matrix is neither Toeplitz nor symmetric, it is shown that using a block matrix inversion formula due to Frobenius and Schur [1], the inverse of the p-dimensional cumulant matrix can be updated from the $(p-1)$-dimensional inverse with $O(p^2)$ operations. When compared to batch mode computation, the proposed algorithm reduces the computational requirement for order-recursive calculation of the AR-parameters. When the cumulant matrix is non-symmetric Toeplitz also, further reduction in computation is obtained using an algorithm due to Trench [7].

I. INTRODUCTION

With the advent of faster and cheaper digital computers, the signal processing community is paying a timely and much well deserved attention to spectrum estimation methods based on higher-order statistics [2-5]. Many problems that are known to be intractable in the second-order domain of auto-correlation and Power Spectrum, can be rather easily handled in the higher-order domains [see examples in 2-4], albeit with increased computational load. The last decade has seen a renewed interest in developing algorithms based on higher-order statistics and with the cost of computing coming down, these methods will be practicable in the near future.

Following the footsteps of success in Power Spectrum estimation, linear time-invariant model-based parametric approaches have been found to be very effective in the higher-order domains also [2-8]. Specifically, estimation of the AR parameters from the third moments plays a key role in AR and ARMA approaches. Almost all the presently available techniques assume the knowledge of the model order (p) and the AR parameters are computed in batch mode by computing the inverse of a $p \times p$ third moment or cumulant matrix (which may or may not be Toeplitz). In practice, of course, the exact model order will not be known and one has to determine it from the available data. The order determination remains an open problem but any order selection approach [6, for example] would most likely require the estimation of the AR parameters for each order starting from order 1. Hence, at each order $k = 1, 2, \ldots, p$, one would have to invert a $k \times k$ cumulant matrix. But the cumulant matrix usually contains a good deal of structure in the sense that the matrix at lower model order remains embedded in the matrix of higher model order. The main purpose of this work is to exploit the structure in the cumulant matrices for order-recursive calculation of the inverse of the cumulant matrix for the k^{th} order based on the inverse calculated at $(k-1)^{th}$ order. The AR parameters are also computed order recursively in the process.

Obviously, the motivation of this work comes from the well known Levinson recursion algorithm for computing the AR parameters using second order statistics or autocorrelation estimates. It should be emphasized here though that unlike the autocorrelation matrix arising in the normal equations, the cumulant matrix is not necessarily Hermitian Toeplitz. Furthermore, the nice orthogonality properties on which the Levinson recursion is based on do not exist here. Instead, a block matrix inversion lemma [1], which is applicable to any general invertible matrix, is utilized here. This enables recursive computation of the AR parameters at increasing model orders without having to invert the whole cumulant matrix from scratch for each model order. In some cases, the cumulant matrix may also have Toeplitz structure and then the order-recursive computation of the inverse of the cumulant matrix and the corresponding AR-parameter vector can be performed using an algorithm due to Trench [9,10]. If the underlying process is of high but unknown model order, the proposed schemes will result in considerable computational savings.

This Section is arranged as follows: In Subsection II, the problem formulation is given. In Subsection III, two order recursion algorithms are given. The first one is applicable for cumulant matrix with general structure

whereas the second one is meant for Toeplitz cumulant matrix. Finally, some discussions regarding the algorithms are given in Subsection IV.

II. Problem Formulation

The third moment (or cumulant) sequence of a real discrete process $x(n)$ is defined as,

$$c(l,m) \triangleq E\{x(n)x(n+l)x(n+m)\}. \tag{1}$$

$E\{\cdot\}$ denotes the expectation operation. We will assume the process $x(n)$ to be zero mean and third-order stationary. Now consider the case when $x(n)$ is a kth order auto-regressive process, *i.e.*, the present value of $x(n)$ is formed by the linear combination of the past k samples plus driving noise sample as described by the following regression formula:

$$x(n) = -\sum_{i=1}^{k} a(i)x(n-i) + w(n). \tag{2}$$

The noise samples $w(n)$ are assumed to be zero-mean, white and non-Gaussinaly distributed with $E\{w^2(n)\} = \sigma^2$ denoting the variance and $E\{w^3(n)\} = \beta$ denoting the third moment. In (2), causal AR model is assumed, though the non-causal case can also be handled by the order-recursion algorithm considered here.

Multiplying both sides of (2) by $x(n-l)x(n-m)$ and taking expectation we get,

$$c(-l,-m) = \sum_{i=1}^{k} a(i)c(i-l,i-m) + \beta\delta(l,m), \quad \text{for } l,m \geq 0. \tag{3}$$

These equations are known as the *third-order recursion* (TOR) equations [2]. There are many possibilities of writing the above set of equations in the matrix equation form in order to solve for the AR parameters. In [2], [3] and [5] two possibilities which include the $l = m = 0$ point (*i.e.*, the origin or of $l-m$ plane) were proposed. In the first case, they chose the $l = m$ line to obtain the following matrix equation for AR model order k :

$$\begin{pmatrix} c(0,0) & \dots & c(k,k) \\ c(-1,-1) & \dots & c(k-1,k-1) \\ \vdots & \ddots & \vdots \\ c(-k,-k) & \dots & c(0,0) \end{pmatrix} \begin{pmatrix} 1 \\ a_1^{(k)} \\ \vdots \\ a_k^{(k)} \end{pmatrix} \triangleq \begin{pmatrix} \beta^{(k)} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{4a}$$

$$\mathbf{C}_T^{(k+1)}\tilde{\mathbf{a}}^{(k)} \triangleq \mathbf{b}^{(k)}. \tag{4b}$$

