INT. J. CONTROL, 1988, VOL. 48, NO. 4, 1537–1550

#23

Inverse scattering and minimal partial realizations

A. M. BRUCKSTEIN†‡, T. K. CITRON†§ and T. KAILATH†

We present an inverse scattering interpretation of the classical minimal partial realization problem posed in a slightly generalized context. Our approach starts by considering a canonical cascade-form structure for the realization of arbitrary transfer functions, where the cascade structure can be interpreted as the description of a *layered* wave scattering medium. In this context the partial realization problem calls for a recursive process of layer identification from a given input–response pair (the *scattering data*). The realization algorithm uses a causality principle to progressively determine the parameters of cascaded linear 2-ports that model the successive wave-interaction layers. This method for approaching the realization problem turns out to fit nicely into a framework that was also used to obtain fast, structured linear estimation algorithms and cascade realizations for digital filters.

1. Minimal partial realization problem

The minimal realization problem of linear system theory is the following: given a finite sequence of numbers, find a minimal-order linear time-invariant system whose impulse response starts with this sequence of values. In the minimal partial realization setting it is assumed that we are given an infinite sequence of numbers and that we wish to find, for increasing values of N, minimal-order linear systems whose impulse responses match the first N lags of the given infinite sequence. Let ord (N) be the order of the minimal partial realization of the first N lags of a given sequence. Then it is interesting to investigate the behaviour of ord (N) as N increases, and also to study the uniqueness of minimal partial realizations. It turns out that the sequence ord (N) increases with N in jumps, and that the minimal partial realizations become unique at the points prior to the jump-points. At all other values of N, there are several different minimal-order partial realizations, leading to the interesting question of characterizing these sets of linear systems, by displaying some structure in which there are sets of free parameters. Any such characterization should show that the number of free parameters decreases as we approach an order-increase point, becoming zero (implying uniqueness of the realization) just prior to the jump point.

Kalman's analysis of the problem, reported in Kalman (1979), led to a recursive construction closely related to the theory of a special type of continued fraction expansion that exists for any formal power series, the so-called P-fraction expansion of Magnus (1962) (for the basic theory of continued fraction expansions see Wall 1948). The relations of P-fraction expansions to the Padé approximation problem had been known for a long time, see for example Jones and Thron (1980), and their immediate connection to the partial realization problem is hardly surprising. The classical approach to minimal partial realization problem (see Kalman 1979, Rissanen

Received 27 October 1987.

† Information Systems Laboratory, Department of Electrical Engineering, Stanford University, CA 94305, U.S.A.

‡ Now at the Department of Electrical Engineering, The Technion, Haifa, Israel.

§ Now at Hughes Aircraft Company, Space and Communications Group, Los Angeles, California, U.S.A.

1971, Kung 1977, Kailath 1980) proceeds via an analysis of the rank structure of the Hankel matrix displaying the impulse response data (the so-called Markov parameters), and realizing that the state-space representations of solutions can be obtained by a block factorization of this Hankel matrix. It was Rissanen (1971) who first noted that the necessary Hankel matrix factorization could be performed recursively, yielding *nested* sequences of partial realizations. The nesting property is desirable because it allows the updating of realizations when more and more lags of the given sequence have to be accounted for.

In this paper we *assume* that a given input–response pair, or the equivalent infinite impulse response sequence, was produced by a linear system having a cascade structure. We identify this system with a layered scattering medium, regarding it as a structure that supports waves (signals) causally propagating in opposite directions, the elementary layers describing local wave-interactions. Inverse scattering problems require the identification of such layered media from scattering data, i.e. from input–response pairs of sequences. We shall see that, for a special type of scattering medium, the minimal partial realization problem naturally falls into an inverse scattering framework. Using this framework, we show that the partial realization process may be performed by recursively determining and *peeling off* the layers of the assumed scattering structure having an infinite extent. This is different from the classical point of view which, as we saw, regards the realization process as one of matching an impulse response sequence over increasing time spans by adding 'correction terms' determined from the observation of the impulse response and the previous partial realization. This classical approach, we shall see, corresponds to an alternative way of solving the inverse scattering problem, via 'layer-adjoining'.

Within this inverse scattering framework we can also directly deal with a more general problem where, instead of being given the impulse response, we are provided with the system response to some arbitrary but known input; in this case no Hankel matrix is immediately at hand and one usually resorts to a deconvolution process to restore things to the usual impulse input problem.

Through the connection to inverse scattering, we can readily derive various algorithmic alternatives for solving the minimal partial realization problem without reference to the usual Hankel matrix rank structure and factorization; therefore it appears that this approach is more intimately related to the inherent structure of the partial realization problem. It is important to point out that the inverse scattering analogy provides not only additional insight into the problem but also a clear connection to other important computationally efficient procedures used in the theory of linear prediction, such as the Schur and Levinson algorithms, and some classical and new algorithms for the synthesis of digital filters in cascade form.

This paper is organized as follows. The next section provides a brief general overview of wave-scattering and layer-peeling algorithms for inverse scattering; alternative layer-adjoining inversion methods are also discussed. Section 3 then introduces a particular type of layered medium that turns out to have several remarkable properties. This medium model is recognized to yield minimal realizations for any finite section of the scattering data; moreover it has (by definition) the cascade structure essential for recursive updating of realizations, when longer sections of the data have to be accounted for. The medium model, together with the inverse scattering methodology, are applied in § 4 and are shown to yield various useful algorithmic alternatives for solving the minimal partial realization problem. The layer-peeling method will be seen to yield what have been called Lanczos-type

algorithms, whereas the layer-adjoining method will be recognized as the much studied Berlekamp–Massey procedure, originally developed by Berlekamp (1968) as a fast algorithm for decoding error correcting codes and reinterpreted by Massey (1969) as a minimal realization algorithm.

