JOSÉ HENRIQUE DE MORAIS GOULART

**DERIVAÇÃO EFICIENTE E UTILIZAÇÃO DE
FILTROS DE VOLTERRA DE REFERÊNCIA NA
AVALIAÇÃO DE FORMALISMOS NÃO-LINEARES**

Dissertação apresentada à Escola Politécnica
da Universidade de São Paulo para obtenção
do Título de Mestre em Engenharia Elétrica.

São Paulo
2012

JOSÉ HENRIQUE DE MORAIS GOULART

**DERIVAÇÃO EFICIENTE E UTILIZAÇÃO DE
FILTROS DE VOLTERRA DE REFERÊNCIA NA
AVALIAÇÃO DE FORMALISMOS NÃO-LINEARES**

Dissertação apresentada à Escola Politécnica
da Universidade de São Paulo para obtenção
do Título de Mestre em Engenharia Elétrica.

Área de Concentração:

Sistemas Eletrônicos

Orientador:

Prof. Dr. Phillip Mark Seymour Burt

São Paulo
2012

Este exemplar foi revisado e alterado em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 24 de setembro de 2012.

Assinatura do autor _____

Assinatura do orientador _____

FICHA CATALOGRÁFICA

Goulart, José Henrique de Moraes

Derivação eficiente e utilização de filtros de Volterra de referência na avaliação de formalismos não-lineares. -- ed.rev. -- São Paulo, 2012.
326 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Telecomunicações e Controle.

1.Sistemas não-lineares (Modelamento) 2.Processamento digital de sinais I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Telecomunicações e Controle II.t.

*Dedico este trabalho aos meus pais,
José Antonio e Delma. Esta conquista
também é de vocês.*

AGRADECIMENTOS

Gostaria de compartilhar a satisfação que sinto pelo encerramento desta jornada com diversas pessoas que, de uma forma ou de outra, foram importantes no decorrer dela.

Sendo assim, agradeço:

Ao meu orientador, professor Phillip Burt, pela dedicação com que me orientou e pela confiança em mim depositada. Sou grato também por todo o aprendizado que tive com ele durante a realização deste trabalho, especialmente sobre como pensar os problemas de forma analítica.

Aos meus pais, José Antonio e Delma, por todo o carinho, tudo o que me ensinaram e ensinam, por terem me incentivado e me auxiliado sempre e por terem me dado exemplos que jamais esquecerei.

À minha esposa, Luana, que esteve ao meu lado durante todo este percurso, pelo carinho, pelo companheirismo, pelo apoio dado sempre que precisei e por toda a compreensão que teve.

À minha irmã, Érika, minha sobrinha, Cecília, e todos os meus familiares, por todo o carinho e incentivo que sempre me deram.

Ao professor Cristiano Panazio e à professora Maria Miranda, pela gentileza e solicitude com que sempre me atenderam.

Aos colegas do LCS, os quais sempre foram prestativos e cordiais. Em especial, aos amigos Lucas Massaroppe e Amanda de Paula.

Aos todos os colegas do CETAI, com os quais aprendi bastante neste último ano. Em especial, ao professor Cláudio Garcia, pela oportunidade concedida.

A todos os grandes amigos que tenho, por terem me incentivado sempre.

RESUMO

O modelamento matemático de sistemas físicos é fundamental para diversas aplicações de processamento digital de sinais (PDS). Em muitos dos problemas enfrentados nesse contexto, para que um modelo seja útil, é necessário que ele represente seu análogo físico com precisão e possua características favoráveis para implementação, como estabilidade e compacidade.

A obtenção de um modelo que atenda a estes requisitos depende da escolha de um formalismo matemático apropriado. Em se tratando do modelamento de sistemas (significativamente) não-lineares, tal decisão é particularmente desafiadora, uma vez que muitos formalismos com propriedades diferentes foram propostos na literatura. Basicamente, isto se deve à inexistência de uma teoria completa e geral para sistemas não-lineares, diferentemente do que ocorre no caso linear.