Note that the $(k+1) \times (k+1)$ cumulant matrix $\mathbf{C}_T^{(k+1)}$ is Toeplitz but, in general, not symmetric. The Toeplitz structure is denoted by the subscript $_T$ in (4). In another representation considered in [2], [3] and [5], l and m values are chosen in the wedge region contiguous to the origin between the $l = m$ line and the $l = 0$ line. Their choice of l and m were as follows :

$$l = 0,\dots,L, \text{ and } m = \begin{cases} 0,\dots,l, & \text{for } l < L \\ 0,\dots,M, & \text{for } l = L, \end{cases} \tag{5a}$$

where L and M are chosen such that $M \leq L$, and

$$k = 1 + M + \frac{(L-1)(L+2)}{2}. \tag{5b}$$

Note that the cumulant matrix formed in the second case is not Toeplitz. For example, if order $k = 3$, one would have $L = 2$ and $M = 0$ and

$$\begin{pmatrix} c(0,0) & \dots & c(2,2) & c(3,3) \\ c(-1,0) & \dots & c(1,2) & c(2,3) \\ c(-1,-1) & \dots & c(1,1) & c(2,2) \\ c(-2,-1) & \dots & c(0,1) & c(1,2) \end{pmatrix} \begin{pmatrix} 1 \\ a_1^{(3)} \\ a_2^{(3)} \\ a_3^{(3)} \end{pmatrix} \triangleq \begin{pmatrix} \beta^{(3)} \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{6a}$$

$$\mathbf{C}_{NT}^{(4)} \bar{\mathbf{a}}^{(3)} \triangleq \mathbf{b}^{(3)}. \tag{6b}$$

The subscript $_{NT}$ denotes the non-Toeplitz structure of this choice of cumulant matrix. Certainly, the above two choices are not unique for determination of the AR-parameters because according to equation (3) one could select any set for the bottom k equations as long as $l, m \geq 0$ (the top equation for $l = m = 0$ is needed to determine β). In fact, one can form another set of equations with Toeplitz cumulant matrix utilizing the third moments on any straight line, $l = m + d$, where $d \geq 0$ is a constant. Also for order $k = 3$, one could use the equation with $l = 2$ and $m = 0$ to replace the fourth row equation in (6). With true cumulant values, any of these sets of equations will give the correct AR parameters as solutions.

Now, let us look at the cumulant matrix for the $p = 4$ case corresponding to (5), then,

$$\begin{pmatrix} c(0,0) & c(1,1) & c(2,2) & c(3,3) & : & c(4,4) \\ c(-1,0) & c(0,1) & c(1,2) & c(2,3) & : & c(3,4) \\ c(-1,-1) & c(0,0) & c(1,1) & c(2,2) & : & c(3,3) \\ c(-2,-1) & c(-1,0) & c(0,1) & c(1,2) & : & c(2,3) \\ \dots\dots & \dots\dots & \dots\dots & \dots\dots & \dots & \dots\dots \\ c(-2,0) & c(-1,1) & c(0,2) & c(1,3) & : & c(2,4) \end{pmatrix} \tag{7}$$

Note that the cumulant matrix $\mathbf{C}_{NT}^{(4)}$ of (6) reappears in (7) as the upper left corner partitioned block. The situation will be the same if we used the Toeplitz structure of (4) or any other set of equations satisfying (3). This property that the cumulant matrix at higher model order contains the matrix of lower order will be utilized later in the recursion algorithm.

Whether the general non-Toeplitz structure of (6) or the Toeplitz structure of (4) is used, we can, in general, write the set of $k + 1$ equations as,

$$\begin{pmatrix} c(0,0) & c(1,1) & \dots & c(k,k) \\ \mathbf{c}^{(k)} & & \mathbf{C}^{(k)} & \end{pmatrix} \begin{pmatrix} 1 \\ a_1^{(k)} \\ a_2^{(k)} \\ \vdots \\ a_k^{(k)} \end{pmatrix} \triangleq \begin{pmatrix} \beta^{(k)} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{8}$$

The elements of $\mathbf{c}^{(k)}$ and $\mathbf{C}^{(k)}$ will be denoted generically as,

$$\mathbf{c}^{(k)} \triangleq [c_1 \; c_2 \; \dots \; c_k]^T \tag{9a}$$

and

$$\mathbf{C}^{(k)} \triangleq \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1k} \\ c_{21} & c_{22} & \dots & c_{2k} \\ \vdots & \vdots & \dots & \vdots \\ c_{k1} & c_{k2} & \dots & c_{kk} \end{pmatrix}. \tag{9b}$$

The AR parameters

$$\mathbf{a}^{(k)} \triangleq [a_1^{(k)} \; a_2^{(k)} \; \dots \; a_k^{(k)}]^T \tag{10}$$

are found by solving the lower k equations of (8), *i.e.*,

$$\mathbf{a}^{(k)} = -\mathbf{C}^{(k)^{-1}}\mathbf{c}^{(k)}. \tag{11a}$$

Then using the estimates of AR parameters,

$$\beta^{(k)} = c(0,0) + \sum_{i=1}^{k} a_i^{(k)} c(i,i). \tag{11b}$$

According to our discussion above $\mathbf{C}^{(k+1)}$ can be written as,

$$\mathbf{C}^{(k+1)} = \begin{pmatrix} & & c_{1(k+1)} \\ & \mathbf{C}^{(k)} & c_{2(k+1)} \\ & & \vdots \\ & & c_{k(k+1)} \\ c_{(k+1)1} & c_{(k+1)2} \cdots & c_{(k+1)(k+1)} \end{pmatrix} \tag{12a}$$

$$\triangleq \begin{pmatrix} \mathbf{C}^{(k)} & \mathbf{c}_1^{(k+1)} \\ \mathbf{c}_{-1}^{(k+1)^T} & c_{(k+1)(k+1)} \end{pmatrix}. \tag{12b}$$

The inverse of this partitioned matrix can be expressed as [1],

$$\mathbf{C}^{(k+1)^{-1}} = \frac{1}{\Delta^{(k+1)}} \begin{bmatrix} \Delta^{(k)}\mathbf{C}^{(k)^{-1}} + \mathbf{c}_2^{(k+1)}\mathbf{c}_{-2}^{(k+1)^T} & \vdots & -\mathbf{c}_2^{(k)} \\ \cdots\cdots\cdots\cdots\cdots\cdots & \vdots & \cdots \\ -\mathbf{c}_{-2}^{(k)^T} & \vdots & 1 \end{bmatrix} \tag{13a}$$

where the scalar Δ which is also known as the *Schur complement* is given by,

$$\Delta^{(k+1)} \triangleq c_{(k+1)(k+1)} - \mathbf{c}_{-1}^{(k+1)^T}\mathbf{C}^{(k)^{-1}}\mathbf{c}_1^{(k+1)} \tag{13b}$$

and,

$$\mathbf{c}_2^{(k+1)} \triangleq \mathbf{C}^{(k)^{-1}}\mathbf{c}_1^{(k+1)} \tag{13c}$$

$$\mathbf{c}_{-2}^{(k+1)^T} \triangleq \mathbf{c}_{-1}^{(k+1)^T}\mathbf{C}^{(k)^{-1}}. \tag{13d}$$

This relationship of the inverses at successive model orders are utilized in the recursive algorithms given later. Now we are ready for the recursive update algorithm for calculating the inverse of the cumulant matrix as well as the AR parameters at $(k+1)$th order based on the same at kth order. First we will give the most general algorithm and later the Toeplitz case will be considered.