2. Wave scattering and inverse problems

Scattering theory deals with the analysis of signal propagation through layered media (Redheffer 1962, Bruckstein and Kailath 1983). Suppose we are given a cascade of 2-port linear processors as shown in Fig. 1. Such a system conceptually describes a

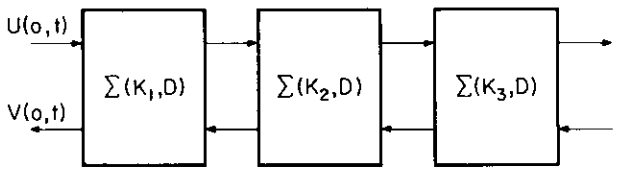

Figure 1. Wave-scattering medium: a cascade of linear 2-port systems.

layered wave-scattering medium, each processing element defining the local interaction between the propagating signals. Wavefronts are assumed to propagate in opposite directions and the typical probing of the medium is done by sending a known *right-propagating* waveform, say $U(0, t)$, into the leftmost processor and observing the causally generated *left-propagating* response or output signal, $V(0, t)$, at the same boundary.

The objective of the inverse scattering problem is to identify the medium and hence the parameters of each processor in the cascade structure of Fig. 1 from the scattering data $\{U(0, t), V(0, t)\}$ (Bruckstein and Kailath 1983, 1986, Bruckstein *et al.* 1985). By assumption, we are also given the structure of the elementary linear 2-port processors. This means that the operation of each processor is described by a parametrized *scattering* matrix, $\Sigma(K, D)$, which is a linear operator relating the outgoing waves to the incoming ones (D denotes a time shift, or delay operator, and, of course any linear operator can be expressed as a function of D). Denoting by $U(n, t)$ the right-propagating signal at depth n, and by $V(n, t)$ the corresponding left-propagating waveform, we have that

$$\begin{bmatrix} U(n+1, t) \\ V(n, t) \end{bmatrix} = \Sigma(K_{n+1}, D) \begin{bmatrix} U(n, t) \\ V(n+1, t) \end{bmatrix} \tag{2.1}$$

A mathematically equivalent way of describing the operation of such a processor is via its *transmission* matrix, $\Theta(K_n, D)$, that provides the relation between the signals at the right and the signals at the left end of a processor (see Fig. 2):

$$\begin{bmatrix} U(n+1, t) \\ V(n+1, t) \end{bmatrix} = \Theta(K_{n+1}, D) \begin{bmatrix} U(n, t) \\ V(n, t) \end{bmatrix} \tag{2.2}$$

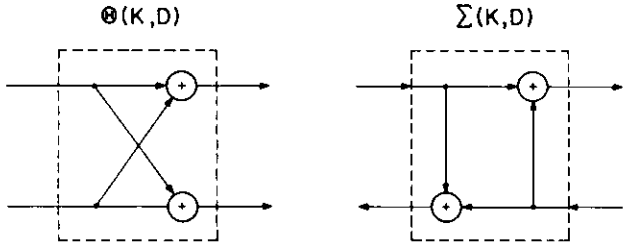

Figure 2. Transmission (*a*) and scattering (*b*) representations.

The relation between the Σ and Θ matrices is given by the following simple (Mason) exchange rule:

if

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

then

$$\Theta = \begin{bmatrix} \sigma_{11} - \sigma_{21}\sigma_{22}^{-1}\sigma_{12} & \sigma_{12}\sigma_{22}^{-1} \\ -\sigma_{22}^{-1}\sigma_{21} & \sigma_{22}^{-1} \end{bmatrix} \qquad (2.3)$$

where σ_{ij} denote linear and time-invariant operators, parametrized by some vector of parameters K. Of course, the exchange rule is the same whether we go from a scattering representation to a transmission matrix or vice versa. Note that for the equivalent representations to exist, the (2, 2)-block entry of the propagation operators (in either of the representations) has to be invertible.

The solution of an inverse scattering problem is readily obtained provided the structure of the medium enables us to determine the parametrization (K_1) of the left-most processor from the scattering data. Once this processor is identified, we can *forward propagate* the scattering data using the processor's transmission matrix, $\Theta(K_1, D)$, to obtain $\{U(1, t), V(1, t)\}$. This leaves us in the same situation we faced in the identification of the first layer, since now $U(1, t)$ and $V(1, t)$ constitute an input–response data for a medium stripped of its first layer. From this 'synthesized' scattering data we can determine K_2 and then use $\Theta(K_2, D)$ to propagate for $U(2, t)$ and $V(2, t)$. Continuing this process we can identify the entire medium one layer at a time. We call this recursive procedure a *layer-peeling* algorithm.

The crucial point in the inversion process is, of course, the possibility of identifying the nearest layer of the medium from the input and output signals appearing at its *left* terminals. This will be possible only for certain types of scattering models. For example, the inversion method works if the input signal starts at $t = 0$, with some non-zero $U(0, 0)$, and the initial portion of the observed response can be attributed solely to the reflections of the first layer and also this part of the response is sufficient to identify its parameters. The implication of this *causality* condition is that, due to propagation delays, returns from deeper layers cannot affect the initial phase of the scattering data. Formally, if we can always compute K_n via some function $\Phi\{\cdot, \cdot\}$ as follows:

$$K_n = \Phi\{U(n-1, t), V(n-1, t)\} \qquad (2.4)$$

where $U(n, t)$ and $V(n, t)$ are the waveforms produced, at depth n, by the probing signal, then we have the following layer identification and peeling algorithm.