Porém, em diversos trabalhos que lidam com aplicações nas quais é necessário modelar dispositivos não-lineares, adota-se alguma representação sem que sejam fornecidas justificativas claras e fundamentadas em características físicas do sistema a ser modelado. Ao invés disso, esse importante aspecto é discutido apenas superficialmente, com base em argumentos informais ou heurísticos. Adicionalmente, a definição de certas características estruturais de um modelo que possuem grande impacto sobre seu desempenho frequentemente não é feita de maneira sistemática, o que dificulta uma compreensão precisa do potencial do formalismo subjacente.

Visando auxiliar na escolha por um formalismo adequado em aplicações de PDS, neste trabalho propõe-se uma metodologia de avaliação do desempenho de formalismos não-lineares que se apoia sobre considerações físicas. Para tanto, emprega-se um modelo físico do sistema de interesse como referência. Mais especificamente, a estratégia adotada baseia-se em fazer uso do método de bilinearização de Carleman para se obter, a partir deste modelo e de um conjunto de parâmetros típicos, um conjunto de núcleos de Volterra de referência.

Uma vez que os núcleos de referência são obtidos, pode-se estimar, por exemplo, a ordem e a extensão de memória mínimas que um filtro de Volterra convencional deve possuir para se atingir o nível de precisão desejado, o que permite avaliar se o uso de modelos deste tipo é viável em termos de custo computacional. Quando este não é o caso, as informações fornecidas pelos núcleos podem ser exploradas para se escolher outra representação, como uma estrutura modular ou uma estrutura de Volterra alternativa. Além disso, os núcleos de referência são úteis ainda para se realizar uma avaliação quantitativa do desempenho da estrutura escolhida e compará-lo com aquele apresentado por um filtro de Volterra convencional.

Para a realização do cômputo dos núcleos de referência, um algoritmo que implementa eficientemente o método de Carleman foi proposto. Tal algoritmo, juntamente com a ideia básica da metodologia desenvolvida, constituem as principais contribuições deste trabalho.

Como estudo de caso, emprega-se um modelo físico para alto-falantes disponível na literatura para a avaliação da adequação de diversas estruturas ao mode-

lamento de dispositivos deste tipo. Com este exemplo, demonstra-se a utilidade dos núcleos de referência para as finalidades supracitadas.

Palavras-chave: Modelamento de sistemas. Sistemas não-lineares. Processamento digital de sinais. Filtros de Volterra.

ABSTRACT

The mathematical modeling of physical systems is essential for several digital signal processing (DSP) applications. In many problems faced in this context, if a model is to be useful, it must represent its physical analog with precision and possess characteristics that favour implementation, such as stability and compactness.

In order to obtain a model that meets those requirements, it is indispensable to choose an appropriate formalism. Regarding the modeling of (significantly) nonlinear systems, this decision is a particularly challenging problem, since many formalisms with different properties have been proposed in the literature. Basically, this is due to the inexistence of a complete and general theory for nonlinear systems as there is in the linear case.

In several works that deal with applications in which it is necessary to model nonlinear devices, some representation is adopted without the provision of clear and physically motivated justifications. Instead, this important aspect is discussed only superficially, based on an informal or heuristic reasoning. Additionally, the definition of certain structural characteristics of a model which have great influence on its performance is frequently done in a non-systematic manner, which difficulties a precise comprehension of the potential of the underlying formalism.

Aiming to assist the choice of an adequate formalism in DSP applications, in this work we propose a methodology for evaluating the performance of nonlinear models that relies on physical considerations. For this purpose, a physical model of the system of interest is used as a reference. Specifically, the adopted strategy is based on using the Carleman bilinearization method for obtaining a set of reference Volterra kernels from that model, considering typical parameter values.