III. The Recursion Algorithms

ALGORITHM 1 : (**General Case**) We will use the notation of equation (9) in describing the algorithm.

$$\mathbf{c}^{(1)} = [c_1]$$

$$\mathbf{C}^{(1)^{-1}} = \frac{1}{c_{11}}$$

$$\mathbf{a}^{(1)} = -\mathbf{C}^{(1)^{-1}}\mathbf{c}^{(1)}$$

$$\beta^{(1)} = c(0,0) + \mathbf{a}^{(1)}c(1,1)$$

for $k = 2, 3, \ldots, p$ **do**

$$\mathbf{c}^{(k)} = [\mathbf{c}^{(k-1)^T} \ : \ c_k]^T$$

$$\mathbf{c}_1^{(k)} = [c_{1k} \ldots c_{(k-1)k}]^T$$

$$\mathbf{c}_{-1}^{(k)} = [c_{k1} \ldots c_{k(k-1)}]^T$$

$$\mathbf{c}_2^{(k)} = \mathbf{C}^{(k-1)^{-1}}\mathbf{c}_1^{(k)}$$

$$\Delta^{(k)} = c_{kk} - \mathbf{c}_{-1}^{(k)^T}\mathbf{c}_2^{(k)}$$

$$\mathbf{c}_{-2}^{(k)^T} = \mathbf{c}_{-1}^{(k)^T}\mathbf{C}^{(k-1)^{-1}}$$

$$\mathbf{C}^{(k)^{-1}} = \frac{1}{\Delta^{(k)}}\begin{bmatrix} \Delta^{(k)}\mathbf{C}^{(k-1)^{-1}} + \mathbf{c}_2^{(k)}\mathbf{c}_{-2}^{(k)^T} & \vdots & -\mathbf{c}_2^{(k)} \\ \cdots\cdots\cdots\cdots\cdots\cdots & \vdots & \cdots \\ -\mathbf{c}_{-2}^{(k)^T} & \vdots & 1 \end{bmatrix}$$

$$\mathbf{a}^{(k)} = -\mathbf{C}^{(k)^{-1}}\mathbf{c}^{(k)}$$

$$\beta^{(k)} = c(0,0) + \sum_{i=1}^{k} a_i^{(k)}c(i,i).$$

ALGORITHM 2 : (**Toeplitz Case**) In this case we use the notation of eq. (4) :

$$\mathbf{c}^{(1)} = [c(-1,-1)]$$

$$\mathbf{C}^{(1)^{-1}} = \frac{1}{c(0,0)}$$

$$\mathbf{a}^{(1)} = -\mathbf{C}^{(1)^{-1}}\mathbf{c}^{(1)}$$

$$\beta^{(1)} = c(0,0) - \mathbf{a}^{(1)}c(1,1)$$

for $k = 2, 3, \ldots, p$ **do**

$$\mathbf{c}^{(k)} = [\mathbf{c}^{(k-1)^T} \ : \ c(-k,-k)]^T$$

$$\mathbf{c}_1^{(k)} = [c(k-1,k-1) \ldots c(1,1)]^T$$

$$\mathbf{c}_{-1}^{(k)} = [c(-k+1,-k+1) \ldots c(-1,-1)]^T$$

$$\mathbf{c}_2^{(k)} = \mathbf{C}^{(k-1)^{-1}}\mathbf{c}_1^{(k)}$$

$$\Delta^{(k)} = c(0,0) - \mathbf{c}_{-1}^{(k)^T}\mathbf{c}_2^{(k)}$$

$$\mathbf{c}_{-2}^{(k)^T} = \mathbf{c}_{-1}^{(k)^T}\mathbf{C}^{(k-1)^{-1}}$$

$$\mathbf{C}^{(k)^{-1}} = \frac{1}{\Delta^{(k)}}\begin{bmatrix} \Delta^{(k)}\mathbf{C}^{(k-1)^{-1}} + -\mathbf{c}_2^{(k)}\mathbf{c}_{-2}^{(k)^T} & \vdots & -\mathbf{c}_2^{(k)} \\ \cdots\cdots\cdots\cdots\cdots\cdots & \vdots & \cdots \\ -\mathbf{c}_{-2}^{(k)^T} & \vdots & 1 \end{bmatrix}$$

$$\mathbf{a}^{(k)} = -\mathbf{C}^{(k)^{-1}}\mathbf{c}^{(k)}$$

$$\beta^{(k)} = c(0,0) + \sum_{i=1}^{k} a_i^{(k)} c(i,i).$$

Though it is not explicit in the algorithm development, it must be emphasized here that a Toeplitz matrix being persymmetric (*i.e.*, symmetric about the cross diagonal), its inverse is persymmetric also [2]. Hence computation of the elements in either the upper left or the lower right triangle of the inverse matrix will suffice. This will provide further computational savings for the Toeplitz case. For more details on this algorithm see [9-12].

IV. Discussion

Both the recursion algorithms outlined here use the inverse of the cumulant matrix at the previous order to update the matrix inverse of successive orders and produce the AR parameter estimates, $\mathbf{a}^{(k)}$, for $k = 1, 2, \ldots, p$. If the true values of the third order moments are not available then these must be estimated from (possibly noisy) data. We should point out that if the true model order is known exactly, there is no gain in using the proposed recursion algorithms and the direct batch mode inverse will be sufficient. But if the true order is not known then one may have to check at different and increasing model orders. In such cases, the proposed update algorithm will reduce the cost of computation at higher model orders. As noted in [1, pp 188-191] for the general case, if the k-dimensional inverse is to be computed based on the known $(k-1)$-dimensional inverse, one would save $(k-1)^3$ computations. When compared to batch mode, which normally requires k^3 operations, the update would require $k^3 - (k-1)^3 = 3k^2 - 3k + 1$ operations. The computational savings will be even more for the Toeplitz case [9,10]. It may also be pointed out that TOR type equations also appear in anti-causal AR models and in ARMA parameter estimation [3, eq (5.31c)] with third-order statistics and the proposed recursion algorithms will be applicable in those cases also. Furthermore, the recursion algorithms are also applicable for order-recursive estimation of the AR part of ARMA models from autocorrelation estimates [12].