Layer-peeling inverse scattering procedure

Given the scattering data $\{U(0, t), V(0, t)\}$.

For $n = 0, 1, 2, \ldots$ do

Step 1. Compute K_{n+1} as $\Phi\{U(n, t), V(n, t)\}$.

Step 2. Using $\Theta(K_{n+1})$ obtain the sequences $U(n+1, t)$ and $V(n+1, t)$.

Layer-peeling inverse scattering methods were first discovered in the context of geophysical inversion problems, where scientists were interested in mapping the properties of stratified earth-structures from echoes of explosions recorded on the earth surface. The acoustic wave-propagation in layered media does indeed enable the recovery of the medium parameters from such scattering data, via layer-peeling, or, as the geophysicists call it, a downward wave-continuation process (Bruckstein *et al.* 1985, Bruckstein and Kailath 1986, Bube and Burridge 1983, Robinson 1982).

There exists, however, a rather obvious alternative way for identifying the scattering medium. Indeed, note that the signals at depth n, $U(n, t)$ and $V(n, t)$ may also be obtained by passing the original scattering data $U(0, t)$ and $V(0, t)$ through the portion of the medium extending from depth 0 to depth n. This portion of the scattering medium has a transfer function which is the composition of the transmission operators $\Theta(K_n, 0), \Theta(K_{n-1}, 0), \ldots, \Theta(K_1, D)$. We can write that the transfer function of the medium portion extending over $[1, n]$, denoted by $M(n, D)$, is

$$M(n, D) = \Theta(K_n, D)\Theta(K_{n-1}, D) \ldots \Theta(K_1, D) \tag{2.5}$$

If $M(n, D)$ is already known, the lags from the sequences $U(n, t)$ and $V(n, t)$ necessary for the computation of the next medium layer parameter K_{n+1} may be obtained by convolving the original data with the weighting functions corresponding to $M(n, D)$. In practical cases of interest we shall need only a few lags of the signals at depth n, therefore we have to perform only a few such convolutions. Then, having identified K_{n+1}, we can compute the transfer function $M(n+1, D)$ as $\Theta(K_{n+1}, D)M(n, D)$. Now the original scattering data may again be used to determine the lags of $U(n+1, t)$ and $V(n+1, t)$ that are necessary to determine the next medium layer, and so on. Note that in this process we do not replace the original scattering data with the signals at increasing depth inside the medium. Instead, we rely on passing the original scattering data through the medium section that has already been identified, in order to get only the minimal amount of information necessary to identify one more medium layer. We call this alternative inverse scattering process a layer-adjoining algorithm, since it is based on recursively determining transfer functions (in the transmission representation) of increasing portions of the medium, by adjoining the newly identified layers to the previously determined sections. The layer-adjoining algorithm can be summarized as follows.

Layer-adjoining inverse scattering procedure

Given the scattering data $\{U(0, t), V(0, t)\}$

For $n = 0, 1, 2, \ldots$ do

Step 1. Determine K_{n+1} as $\Phi\{U(n, t), V(n, t)\}$, computing the necessary lags of $U(n, t)$ and $V(n, t)$ via the following convolution relations

$$\begin{bmatrix} U(n, t) \\ V(n, t) \end{bmatrix} = M(n, D)^* \begin{bmatrix} U(0, t) \\ V(0, t) \end{bmatrix}$$

Step 2. Using $\Theta(K_{n+1})$ obtain $M(n+1, D)$ as $\Theta(K_{n+1})M(n, D)$.

3. Particular layered medium model and its properties

There are many models that we can assume for the elementary layers or processors of the cascade system of Fig. 1. For reasons that will soon emerge, we shall consider a model with the scattering description

$$\begin{bmatrix} U(n+1, t) \\ V(n, t) \end{bmatrix} = \begin{bmatrix} \dfrac{D^{\alpha_{n+1}}}{T_{n+1}(D)} & -\dfrac{D^{\alpha_{n+1}}}{T_{n+1}(D)} \\ \dfrac{D^{\alpha_{n+1}}}{T_{n+1}(D)} & -\dfrac{D^{\alpha_{n+1}}}{T_{n+1}(D)} \end{bmatrix} \begin{bmatrix} U(n, t) \\ V(n+1, t) \end{bmatrix} \tag{3.1}$$

where we recall that D formally represents a time-delay operator, α_n is a positive integer and $T_n(D)$ is a polynomial of degree α_n in D. The diagram of Fig. 3 shows that the input $U(0, t)$ to response $V(0, t)$ mapping for this medium corresponds to a cascade connection of linear systems having transfer functions $1/(z^{\alpha_n} T_n(z^{-1}))$ (in terms of the usual z-transform notation, a delay corresponds to multiplication by z^{-1}) each acting as a feedback loop for its predecessor. From the structure of the elementary layers of the above-defined medium it becomes clear that both the signals passing through and those reflected by the nth layer are delayed by α_n time units. This immediately implies that the response of the medium to a causal input signal $U(0, t)$, i.e. an input sequence that is zero prior to time 0 (and $U(0, 0) \neq 0$), will appear after a delay of α_1 time units, and will be undisturbed by the echoes from the next layer for $\alpha_1 + \alpha_2$ more time units. Now since α_2 is at least 1, by assumption, and we need to identify the $\alpha_1 + 1$ parameters describing the first medium layer (the coefficients of the polynomial $T_1(D)$, of degree α_1), it follows that we have (at least from a parameter counting point of view) enough data to determine the first medium layer from the response $V(0, t)$. In fact writing out that the convolution of $U(0, t)$ with the impulse