Once the reference kernels are obtained, we can estimate, for instance, the order and the minimal memory extension that a conventional Volterra filter must have in order to achieve the desired precision level, which allows us to assess whether using models of this type is feasible in terms of computational cost. When this is not the case, the information provided by the kernels may be exploited for choosing another representation, as a modular structure or an alternative Volterra structure. Furthermore, the reference kernels are also useful for quantitatively evaluating the performance of the chosen structure and for comparing it with a conventional Volterra filter.

To perform the reference kernels computation, an efficient algorithm for the Carleman method is proposed. This algorithm, together with the basic idea of the developed methodology, constitute the main contributions of this work.

As a case study, a physical model for loudspeakers available in the literature is employed for assessing the suitability of several structures for modeling devices of this kind. With this example, we show the utility of the reference kernels for the aforementioned purposes.

Keywords: System modeling. Nonlinear systems. Digital signal processing. Volterra filters.

LISTA DE FIGURAS

1.1	Diagrama de blocos representando a configuração típica do problema de controle ativo de ruído acústico.	24
1.2	Diagrama de blocos representando a configuração típica do problema de cancelamento de eco acústico.	25
1.3	Diagrama de blocos representando a configuração típica do problema de cancelamento de vazamento de instrumentos musicais. .	27
3.1	Estrutura composta pela interconexão de um operador não-linear seguido de um linear.	81
3.2	Estrutura composta pela interconexão de um operador linear seguido de um não-linear.	82
3.3	Estrutura de um modelo N-L ou de Hammerstein.	110
3.4	Estrutura de um modelo de Hammerstein generalizado ou modelo de Uryson.	114
3.5	Estrutura de um modelo L-N.	116
3.6	Estrutura de um modelo L-N-L.	118
3.7	Estrutura de um modelo de Wiener.	123
4.1	Valores da função $C(N, p)$ definida em (4.13) para $p = 1, 2, 3, 4$. .	132
4.2	Representação de um núcleo de segunda ordem na forma triangular através de coordenadas diagonais.	136
4.3	Valores absolutos dos coeficientes de um núcleo de segunda ordem calculado a partir do modelo físico apresentado na Seção 6.1. . . .	140
4.4	Esquema esparso com interpolação para a aproximação de um filtro FIR.	144
4.5	Efeito do esquema esparso com interpolação linear sobre um filtro FIR.	145
4.6	Modelo de Wiener equivalente ao modelo de Volterra homogêneo de grau dois cujo núcleo é dado por (4.62).	154

4.7	Estrutura de Volterra alternativa PARAFAC-Volterra.	159
4.8	Modelo de modos dinâmicos principais.	166
5.1	Estrutura considerada para obtenção dos núcleos de referência de um sistema físico no contexto de aplicações de PDS.	176
5.2	Inclusão de um filtro linear que representa fatores ambientais na estrutura da Figura 5.1.	180
5.3	Conexão da estrutura da Figura 5.1 em série com um filtro linear de tempo discreto que representa fatores ambientais.	181
5.4	Resposta impulsiva de uma versão linearizada de um modelo físico para alto-falantes.	235
5.5	Comparação entre a resposta impulsiva do modelo físico do alto-falante linearizado e núcleos de primeira ordem calculados através da expansão em série de Taylor.	236
6.1	Esquema seccional de um alto-falante eletrodinâmico típico. . . .	239
6.2	Esquema de circuito elétrico análogo a um alto-falante eletrodinâmico.	240
6.3	Comportamento da função de indutância não-linear, suas derivadas primeira e segunda.	246
6.4	Comparações entre a função de indutância não-linear e sua aproximação e entre suas respectivas derivadas primeira e segunda. . .	247
6.5	Entrada usada para a validação do modelo físico do alto-falante. .	248
6.6	Comparação entre respostas produzidas pelo MF original e pelo MF ajustado.	248
6.7	Resposta em frequência do filtro incorporado nos núcleos de referência.	253
6.8	Núcleo de primeira ordem de referência obtido a partir do modelo (6.28)–(6.29).	254
6.9	Núcleo de segunda ordem de referência obtido a partir do modelo (6.28)–(6.29).	254