References

1) E. Bodewig, *Matrix Calculus*, North-Holland Publishing Co., Amsterdam, 1956.

2) M. R. Raghuveer and C. L. Nikias, "Bispectrum Estimation : A Parametric Approach," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. ASSP-33, no. 4, pp. 1213-1230, Oct., 85.

3) C. L. Nikias and M. R. Raghuveer, "Bispectrum Estimation : A Digital Processing Framework," *IEEE Proceedings*, vol. 75, no. 7, pp. 869-891, Jut., 87.

4) M. R. Raghuveer and C. L. Nikias, "Bispectrum Estimation via AR Modeling," *Signal Processing*, vol. 10, no. 1, pp. 35-48, Jan., 85.

5) G. B. Giannikas and J. M. Mendel, "Identification of Non-minimum Phase Systems Using Higher Order Statistics," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. ASSP-37, no. 3, pp. 360-377, Mar., 89.

6) J. K. Tugnait, "Consistent Parameter Estimation for Non-Causal Autoregressive Models via Higher-Order Statistics," *Automatica*, vol. 26, no. 1, pp. 51-61, Jan. 1990.

7) J. K. Tugnait, "Consistent Order Selection for Non-Causal Autoregressive Models via Higher-Order Statistics," *Automatica*, vol. 26, no. 2, pp. 311-320, 1990.

8) Special issue on Higher order spectrum, *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. ASSP-37, no. 3, pp. 360-377, Oct., 85.

9) S. Zohar, "Toeplitz Matrix Inversion : The Algorithm of W. F. Trench," *Journal of the Association for*

Computing Machinery, vol. 16, no. 4, pp. 592-601, Oct. 1969.

10) S. Zohar, "The Solution of Toeplitz Toeplitz Set of Linear Equations," *Journal of the Association for Computing Machinery*, vol. 21, no. 2, pp. 272-276, Apr., 1974.

11) W. F. Trench, "An Algorithm for Inversion of Toeplitz Matrices," *J. Soc. Ind. Appl. Math.*, vol. 12, pp. 515-522, Sept., 1964.

12) S. M. Kay, *Modern Spectral Estimation : Theory and Application*, Prentice-Hall, NJ, 1988.

SECTION 3.5 : VOWEL AND PHONEME RECOGNITION WITH A TIME-DELAY NEURAL NETWORK

SUMMARY

A Time-Delay Neural Network architecture is used for speaker independent recognition of the long vowel sounds a, e, i, o and u. The present work extends the work in [1-4] by training the network with the LPC coefficients of speech along with FFT bin energies and by allowing longer and variable length utterances. Furthermore, the training has been performed with multiple speakers using English rather than Japanese speech. With this modifications, we have obtained 100% recognition of all vowels spoken by two speakers.

I. Introduction

Many attempts have been made to use neural networks for speech recognition. Most of these attempts have only been successful in limited situations. Often one of the limiting factors has been the requirement of precise temporal alignment. This is a necessary requirement for most recognition algorithms since recognition is achieved by matching known features with observed features. Neural networks are not different in this respect. They have been found to be very sensitive to spatial shifts. As a solution to this problem many speech recognizers have made use of segmentation algorithms to pre-segment utterances in order to find the beginning and ending of a sound, word or phrase precisely [7]. Once this is done the signal features go through nonlinear time warping algorithm in order to normalize the duration of the utterance of interest, thus achieving temporal alignment. Segmentation, however, is erroneous in itself and, when in error, causes recognition failure due to mismatch between the training utterance and the testing utterance. From this it is apparent that a recognition system that is not dependent on temporal alignment needs to be developed before speech recognition can be a reliable process.

One approach to eliminate the problem of temporal alignment as discussed above would be to classify the features based on a single time frame of an utterance. Although this would solve the alignment problem the classification performance will certainly be degraded since much of the information about the signal is found from the manner in which the features vary from one time frame to the next. For this reason it seems that this idea needs to be modified such that the recognition algorithm looks at multiple time frames at one time in order to make use of the inter-frame variations of the input features. This can be achieved by having a large window that encompasses a number of time frames. This large window is then passed over the incoming data looking for local movements in the features, thereby eliminating the need for any temporal alignment as long as the time frames are sufficiently small. The requirement for a small time frame is to ensure that the signal can be considered to be stationary during that time. For speech it has been found [7] that time frames of 10ms are adequate to meet this requirement.

The alignment problem and its solution as discussed above are quite well suited for a particular type of neural network architecture known as the Time-Delay Neural Network (TDNN). A number of recent studies [1-4,8,9] have demonstrated the ability of TDNN to handle the dynamic nature of speech for phoneme recognition. As with most neural network architectures, complex constraint satisfaction is obtained via a massively parallel configuration but in the case of TDNN the constraints can occur over a period of time as desired for speech recognition. This gives the network the ability to represent relationships between events in time and at the same time TDNN allows for invariance of these events under translation in time. With this translation invariance, the network does not require precise temporal alignment, therefore the network is able to simply scan the input features for clues. This is a necessary requirement for efficient continuous speech recognition.

The work presented here was motivated by the work reported by Waibel and his colleagues [1-4,8,9] on TDNN

†This work was partly supported by AFOSR Grant AFOSR-89-0291

for phoneme recognition. Our goal was to improve the performance of TDNN by increasing the amount of data supplied to the network. This was done by including the LPC coefficients along with the FFT bin energies. We have also trained the network with utterances having variable durations. We feel that it is an important aspect due to the extreme variability in speaking rate of different speakers at different situations. Also, restricting the utterance length to 150ms to make a decision about a vowel may limit the network's performance because depending on speaking style and the spoken word the 150ms portion may not contain all the key information to recognize the vowel. We have also trained the network with more than one speaker and have obtained 100% recognition rate.