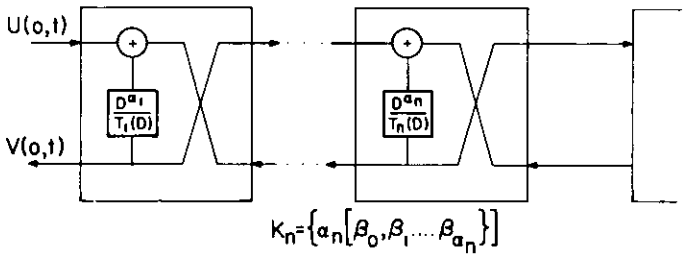

Figure 3. Nested-feedback type of scattering medium.

response corresponding to $1/(z^{\alpha_1} T_1(z^{-1}))$ yields the first $2\alpha_1 + 1$ lags of $V(0, t)$, we obtain the equations

$$\begin{bmatrix} V(0,\alpha_1) & 0 & 0 & \dots & 0 \\ V(0,\alpha_1+1) & V(0,\alpha_1) & 0 & \dots & 0 \\ & & V(0,\alpha_1) & & \\ \vdots & \vdots & \vdots & & \vdots \\ V(0,2\alpha_1) & V(0,2\alpha_1-1) & V(0,2\alpha_1-2) & \dots & V(0,\alpha_1) \end{bmatrix} \begin{bmatrix} \beta_0^{(1)} \\ \beta_1^{(1)} \\ \vdots \\ \beta_{\alpha_1}^{(1)} \end{bmatrix} = \begin{bmatrix} U(0,0) \\ U(0,1) \\ \vdots \\ U(0,\alpha_1) \end{bmatrix} \tag{3.2}$$

for the coefficients of the polynomial $T_1(z^{-1})$, denoted by $\beta_i^{(1)}$, $i = 0, 1, \dots, \alpha_1$. Since $V(\alpha_1) \neq 0$, being the first non-zero lag of $V(0, t)$, the above equations can always be solved for the unknown $\beta^{(1)}$'s. As we shall see in the next section, the identification of the coefficients of $T_1(z^{-1})$, can also be done in a nested way, by a layer-peeling process that recursively solves the Toeplitz system of equations (3.2). We shall call this a micro-level layer-peeling algorithm because it will be an intermediate step in the identification of a (macro) elementary layer.

The above discussion shows that a layer-peeling inverse scattering algorithm may be applied to determine the medium from its response to any known signal. However, this property is shared by many types of media, see e.g. Bruckstein and Kailath (1987). What makes this medium special is that, at each point in time, the linear system that is determined using the data up to that point is a *minimal-order* realization of the I/O data. To prove this claim we shall next show several interesting and striking properties of the type of medium we postulated above.

The transmission representation of the elementary layers is given by (cf. (1.3))

$$\Theta(K_n) = \begin{bmatrix} 0 & 1 \\ 1 & -D^{-\alpha_n} T_n(D) \end{bmatrix} \tag{3.3}$$

Note that $D^{-\alpha} T_n(D)$ is a polynomial of degree α_n in D^{-1}, since we had to have the free-term of $T_n(D)$ different from zero (otherwise we would have had a layer with degree $\alpha_n - 1$). Having found the transmission representation of an elementary layer, we can proceed and compute the transmission representation of the transfer function of the first N layers as

$$\begin{aligned} M(N, D) &= \begin{bmatrix} m_{11}(N, D) & m_{12}(N, D) \\ m_{21}(N, D) & m_{22}(N, D) \end{bmatrix} = \Theta(K_N)\Theta(K_{N-1}) \dots \Theta(K_1) \\ &= \begin{bmatrix} 0 & 1 \\ 1 & -D^{-\alpha_N} T_N(D) \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & -D^{-\alpha_{N-1}} T_{N-1}(D) \end{bmatrix} \dots \begin{bmatrix} 0 & 1 \\ 1 & -D^{-\alpha_1} T_1(D) \end{bmatrix} \end{aligned} \tag{3.4}$$

Consider a cascade system composed only of the first N layers of the medium. Given the input $U(0, t)$ to this system, we may ask how will the response of this system, denoted by $V^N(0, t)$ differ from $V(0, t)$. The answer is found by examining the structure of $M(N, D)$ in its scattering representation. From the Mason exchange rule,

and the structure of the elementary layers (recall causality and propagation delay) it follows that

$S(N, D)$ = Mason exchange applied to $M(N, D)$

$$= \begin{bmatrix} D^{\Lambda_N}\{\text{a causal transfer function}\} & \text{a causal transfer function} \\ -\dfrac{m_{21}(N, D)}{m_{22}(N, D)} & D^{\Lambda_N}\{\text{a causal transfer function}\} \end{bmatrix} \tag{3.5}$$

where $\Lambda_N = \alpha_1 + \alpha_2 + \ldots + \alpha_N$. This representation shows that

$$V^N(0, t) = -\frac{m_{21}(N, D)}{m_{22}(N, D)} U(0, t) \tag{3.6}$$

However, we also have that

$$V(0, t) = -\frac{m_{21}(N, D)}{m_{22}(N, D)} U(0, t) + D^{2\Lambda_N + \alpha_{N+1}}\{\text{a causal transfer function}\} U(0, t) \tag{3.7}$$

since the response of the entire medium can be obtained as the response of the first N layers, added to the response due to a feedback from $U(N, t)$ to $V(N, t)$ through the response of the medium portion starting at the $(N + 1)$ layer (see Fig. 4). This portion