6.10	Valores do núcleo de primeira ordem de referência para índices próximos ao seu comprimento.	256
6.11	Comparação entre valores do núcleo de primeira ordem de referência obtidos com e sem janelamento senoidal, para índices próximos ao seu comprimento.	256
6.12	Índices de ajuste das aproximações dadas por um filtro de Volterra convencional composto pelos núcleos de referência.	261
6.13	Comportamento da dispersão relativa observada na simulação Monte Carlo feita para a avaliação do filtro de Volterra convencional de referência.	263
6.14	Comportamento das aproximações fornecidas pelo FV convencional de referência em um intervalo de potências maior.	264
6.15	Índices de ajuste das aproximações dadas por filtros de Volterra com diferentes extensões de memória.	267
6.16	Índices de ajuste das aproximações dadas por filtros de Volterra com diferentes extensões de memória.	267
6.17	Índices de ajuste das aproximações dadas pelo FV convencional de referência $\mathcal{F}_2$ e pelo filtro diagonal aproximado $\mathcal{D}_1$	272
6.18	Índices de ajuste das aproximações dadas pelo FV convencional de referência $\mathcal{F}_2$ e por outro filtro diagonal aproximado, $\mathcal{D}_2$	272
6.19	Índices de ajuste das aproximações dadas pelo FV convencional de referência $\mathcal{F}_2$ e pelo filtro esparsos com interpolação $\mathcal{S}_1$	275
6.20	Índices de ajuste das aproximações dadas pelo FV convencional de referência $\mathcal{F}_2$ e por outro filtro esparsos com interpolação, $\mathcal{S}_2$	276
6.21	Índices de ajuste das aproximações dadas pelo FV convencional de referência $\mathcal{F}_2$ e pela estrutura PARAFAC-Volterra $\mathcal{P}_1$	283
6.22	Índices de ajuste das aproximações dadas pelo FV convencional de referência $\mathcal{F}_2$ e por outra estrutura PARAFAC-Volterra, $\mathcal{P}_2$	283
6.23	Índices de ajuste das aproximações dadas pelo FV convencional de referência $\mathcal{F}_2$ e pela estrutura MDP $\mathcal{M}_1$	288

6.24	Índices de ajuste das aproximações dadas pelo FV convencional de referência $\mathcal{F}_2$ e por outra estrutura MDP, $\mathcal{M}_2$	289
6.25	Comparação entre os desempenhos das estruturas avaliadas nas aproximações de primeira e de segunda ordens.	291
6.26	Comparação entre os desempenhos das estruturas avaliadas nas aproximações de terceira e de quarta ordens.	292

LISTA DE TABELAS

4.1	Comparação entre os números de coeficientes de filtros de Volterra diagonais e de filtros de Volterra convencionais.	141
4.2	Comparação entre os números de coeficientes de filtros de Volterra esparsos com interpolação e de filtros de Volterra convencionais. .	153
4.3	Comparação entre os números de coeficientes de modelos PARAFAC-Volterra e de filtros de Volterra convencionais.	162
4.4	Comparação entre os números de coeficientes de modelos de modos dinâmicos principais e de filtros de Volterra convencionais. . . .	173
5.1	Comparação entre os números de elementos do vetor de estados não-redundante e do vetor de estados redundante.	214
5.2	Números de operações aritméticas envolvidas nas implementações direta e eficiente do método de Carleman.	218
6.1	Valores numéricos dos parâmetros para um auto-falante de graves automotivo.	245
6.2	Valores estimados para a aproximação da função de indutância não-linear dada por (6.24).	247
6.3	Número de coeficientes de cada um dos núcleos de referência calculados.	265
6.4	Número de coeficientes dos núcleos do filtro de Volterra convencional $\mathcal{F}_2$	266
6.5	Número de coeficientes dos núcleos do filtro de Volterra convencional $\mathcal{F}_3$	268
6.6	Tempo médio de cálculo das aproximações fornecidas pelos FVs convencionais $\mathcal{F}_1$, $\mathcal{F}_2$ e $\mathcal{F}_3$	269
6.7	Número de coeficientes dos núcleos do FV convencional $\mathcal{F}_2$ e dos filtros diagonais aproximados $\mathcal{D}_1$ e $\mathcal{D}_2$	273