The rest of this Section is organized as follows. A brief introduction to the TDNN architecture is given in Subsection II. Subsection III gives a description of the data used for the simulation. In Subsection IV, the experiment is discussed and some results are given along with some comparison to related work. Finally, in Subsection V some conclusions and future direction of this work are given.

II. Network Architecture

The network architecture used in this work is similar to that used by Waibel and his colleagues [1-4,8,9] for phoneme recognition of Japanese speech. The differences lie mainly in the number of inputs applied to the network and the ability to accept longer and variable length utterances. A brief description of the network architecture follows.

The TDNN is basically a modified version of the standard multi-layer feed-forward neural network [6]. The difference between the two networks lies in the way the layers are interconnected and in the incorporation of the time delays. The input to the network is divided into a sequence of 10ms time segments. As the time segments are fed to the network, they pass through a series of two time delays as shown in Figure 1. This series of time delays acts as a time window passing over the data which effectively gives the network 30ms of data at any one time, as shown in Figure 2. This entire window is fully forward connected to a hidden layer of eight units. The hidden units are then combined similarly in a five frame window, in the same fashion as in the input layer. This frame is then fully forward connected to the output layer. This then gives a sequence of outputs which are averaged together for a final classification for the entire utterance.

III. Data Description

The database used for this work was the TI 46-word speech database. The input to the network consists of 28 values. These values consist of 16 spectral coefficients, 12 auto-regressive (AR) coefficients and average power. The spectral coefficients are found by taking the Fast Fourier Transform (FFT) of each 10ms time frame of data. This data is then combined, as shown in Table 1, to form the coefficients. The computation of the AR coefficients is a standard procedure in digital signal processing [7]. These coefficients are the linear prediction coefficients that model the vocal tract as an all pole filter. It is well known that 10-12th order filter (10-12 LPC coefficients) is sufficient to model the vocal tract. We have used 12 coefficients in our simulations. The complete set of coefficients is normalized (excluding the average power) such that the average value for each set of coefficients is zero and lies within the range of -1 and +1.

One key difference in our work when compared to the ones reported in [1-4] is the utilization of complete utterances to make the recognition decisions. The referenced works restrict this length to 150ms in duration. We feel that restricting the utterance length may limit the performance of the network. This is motivated by examining the utterance length statistics for the data used in this experiment, as shown in Table 2. From the table it is clear that there are considerable variations in lengths of different utterances. It may also be observed that almost all utterances are much longer than the 150ms, as used in [1-4,8,9]. For this reason, we have trained

and tested the TDNN with complete utterances of the vowels.

IV. Experiment and Results

In our experiments, we have tested the TDNN by training and testing it to recognize the long vowel sounds a, e, i, o and u. The data consisted of 10 training and 16 testing utterances for each of the five vowels and for each of the 2 female speakers. This gives a total of 100 training and 160 testing utterances.

The training algorithm used was a modified version of the back propagation algorithm. The modifications were made in order to adapt the algorithm to the TDNN architecture [12]. One other modification was to skip the learning cycle if the output error was less than 1% [2]. This was done in order to speed up the training time with the anticipation that if the network has satisfactorily learned the particular training data then no further training will be necessary. For each training cycle, the forward pass of the network computes a sequence of three outputs. The error and weight adjustments are computed for each of the three outputs. These three sets of weight adjustments are then averaged and the final weight adjustment is made using the averaged value. By training the network in this way, it is taught based on 90ms of data rather than 70ms with very little increase in computation.

With the network architecture as described above, simulations were performed in an attempt to recognize the set of vowel sounds. The simulation was written in FORTRAN and is running on a VAX 8550. With the data described above we were able to train the network with 100% recognition for two female speakers. This result compares well with that of Waibel [3], in which he has reported a recognition rate of 98.6% for a single speaker. Our experiments indicate that by including the LPC coefficients and by using the total utterance, the network was able to correctly learn all the vowels for multiple speakers.

Two different approaches were tested to train this network with utterances of multiple speakers. First the network was trained with only one speaker and 100% recognition was achieved within 75 iterations and 1:06:57.07 (denotes hr:min:sec) of CPU time. This network was then continued to be trained with the database extended to two speakers. This resulted in 100% recognition for the two speakers in just 12 more iterations (0:27:02.33 CPU time). Hence, as a whole, a total of 1:33:59.40 CPU time was required to train 5 long vowel sounds spoken by two speakers for 100% recognition rate. The second training method used a training set from both speakers from the very beginning of the training cycle. This resulted in a training cycle of 255 iterations (9:40:01.50 CPU time). Hence, the total training time in this later approach was more than 6 times greater than that of the first method. Another interesting observation we made was that the second speaker never attained 100% recognition from its database alone with the same order of number of training iterations as required for the first speaker. This fact demonstrates two facts, 1) the database of the second speaker may not be totally representative of its testing database and 2) the network does generalize on the data that it learns. This is evident from the fact that when trained with the combined database for both speakers, the network was able to learn the vowels of both speakers rather quickly though it was not learning the second speaker when it was being trained separately.

V. Conclusion

This work shows that the increased amount of information used in this work has the potential to increase the recognition capabilities. At the present time we are training the network on an extended number of speakers. Although it is making progress, the learning rate is slowed considerably with the larger data sets. But considering its present performance, we anticipate that these results may be improved upon in the future. We have also shown elsewhere [11] that the network used here can be further generalized to learn different phonemes in English and that groups of previously trained networks can be combined to develop very large networks. Therefore we conclude by stating that small networks, such as the one used here, may be trained on small databases with limited recognition abilities and may later be combined to perform larger tasks.