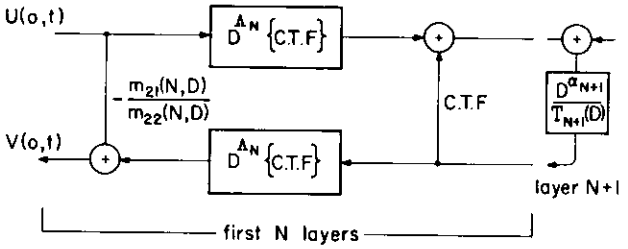

Figure 4. Scattering representation of the first N medium layers. (C.T.F. stands for 'causal transfer function').

of the medium has a causal transfer function that induces an additional delay of α_{N+1}, explaining the $2\Lambda_N + \alpha_{N+1}$ exponent of D in (3.7). (3.6) and (3.7) prove that the first $2\Lambda_N + \alpha_{N+1}$ lags of $V(0, t)$ coincide with the corresponding lags of $V^N(0, t)$. Hence we have obtained a partial realization of the response up to at least $2\Lambda_N + 1$ lags, with a cascade system of the type postulated above. This response matching result is also an expected one, since the whole purpose of the inverse scattering process was the determination of a medium that would produce the response seen to the given input. However, the type of elementary layers assumed has a crucial influence on how economically we succeed in matching a certain number of lags of the given I/O map, in terms of the order of the linear system that provides cascaded partial realization. Let us analyse this problem further. We already know that (in z^{-1} transform notation)

$$-\frac{m_{21}(N, z^{-1})}{m_{22}(N, z^{-1})} = \frac{V(0, z)}{U(0, z)} \bmod \{z^{-(2\Lambda_N + 1)}\} \tag{3.8}$$

From the definition of $M(N, D)$ we can see that $m_{22}(N, D)$ and $m_{21}(N, D)$ are

polynomials in D^{-1} of degrees Λ_N and $\Lambda_N - \alpha_0$ respectively. Furthermore from the definition of $M(N, D)$, as

$$M(N, D) = \begin{bmatrix} 0 & 1 \\ 1 & -D^{-\alpha_N} T_N(D) \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & -D^{-\alpha_{N-1}} T_{N-1}(D) \end{bmatrix} \cdots \begin{bmatrix} 0 & 1 \\ 1 & -D^{-\alpha_1} T_1(D) \end{bmatrix} \tag{3.9}$$

we can also conclude that $m_{22}(N, D)$ and $m_{21}(N, D)$ are coprime polynomials, for *any* choice of polynomials $T_i(D)$. Indeed, building up the product (3.9) from left to right shows that the polynomials $m_{22}(N, D)$ and $m_{21}(N, D)$ in D^{-1} (or z), are the result of propagating the following recursions:

$$\begin{cases} \gamma_{21}(n+1, D) = \gamma_{22}(n, D) \\ \gamma_{22}(n+1, D) = \gamma_{21}(n, D) + \gamma_{22}(n, D)\{-D^{\alpha_{N-n+1}} T_{N-n+1}(D)\} \end{cases} \tag{3.10}$$

for $n = 0, 1, 2, \ldots, N$, starting with the initial conditions

$$\gamma_{21}(0, D) = 0, \quad \gamma_{22}(0, D) = 1 \tag{3.11}$$

But (3.10) may be interpreted as an euclidean algorithm in reverse (building up the polynomials rather than reducing them), starting from the pair $\{0, 1\}$. Hence, the polynomials $m_{22}(N, D) = \gamma_{22}(N, D)$ and $m_{21}(N, D) = \gamma_{21}(N, D)$ have a g.c.d. equal to $\gamma_{21}(1, D) = \gamma_{22}(0, D) = 1$, i.e. they are coprime.

The above simple observations show that the first N layers of the medium provide an irreducible and strictly proper transfer function realization of order Λ_N that matches the response $V(0, t)$ at least up to the lag $2\Lambda_n + 1$. A simple argument, based on the Hankel matrix of Markov parameters, shows that if an irreducible system of order Λ_N produced a sequence of $2\Lambda_N$ Markov parameters, then the order of any realization of that sequence has to be at least Λ_N. (If an irreducible realization of order Λ_N produced a given response for $2\Lambda_N$ lags then the Hankel matrix of Markov parameters or impulse response lags $h_1, h_2, \ldots, h_{2\Lambda_N - 1}$

$$H_{\Lambda_N} = \begin{bmatrix} h_1 & h_2 & \cdots & h_{\Lambda_N} \\ h_2 & h_3 & \cdots & h_{\Lambda_{N+1}} \\ \vdots & \vdots & & \vdots \\ h_{\Lambda_N} & h_{\Lambda_{N+1}} & \cdots & h_{2\Lambda_N - 1} \end{bmatrix} \tag{3.12}$$

is full rank, being the product of the observability and controllability matrices of an irreducible, hence minimal realization. No lower-order realization can exist for any data set corresponding to an impulse response starting with $h_1, h_2, \ldots, h_{2\Lambda_N - 1}$, because H_{Λ_N} would then be of rank less than Λ_N. Another way of seeing this is by assuming that a lower-order realization exists and showing that this fact is equivalent to having a linear system of order less than $2\Lambda_N$, whose impulse response starts with more than $2\Lambda_N$ zeros!)

Therefore, identifying a layered system having the special form described above, we obtain realizations $-m_{21}(N, z)/m_{22}(N, z)$ that are minimal-order realizations (irreducible and obeying (3.9)) accounting for $2\Lambda_N + \alpha_{N+1}$ lags of the input–response data, the Λ_N's increasing in jumps of α_i that are also determined in the layer-peeling process. What remains to be discussed are the algorithmic aspects of determining the partial realizations, i.e. various procedures for solving the inverse scattering problem.