6.8	Tempo médio de cálculo das aproximações fornecidas pelo FV convencional $\mathcal{F}_2$ e pelos filtros diagonais aproximados $\mathcal{D}_1$ e $\mathcal{D}_2$	274
6.9	Número de coeficientes dos núcleos do FV convencional $\mathcal{F}_2$ e dos filtros esparsos com interpolação $\mathcal{S}_1$ e $\mathcal{S}_2$	278
6.10	Tempo médio de cálculo das aproximações fornecidas pelo FV convencional $\mathcal{F}_2$ e pelos filtros esparsos com interpolação $\mathcal{S}_1$ e $\mathcal{S}_2$. .	278
6.11	Número de coeficientes dos núcleos do FV convencional $\mathcal{F}_2$ e das estruturas PARAFAC-Volterra $\mathcal{P}_1$ e $\mathcal{P}_2$	285
6.12	Tempo médio de cálculo das aproximações fornecidas pelo FV convencional $\mathcal{F}_2$ e pelas estruturas PARAFAC-Volterra $\mathcal{P}_1$ e $\mathcal{P}_2$. . .	285
6.13	Número de coeficientes dos núcleos do FV convencional $\mathcal{F}_2$ e das estruturas MDP $\mathcal{M}_1$ e $\mathcal{M}_2$	290
6.14	Tempo médio de cálculo das aproximações fornecidas pelo FV convencional $\mathcal{F}_2$ e pelas estruturas MDP $\mathcal{M}_1$ e $\mathcal{M}_2$	290
6.15	Comparação entre os números de coeficientes que compõem os melhores modelos das estruturas avaliadas.	293

LISTA DE ABREVIATURAS

PDS:	Processamento Digital de Sinais
MF:	Modelo Físico
FV:	Filtro de Volterra
MOB:	Modelo Orientado a Bloco
EVA:	Estrutura de Volterra Alternativa
FRFG:	Função de Resposta em Frequência Generalizada
EQMN:	Erro Quadrático Médio Normalizado
HOSVD:	<i>High-order Singular Value Decomposition</i>
PARAFAC:	<i>PARalell FAcTOR decomposition</i>
MDP:	Modos Dinâmicos Principais
FIR:	<i>Finite Impulse Response</i>
IIR:	<i>Infinite Impulse Response</i>
TFD:	Transformada de Fourier Discreta
TFTD:	Transformada de Fourier de Tempo Discreto
NARX:	<i>Nonlinear Auto-Regressive model with eXogenous inputs</i>
LIT:	Linear e Invariante no Tempo

SUMÁRIO

Capítulo 1—Introdução	21
1.1 Motivação	23
1.1.1 Linearização de alto-falantes	23
1.1.2 Cancelamento de eco acústico	24
1.1.3 Simulação digital de dispositivos analógicos	26
1.1.4 Cancelamento de vazamento de instrumentos musicais . . .	27
1.1.5 Controle	28
1.1.6 Outras aplicações	28
1.2 Justificativa	29
1.3 Objetivos	31
1.4 Contribuições	33
1.5 Organização	34
Capítulo 2—Aspectos básicos sobre o modelamento de sistemas não-lineares	36
2.1 Conceitos básicos	36
2.1.1 Sistemas	36
2.1.2 Modelos matemáticos	38
2.2 Uso de informações prévias no modelamento	40
2.3 Etapas típicas do processo de modelamento	42
2.4 Uso prático dos modelos	43
2.5 Abordagens de modelamento	45
2.5.1 Abordagem não-paramétrica	45
2.5.2 Abordagem paramétrica	46
2.5.3 Abordagem modular	47
2.5.4 Abordagem conexionista	48