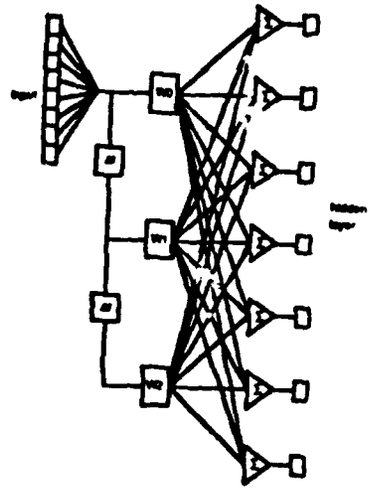

Figure 1. TDNN input

Coef #	Freq. Range
1	0 - 146.5
2	146.5 - 293.0
3	293.0 - 488.3
4	488.3 - 683.6
5	683.6 - 976.6
6	976.6 - 1296.5
7	1296.5 - 1660.2
8	1660.2 - 2050.8
9	2050.8 - 2441.4
10	2441.4 - 2929.7
11	2929.7 - 3418.0
12	3418.0 - 3906.3
13	3906.3 - 4492.2
14	4492.2 - 5078.1
15	5078.1 - 5664.1
16	5664.1 - 6250.0

Table 1 : Spectral compression scheme.

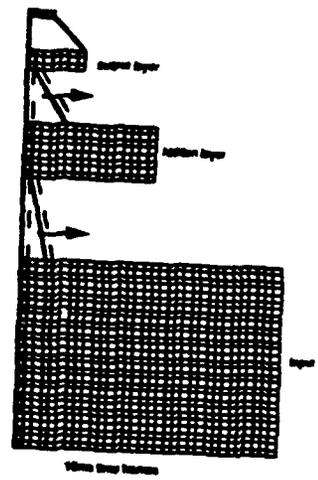

Figure 2. TDNN architecture

length(ms)	Training Set	Testing Set
minimum length	250	270
average length	421.33	423.96
maximum length	740	830
standard deviation	50.98	52.65

Table 2. Utterance length statistics

References

[1] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme Recognition using Time-Delay Neural Networks" IEEE Trans. Acoust., Speech, Signal Processing, vol. 37, pp. 328-339, Mar. 1989.

[2] H. Sawai, A. Waibel, P. Haffner, M. Miyatake, K. Shikano, "Parallelism, Hierarchy, Scaling in Time-Delay Neural Networks for Spotting Japanese Phonemes/ CV- Syllables", Proc. Int. Joint Conf. Neural Networks 1989, Vol. II, pp.II-81.

[3] A. Waibel, H. Sawai, K. Shikano, "Modularity and Scaling in Large Phonetic Neural Networks", IEEE Trans. Acoust., Speech, Signal Processing, vol. 37, pp. 1888- 1898, Dec. 1989.

[4] A. Waibel, "Modular Construction of Time-Delay Neural Networks for Speech Recognition", Neural Computation, vol. 1, pp. 39-46, 1989.

[5] R. P. Lippmann, "An Introduction to Computing with Neural Nets", IEEE ASSP Mag., pp. 4-22, Apr. 1987.

[6] D. E. Rumelhart and J. L. McClellan, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. I and Vol. II. Cambridge, MA: M.I.T. Press, 1986.

[7] L. R. Rabiner, R. W. Schafer, "Digital Processing of Speech Signals", Prentice-Hall, Inc., Englewood Cliffs, NJ 07632.

[8] H. Sawai, A. Waibel, P. Haffner, M. Miyatake, K. Shikano, "Parallelism, Hierarchy, Scaling in Time-Delay Neural Networks for Spotting Japanese Phonemes/CV Syllables", Proc. Int. Joint Conf. Neural Networks 1989, Vol. II, pp.II-81.

[9] J. B. Hampshire II, A. H. Waibel, "A Novel Objective Function for Improved Phoneme Recognition Using Time-Delay Neural Networks", Proc. Int. Joint Conf. Neural Networks 1989, Vol. I, pp.II-235.

[10] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", J. Acoust. Soc. Am. Vol. 55, No. 6, pp. 1304-1312, June 1974.

[11] A. K. Shaw and R. A. Mitchell, "Phoneme Recognition with a Time-Delay Neural Network", accepted for presentation, *International Conference on Neural Networks*, San Diego, June, 1990.

[12] R. A. Mitchell, *Speaker and Speech Recognition using Time-Delay Neural Networks*, M.S. Thesis, Wright State University, 1990.

SECTION 3.6 : SPEECH ANALYSIS AND SYNTHESIS WITH NON-LINEAR PREDICTION

SUMMARY

One of the common assumptions in speech has been that speech production and perception are essentially linear processes and hence one can accurately model speech data using 'linear prediction' based methods [1, 3]. Physiological evidence indicate that some nonlinear operation does occur in speech production and perception [4,14,15]. It is also known that linear models perform poorly for certain types of speech. Recently a non-parametric nonlinear chaos model showed significant improvement over the usual linear models [11].

In this Section we consider the applicability of certain *parametric* nonlinear models for the purpose of analysis/prediction/synthesis/coding of speech signals. To the best of our knowledge, these models have not yet been exploited for speech modeling. Several algorithms for simultaneous estimation of the *nonlinear* as well as the *linear* prediction parameters of speech signals are being considered. Preliminary studies indicate that the nonlinear models retain substantially more information when compared to linear-only models. Preliminary experiments on telephone quality speech data clearly and consistently indicate that there is a significant reduction in the prediction error when the *bilinear* prediction components are included along with the LPC part. The results of this work may have significant effect on the performance accuracy of any speech recognition/synthesis/coding system that currently relies on linear prediction only.

Linear Prediction : In linear prediction, the speech sample at a time instant n is modeled as a linear combination of past p speech samples as,

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + e(n), \tag{1}$$

where, the a_k's are the Linear Prediction Coefficients (LPC), $e(n)$ denotes the prediction error and the first term is the predicted value $\tilde{s}(n)$,

$$\tilde{s}(n) \triangleq -\sum_{k=1}^{p} a_k s(n-k). \tag{2}$$

The LP coefficients are found by minimizing the prediction error power $\sum_{n=0}^{N-1} e^2(n)$ over the entire data length. This results in a set of linear equations,

$$\mathbf{Ra} = \mathbf{r} \tag{3}$$

where the matrix $\mathbf{R}$ and the vector $\mathbf{r}$ contain estimated correlations of speech data. The predictor coefficients are then found simply by inverting the $\mathbf{R}$ matrix and forming,

$$\mathbf{a} = \mathbf{R}^{-1}\mathbf{r}. \tag{4}$$

Motivation for Nonlinear Prediction : There are several practical and theoretical grounds behind the assumption of linearity even though there is physiological evidence that nonlinearity does play a role in speech production and perception [4,14,15] and modeling [11]. Linear assumption makes processing relatively inexpensive because it only requires the solution of a set of linear equations in (3). Also, there indeed is a major contribution from the linear components within the overall nonlinear structure. Hence, linear prediction provides a practical solution with reasonably good performance. But processing speeds on digital computers have increased dramatically in the two decades following the introduction of linear prediction to the speech community [1]. Over the last decade significant advances have also been made in nonlinear parameter estimation methods. In fact, there already exists a rich mathematical theory on nonlinear modeling as well as nonlinear system identification

[2,5-10]. It appears that the potential offered by these techniques are yet to be exploited in analyzing and processing speech data for recognition, coding or synthesis purposes.