4. Minimal partial realizations via layer-peeling and layer-adjoining

The usual z^{-1} ($= D$) transform notation will be used throughout. The developments of § 3 in fact showed that any transfer function can be uniquely realized as a continued fraction which has the 'feedback' cascade structure shown in Fig. 3. We have that

$$\frac{V(0,z)}{U(0,z)} = \cfrac{H_1(z)}{1 + H_1(z)\cfrac{H_2(z)}{1 + H_2(z)\cfrac{H_3(z)}{1 + H_3(z)\ \ldots}}} \tag{4.1}$$

where

$$H_i(z) = \frac{z^{-\alpha_i}}{\beta_0^{(i)} + \beta_1^{(i)} z^{-1} + \ \ldots\ + \beta_{\alpha_i}^{(i)} z^{-\alpha_i}} = \frac{1}{T_i^n(z)} \tag{4.2}$$

The order of each section's elementary transfer function, $H_i(z)$, will in general differ. Equation (4.1) may also be rewritten in terms of the 'mirror' polynomials $T_i^n(z)$ as follows:

$$\frac{V(0,z)}{U(0,z)} = \cfrac{1}{T_1^n + \cfrac{1}{T_2^n(z) + \cfrac{1}{T_3^n(z) + \ \ldots}}} \tag{4.3}$$

which is the classical form of a P-fraction expansion (see Magnus 1962, Wall 1948, Jones and Thron 1980). It is clear that setting $H_{N+1}(z)$ to zero in (4.1) will yield a terminating expansion having the rational transfer function of

$$\frac{V^N(0,z)}{U(0,z)} = \frac{m_{21}(N, z^{-1})}{m_{22}(N, z^{-1})} \tag{4.4}$$

therefore we readily have the approximants of the continued fractions (4.1) from the scattering medium transmission matrices. In this section we shall apply the layer-peeling inverse scattering procedure to determine the P-fraction expansion parameters, for a given pair of scattering data. This method of doing partial realization does not provide the realizations as rational approximants. To do this, recall that we may propagate the alternative layer-adjoining procedure.

To apply the layer-peeling realization procedure is rather immediate. We have postulated a scattering medium that is described as a cascade of processors of the form appearing in Fig. 3, and have also shown how the parameters of each layer can indeed be identified. We have seen that as the layer identification algorithm proceeds it determines the minimal order that is required for the next section and then its other defining parameters, thereby ensuring the overall minimality of the partial realization sequence.

Layer-peeling algorithm

The nth processor in the cascade representing the medium is parameterized by the vector K_n comprising the order α_n and $\alpha_n + 1$ further parameters, $\beta_0^{(n)}, \beta_1^{(n)}, \ldots, \beta_{\alpha_n}^{(n)}$. Its scattering matrix is given by

$$\Sigma(K, z) = \begin{bmatrix} H(z) & -H(z) \\ H(z) & -H(z) \end{bmatrix} \quad \text{with} \quad H(z) = z^{-\alpha}\frac{1}{\beta_0 + \beta_1 z^{-1} + \ \ldots\ + \beta_\alpha z^{-\alpha}} \tag{4.5}$$

Due to this structure, recall that z^{-1} now represents the D or delay operator, an initial portion of the scattering medium response is due only to the first layer of the cascade structure. Suppose $U(0, 0) \neq 0$. Then, the number of leading zeros in the sequence $V(0, t)$ identifies the order k of the first layer. Then, we can proceed and recursively identify the parameters $\beta_0, \beta_1, \ldots$ of this layer using increasing portions of the response. Once the first layer is identified, its effect on the data can be removed by propagating the sequences U and V through the transmission matrix corresponding to the first section. The transmission matrix is

$$\Theta(K, z) = \begin{bmatrix} 0 & 1 \\ 1 & T^{\alpha}(z) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & \beta_0 z^{\alpha} + \beta_1 z^{\alpha-1} + \ldots + \beta_{\alpha} \end{bmatrix} \tag{4.6}$$

With the propagated data we face the same situation as before, i.e. we need to solve a new minimal partial realization problem. Proceeding in this way, we recursively determine the parameters of the entire cascade solution to the partial realization problem. It is easy to see that the process of identifying deeper layers effectively requires using more and more values from the initial scattering data.

A straightforward representation of the transmission matrix $\Theta(K)$ is given in Fig. 5 (*a*); however, we also have for it the following useful multiplicative decomposition (Citron and Kailath 1985) which is depicted in Fig. 5 (*b*):

$$\begin{bmatrix} 0 & 1 \\ 1 & T^{\alpha}(z) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & z^{\alpha} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \beta_0 & 1 \end{bmatrix} \begin{bmatrix} z^{-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \beta_1 & 1 \end{bmatrix} \cdots \begin{bmatrix} z^{-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \beta_k & 1 \end{bmatrix} \tag{4.7}$$

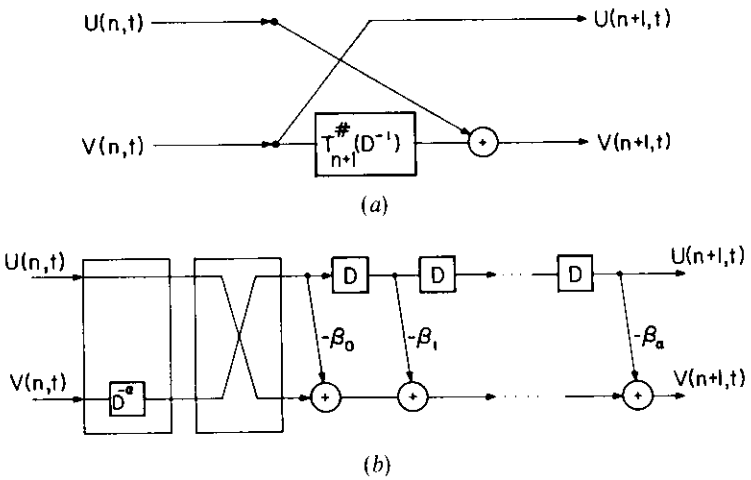

Figure 5. Immediate (*a*) and equivalent (*b*) representation of the transmission properties of a basic layer.