2.6	Aspectos sobre estimação dos parâmetros	49
2.6.1	O estimador de mínimos quadrados	50
2.6.2	Adaptação de parâmetros	51
2.6.3	Aspectos estatísticos relacionados à estimação	52
Capítulo 3—Representações não-lineares		55
3.1	Modelos baseados na série de Volterra	56
3.1.1	A série de Volterra	57
3.1.1.1	Unicidade da representação	58
3.1.1.2	Causalidade e variação no tempo	61
3.1.1.3	Estabilidade	62
3.1.2	Cálculo da série para sistemas bilineares	64
3.1.3	Análise no domínio da frequência	68
3.1.4	Caracterização de distorções não-lineares no domínio da frequência	71
3.1.5	A transformada de Laplace multidimensional	74
3.1.6	Caracterização da resposta de um sistema de Volterra para uma combinação de exponenciais	75
3.1.7	Núcleos resultantes de algumas interconexões entre opera- dores	78
3.1.8	Questões de convergência	82
3.1.9	Aproximação pela série truncada	85
3.1.10	O Filtro de Volterra	88
3.1.10.1	Funções de resposta em frequência generalizadas de tempo discreto	91
3.1.10.2	Discretização de núcleos de Volterra	93
3.2	Modelos físicos	98
3.2.1	Definição	98
3.2.2	Núcleos de Volterra de modelos físicos	99
3.2.3	Uso e estimação de modelos físicos	100
3.3	Modelos NARX	101

3.3.1	Definição	101
3.3.2	Definição da estrutura de um modelo NARX	103
3.3.3	Relação com filtros de Volterra	104
3.4	Modelos modulares	107
3.4.1	Modelo N-L ou de Hammerstein	109
3.4.2	Modelo de Hammerstein generalizado	113
3.4.3	Modelo L-N	116
3.4.4	O modelo L-N-L	118
3.4.5	O modelo de Wiener	122
Capítulo 4—Estruturas de Volterra alternativas		127
4.1	Implementação de filtros de Volterra convencionais	128
4.1.1	Implementação direta	128
4.1.2	Implementação baseada na TFD	130
4.2	Representação dos núcleos em coordenadas diagonais	133
4.2.1	Implementação de filtros de Volterra na representação diagonal	134
4.2.2	Filtros de Volterra diagonais aproximados	137
4.2.2.1	Complexidades estrutural e computacional	139
4.3	Filtro de Volterra com interpolação de núcleos esparsos	142
4.3.1	Aproximação de filtros lineares pelo esquema esparsos com interpolação	143
4.3.2	Aproximação de filtros de Volterra pelo esquema esparsos com interpolação	146
4.3.2.1	Complexidades estrutural e computacional	151
4.4	Estruturas baseadas no modelo de Wiener	153
4.4.1	O modelo PARAFAC-Volterra	155
4.4.1.1	Obtenção dos fatores a partir de um FV de referência	159
4.4.1.2	Complexidades estrutural e computacional	162
4.4.2	O modelo de modos dinâmicos principais	163