It may be noted here that recently Townshend [11] presented a nonparametric and nonlinear prediction model for speech. But in [11] speech was modeled as deterministic chaos signal. Here we model speech as stochastic signal, the same assumption traditionally made in linear prediction.

General Non-linear Case : Volterra Series : In general, the nonlinear relationship between $s(n)$ and $e(n)$ is given by a Volterra series of the form [12, 13],

$$s(n) = \sum_{k=0}^{\infty} g_k e(n-k) + \sum_{l=0}^{\infty}\sum_{m=0}^{\infty} g_{lm} e(n-l)e(n-m) + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} g_{ijk} e(n-i)e(n-j)e(n-k) + \ldots. \quad (5)$$

Note that there is no auto-regression of the output in this model and hence it may be considered a general nonlinear Moving-Average (MA) model. Estimation of the infinite set of kernel parameters $g_k, g_{lm}, g_{ijk}, \ldots$ appears computationally difficult because the kernels $g_{ijk\ldots}$ can not be estimated independently. But there are certain special cases when one can estimate these separately [6] and we intend to analyze if those special assumptions are applicable to speech signals.

Bilinear Modeling with Finite Number of Parameters : Equation (5) is a general nonlinear MA-type Volterra model. But following the trends in the linear case, it seems natural to expect that if we include output regression as well, *i.e.*, if we employ an AR or ARMA-type nonlinear model, we may obtain equivalent results even with finite number of nonlinear parameters. In fact there exists one such model called 'Bilinear model' which has appeared in control theory, nonlinear system theory and time-series analysis literature [2, 5-10]. The general bilinear predictor is expressed as,

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) + \sum_{j=1}^{q} c_j e(n-j) + \sum_{i=1}^{m}\sum_{j=1}^{k} b_{ij} s(n-i)e(n-j) + e(n). \quad (6)$$

Note that the first term in R.H.S. is the AR (or LPC) part, the second term is the MA part and the third term is the bilinear (multiplicative ARMA) part and it is denoted as BL(p, q, m, k) model. The finite order bilinear model is a parsimonious but powerful nonlinear model. In fact, it has been shown that the bilinear model with a finite parameter set can approximate any 'well behaved' infinite Volterra model of (5) to an arbitrary degree of accuracy [2].

The parameter estimation methods and the order selection methods for the BL$(p, 0, m, k)$ case have been studied by Subba Rao in [9,10]. The idea again is to estimate the linear as well as the nonlinear coefficients simultaneously by minimizing the prediction error power $\sum_{n=0}^{N-1} e^2(n)$ over the entire data length. This leads to a nonlinear optimization problem and Subba-Rao has proposed a repeated residue method and a Newton-Raphson based method.

Apart from the nonlinear models (5) and (6), there are several other specialized nonlinear models such as Diagonal Bilinear model, Subset Bilinear model and Threshold AR model [9, 10]. These models may or may not be appropriate for speech signals. We are currently investigating these models.

Simulation Results : In order to compare the performance accuracy of the linear and bilinear prediction methods preliminary tests have been carried out using the word 'STOP' from the Texas Instruments 46-Word isolated speech database. The sampling rate is 12.5 kHz. About 200 msec of speech data (Hamming windowed and 4KHz Lowpass filtered) was sectioned into 20 msec blocks with overlaps of 10 msec between adjacent data blocks. This gave us 20 data blocks to perform the comparison experiments. Our first task was to select an

'appropriate' linear predictor model order (p) which will be effective for all the data blocks. The optimum order was decided according to the MDL (Minimum Description Length) criterion though AIC (Akaike Information Criteria) was also considered. The 12th order AR model seemed to be an optimum choice for many of the blocks.

To compute the Bilinear model BL$(12, 0, m, k)$ parameters, the best values of m and k (with $m, k \leq 5$) were determined for each block of data. The parameter estimates were found for each m and k first by the repeated residue method and then using the Newton-Raphson technique [9, 10, 6]. The optimum m and k were determined for each data block according to the minimum of AIC. The iterative methods converged within 5-20 iterations. It was found consistently for each data block that the prediction errors for the bilinear case were around 3 dB lower than the 12th order AR counterpart.

The speech data for a part of the word 'STOP' is shown in Fig. 1a. In Fig. 1b the prediction errors with the 12th order AR model are plotted. Fig. 1c depicts the errors in the case of the Bilinear model BL$(12, 0, m, k)$ with optimum m and k (both ≤ 5). Comparison of the error variances for several blocks for the 12th order AR and the Bilinear model is shown in Fig. 1d. Equivalent results for another speech sound 'NO' is shown in Figures 2a-2d. These figures clearly show that the bilinear model outperforms the linear prediction model in all cases. Specifically, about 3dB improvement in prediction was observed in almost all cases. Similar results were found at other AR model orders and for other speech signals also. Currently we are exploring the effect of nonlinear modeling in speech synthesis and coding. Considering the improvement in prediction that we have already observed, we hope to see equivalent superior performance in those cases also.