From the scattering representation of the elements appearing in the multiplicative decomposition (5), it becomes clear how to recursively determine the coefficients β_i too. The response V is causally generated by the medium. Further, we have that the first block of advance operators in (5) effectively removes the leading zeros of V. The

next block interchanges the sequences U and V, and then we have a simple causality picture showing that $-\beta_0$ will simply be the ratio of the now leading terms in V and U. Then we can forward propagate the sequences U and V and identify the next parameters in a completely similar way. The above process is in fact a *micro-level* inverse scattering algorithm, being performed as a method of identifying the parameters of one elementary section. We note that stopping the micro-level identification process at any point leaves a certain number of unidentified parameters. The result: the portion of the scattering data that is fitted up to this point has a non-unique partial realization, that is naturally parametrized in terms of the, as yet, unidentified β's (see Kalman 1979). The points where all the sections are uniquely determined and we need (due to the appearance of a non-zero response in $V(n, t)$) to start identifying a new section of the cascade structure are the so-called 'jump points' of the realization process. The orders of the successive partial realizations jump by the α of the next layer.

While propagating the layer-peeling process we use more and more lags of the scattering data. To identify the first N layers we need $2\Lambda_N + 1$ lags of the data, and between these points we shall be able to identify the scatters only partially, leaving the realizations non-unique. We use $2\Lambda_N + 1$ lags of the data to completely identify the first N sections, then up to the lag $2\Lambda_N + \alpha_{N+1}$ we have matching response from the uniquely identified realization, and then up to $2\Lambda_N + 2\alpha_{N+1} + 1$ lags we identify the coefficients of the next layer (having, in between, non-unique realizations parametrized naturally by the yet unidentified coefficients).

The above discussion completes the description of the minimal partial realization algorithm. Several issues concerning the relations of this approach to the ones described in the literature, in particular to the Berlekamp–Massey algorithm (Berlekamp 1968, Massey 1969) and the generalized Lanczos (Kung 1977) method will be discussed next.

Layer-adjoining and some other topics

Note that the inverse scattering algorithm that we described does not provide the rational form of the partial realization transfer function. Rather, it uses the data to identify a cascade realization for the linear system having the prescribed response. A recursive layer-adjoining algorithm can however be invoked to determine the numerator and denominator polynomials of the successively determined partial realizations. Indeed, the transfer matrix for a cascade of already identified processors up to a certain depth is simply the product

$$\begin{bmatrix} 0 & 1 \\ 1 & T_N^{=}(z) \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & T_{N-1}^{=}(z) \end{bmatrix} \cdots \begin{bmatrix} 0 & 1 \\ 1 & T_1^{=}(z) \end{bmatrix} = M(N, z) = \begin{bmatrix} m_{11}(N, z) & m_{12}(N, z) \\ m_{21}(N, z) & m_{22}(N, z) \end{bmatrix} \tag{4.8}$$

We note that in fact we may also define the transmission matrices over portions of the medium having the last section only partially identified, by using products of the elementary matrices appearing in the multiplicative decomposition of (4.7) and then we shall be able to characterize non-unique partial realizations too. We are interested in the approximants corresponding to identified portions of the scattering medium, that are given by (recall 3.5)):

$$P_n(z) = -\frac{m_{21}(n, z)}{m_{22}(n, z)} \tag{4.9}$$

Thus an updating of the above-defined transfer matrix provides the numerator and denominator polynomials of the partial realizations. Updating this transfer matrix is however easy, since we just have to multiply the transmission matrices that are recursively identified in the layer-peeling process, or alternatively we can propagate the layer-adjoining algorithm. The Berlekamp–Massey algorithm does just that, at the micro-level, and the identification of the layers is done by propagating the scattering data through the already identified transfer matrices $M(n, z)$. This amounts to the *convolution* steps that are necessary in that classical algorithm. The generalized Lanczos algorithm described in Kung (1977) and the partial realization solution proposed by Kalman (1979), are closer in spirit to the layer-peeling approach. However, those algorithms are still tied to the Hankel matrix formalism, thereby hiding the nice and more general recursive layer identification step. Also, in the classical approach the determination of the numerator polynomials is done via a Toeplitz set of equations. In our view, the only completely recursive solution of the partial realization problem is a combined propagation of the layer-peeling and layer-adjoining algorithms. This idea can be interpreted as underlying the recent development of an efficient and pipelinable parallel algorithm for decoding BCH codes, where a partial realization step is required (Citron 1984). Another outcome of the scattering approach has been the observation that, in hindsight, the Berlekamp–Massey algorithm, and in fact the partial realization process are particular micro-level implementations of Euclid's algorithm (see Citron 1984, Citron and Kailath 1985).

We note that the partial realization process described above shows that we do not need the impulse response data, i.e. that we can work equally well with any input–response sequence. The inverse scattering process thus implicitly performs a deconvolution, and this is clearly a fact that cannot be seen when proceeding via the classical Hankel matrix approaches.

5. Concluding remarks

The process of recursively identifying and peeling off layers of a scattering medium is most natural for the so-called *transmission-line* processing models. There, the cascaded 2-port processors are simply static-gain scattering matrices followed by a delay element as in Fig. 6. For such simple models the layer-peeling process turns out to be a classical algorithm of Schur, dating back to 1917, and the corresponding adjoining process is the celebrated Levinson algorithm for factoring Toeplitz matrices (Bruckstein and Kailath 1986, Kailath *et al.* 1986). These ideas have been implicit in many developments in the field of linear estimation, where fast factorization algorithms are needed for close-to-Toeplitz matrix structures (Kailath 1982, Lev-Ari and Kailath 1984, Delosme and Morf 1980, Dewilde and Dym 1981), and were

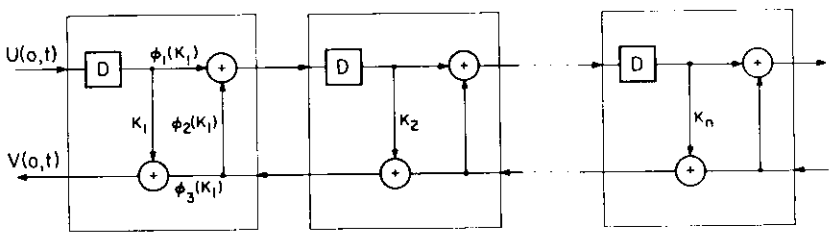

Figure 6. Transmission-line-type scattering medium.

explicitly used recently for digital filter design in VLSI (Rao and Kailath 1984). The partial realization algorithm, for the canonical case, in which all order jumps are by $\alpha = 1$, can in fact be interpreted as a generalized Brune extraction at infinity, in circuit theory jargon (Rao and Kailath 1984). In this case it is also connected to the Hankel moment problem and the corresponding orthogonal polynomials, and some interesting aspects of this problem are discussed in Kailath and Porat (1983). It is indeed pleasing to realize that so many fields of practical and theoretical interest can be so nicely related via a general inverse scattering framework.

The present paper is an extended version of a paper presented at the 23rd CDC Conference in Las Vegas in December 1984. Since then an interesting paper by Antoulas (1986) has discussed the connections between partial realizations, nestedness and recursiveness, linear fractional representations and continued fractions for the multivariable case. His paper also stresses the fact that it is not necessary to address the partial realization problem via the classical Hankel matrix formalism; however, the algorithmic alternatives of solving the realization problem, via layer-peeling or layer-adjoining procedures, are not discussed.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant DCI-84-21315-A1 and the US Army Research Office under Contract DAAL03-86-K-0045.

References

Antoulas, A. C., 1986, *I.E.E.E. Trans. autom. Control*, **31**, 1121.

Berlekamp, E., 1968, *Algebraic Coding Theory* (New York: McGraw-Hill).

Bruckstein, A. M., and Kailath, T., 1983, Spatio-temporal scattering and inverse problems, ISL Report, Department of Electrical Engineering, Stanford University; 1986, *SIAM Rev.*, **29**, 359; 1987, *Acoust. Speech Signal Process. Magazine*, **4**, 6.

Bruckstein, A. M., Levy, B. C., and Kailath, T., 1985, *SIAM J. appl. Math.*, **45**, 312.

Bube, K. P., and Burridge, R., 1983, *SIAM Rev.*, **25**, 497.

Citron, T. K., 1984, Algorithms and architectures for error-correcting codes. Thesis, Department of Electrical Engineering, Stanford University.

Citron, T. K., and Kailath, T., 1985, *I.E.E.E. Trans. Inf. Theory*, to be published.

Delosme, J. M., and Morf, M., 1980, *Proc. 19th Conf. on Decision and Control*, Albuquerqe.

Dewilde, P., and Dym, H., 1981, *Circuit Theory Applic.*, **9**, 135.

Jones, W. B., and Thron, W. J., 1980, *Continued Fractions*. Encyclopedia of Mathematics and its Applications, Vol. 11, edited by G.-C. Rota (Reading, Mass: Addison-Wesley).

Kailath, T., 1980, *Linear Systems* (New York: Prentice Hall), Chap. 5; 1982, *Mathematical Tools and Models for Control Systems Analysis and Signal Processing*, Volume 2 (Paris: Editions CNRS), p. 417.

Kailath, T., Bruckstein, A. M., and Morgan, D., 1986, *Linear Algebra Applic.*, **75**, 1.

Kailath, T., and Porat, B., 1983, In *Prediction Theory and Harmonic Analysis*, The Pesi Masani Volume (Amsterdam: North-Holland).

Kalman, R. E., 1979, *Acta Polytech. Scand.*, **MA31**, 9.

Kung, S. Y., 1977, Multivariable and multidimensional systems, Chap. 4. Thesis, Department of Electrical Engineering, Stanford University.

Lev-Ari, H., and Kailath, T., 1984, *I.E.E.E. Trans. Inf. Theory*, **30**, 2.

Magnus, A., 1962, *Math. Z.*, **80**, 209.

Massey, J. L., 1969, *I.E.E.E. Trans. Inf. Theory*, **15**, 122.

Rao, S. K., and Kailath, T., 1984, Digital filtering in VLSI. ISL Report, Department of Electrical Engineering, Stanford University.

Redheffer, R., 1962, *J. Math. Phys.*, **XLI**, 1.

Rissanen, J., 1971, *SIAM J. Control*, **9**, 420.

Robinson, E., 1982, *Proc. Inst. elect. electron. Engrs.*, **70**, 1039.

Schur, I., 1917, *J. reine angew. Math.*, **147**, 205.

Wall, H. S., 1948, *Analytic Theory of Continued Fractions* (New York: Chelsea).