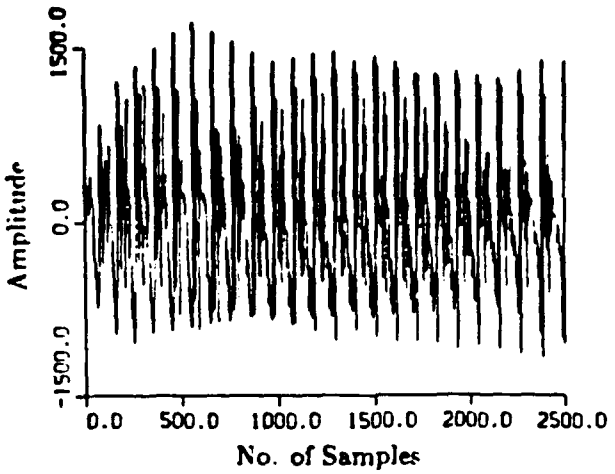

Fig. 1. The test data

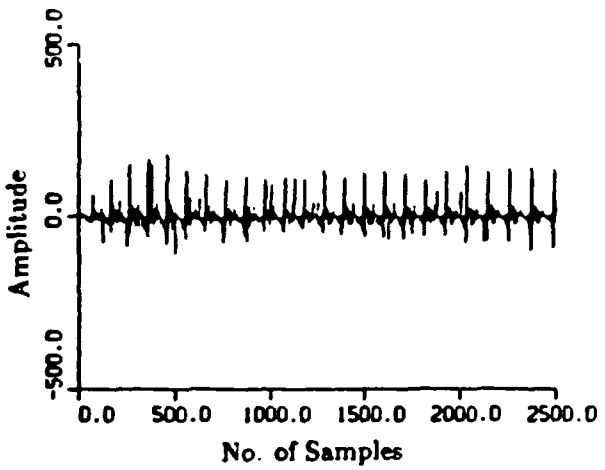

Fig. 2. The prediction errors with 12th order AR model

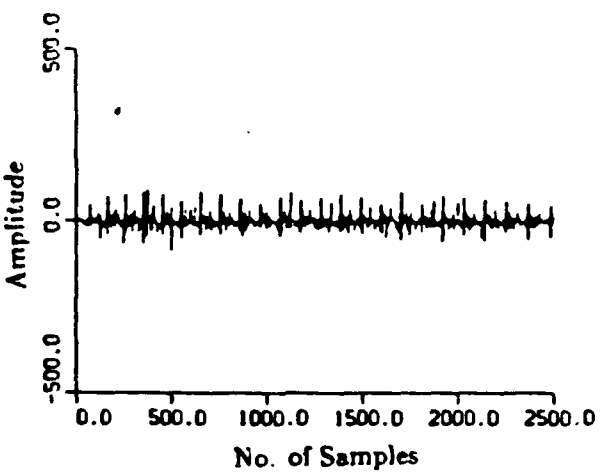

Fig. 3. The prediction errors with Bilinear model

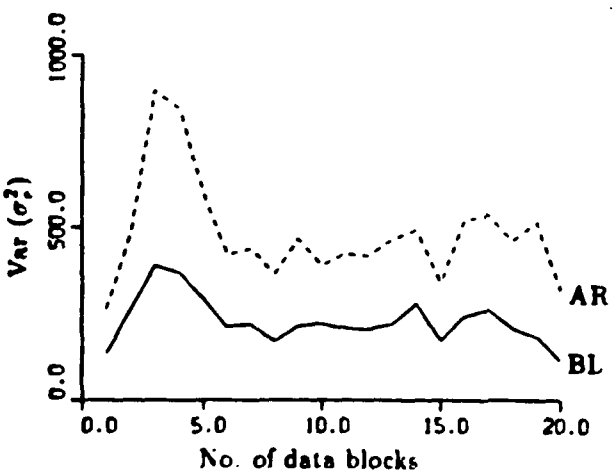

Fig. 4. Error powers for the AR and the Bilnear model

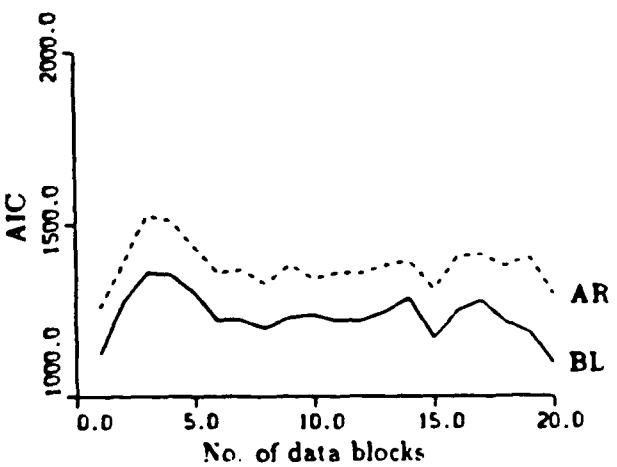

Fig 5. AICs of the AR and the Bilinear model

References

1. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. of Acoust. Soc. of Amer.*, vol. 50, pt. 2, pp. 637-655, Aug. 1971.
2. R. Brockett, "Volterra Series and Geometric Control Theory," *Automatica*, vol.12, pp. 167-176, 1976.
3. J. Makhoul, "Linear Prediction : A Tutorial Review," *Proc. IEEE*, Vol. 63, pp. 561-580, 1975.
4. R. J. Meddis, *J. Acoust. Soc. Amer.*, vol.79, pp. 703-711,, 1986.
5. R. R. Mohler, *Bilinear Control Processes*, Academic Press, London and New York, 1973.
6. M. B. Priestley, *Spectral Analysis and Time Series*, 2 Vols. Academic Press, London and New York, 1980.
7. M. B. Priestley, *Non-linear and Non-Stationary Time Series Analysis*, Academic Press, 1988.
8. A. Ruberti, A. Isidori and P. d'Allessandro, *Theory of Bilinear Dynamical Systems.*
9. T. Subba Rao, " On the Theory of Bilinear Time Series Models," *J. Roy. Stat. Soc. (B)*, Vol.43, no.2, pp.244-255, 1981.
10. T. Subba Rao and M. M. Gabr, *An Introduction to Bispectral Analysis*, Springer-Verlag, Germany, 1984.
11. B. Townshend, "Nonlinear Prediction of Speech," *ICASSP-91*, pp. 425-428, May, 1991.
12. V. Volterra, *Theory of Functional and of Integro-Differential Equations*, Dover, New York, 1959.
13. N. Wiener, *Non-linear Problems in Random Theory*, MIT Press, Cambridge, Mass, 1958.
14. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, New Jersey, 1978.
15. J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, Springer-Verlag, New York, 1972.