4.4.2.1	Obtenção dos modos dinâmicos principais	167
4.4.2.2	Complexidades estrutural e computacional	172
Capítulo 5—Derivação eficiente de filtros de Volterra de referência		175
5.1	Cálculo de núcleos de referência em aplicações de PDS	176
5.1.1	Definição do problema	176
5.1.2	Fatores ambientais	179
5.1.3	Forma considerada para o modelo físico	185
5.2	Abordagem da exponencial crescente	188
5.2.1	Implementação computacional	193
5.2.2	Obtenção das FRFGs de sistemas bilineares	195
5.3	Método de bilinearização de Carleman	197
5.3.1	Implementação computacional	201
5.3.1.1	Implementação direta	201
5.3.1.2	Implementação eficiente	202
5.3.1.3	Algoritmo para a implementação eficiente	212
5.3.1.4	Comparação entre as complexidades do algoritmo eficiente e da implementação direta	214
5.3.2	Exatidão dos núcleos e das FRFGs resultantes	219
5.3.3	Aproximação de dinâmica não-linear por uma equação di- ferencial bilinear	226
5.4	Outros métodos estudados para o cálculo dos núcleos	228
5.4.1	Método da função de fluxo	228
5.4.1.1	Fórmulas analíticas	229
5.4.1.2	Cômputo numérico através das fórmulas	231
5.4.2	Cálculo pela série de Taylor dos núcleos	234
Capítulo 6—Avaliação do desempenho de formalismos não-lineares com o uso de núcleos de referência		237
6.1	Um modelo físico para alto-falantes	238
6.1.1	Princípios do modelo físico	238

6.1.2	Resultados apresentados pelo modelo físico	244
6.1.3	Ajuste do modelo físico para colocá-lo na forma linear-analítica	245
6.2	Derivação dos núcleos de referência	249
6.2.1	Obtenção das FRFGs associadas ao modelo físico	250
6.2.2	Cálculo dos núcleos de referência	251
6.2.2.1	Definição dos filtros G_1 e G_2	251
6.2.2.2	Núcleos resultantes	253
6.2.3	Inadequação de modelos modulares simples	256
6.3	Avaliação do desempenho de um FV composto pelos núcleos de referência	258
6.3.1	Aproximação fornecida pelo FV	258
6.3.1.1	Discussão dos resultados	261
6.3.2	Comportamento das aproximações em um intervalo de potências maior	264
6.3.3	Avaliação de complexidades estrutural e computacional do FV	265
6.4	Avaliação do desempenho de estruturas de Volterra alternativas .	269
6.4.1	Filtro de Volterra diagonal aproximado	270
6.4.1.1	Avaliação das aproximações obtidas	270
6.4.1.2	Avaliação da redução de complexidades estrutural e computacional	273
6.4.2	Filtro de Volterra esparso com interpolação	274
6.4.2.1	Avaliação das aproximações obtidas	275
6.4.2.2	Avaliação da redução de complexidades estrutural e computacional	277
6.4.3	Modelo PARAFAC-Volterra	279
6.4.3.1	Avaliação das aproximações obtidas	282
6.4.3.2	Avaliação da redução de complexidades estrutural e computacional	284
6.4.4	Modelo de modos dinâmicos principais	285

6.4.4.1	Avaliação das aproximações obtidas	288
6.4.4.2	Avaliação da redução de complexidades estrutural e computacional	289
6.5	Discussão dos resultados	290
Capítulo 7—Conclusões e perspectivas		296
7.1	Perspectivas de trabalhos futuros	299
Apêndice A—Revisão de conceitos matemáticos e notação		309
A.1	Álgebra de tensores	309
A.1.1	Notação e conceitos básicos	309
A.1.2	Produto e decomposição de Tucker	311
A.1.3	Produto externo	313
A.1.4	Produto de Kronecker	314
A.1.5	Operador vec	316
A.2	Expansão de Taylor de funções multivariadas	317
A.2.1	Expansão de Taylor de funções reais multivariadas	317
A.2.2	Expansão de Taylor de campos vetoriais	319
Apêndice B—Publicação		322

Capítulo 1

Introdução

Modelos matemáticos desempenham um papel fundamental nos desenvolvimentos científico e tecnológico. Por descreverem os fenômenos naturais de forma quantitativa, eles são elemento chave para a teoria de sistemas, possibilitando uma compreensão precisa do comportamento de sistemas físicos. Adicionalmente, devido a tal característica, modelos matemáticos são capazes de prever o comportamento de muitos sistemas reais, o que lhes atribui também enorme importância prática.

Para se obter um modelo matemático de um sistema físico, deve-se conduzir um processo de modelamento, para o qual diversas metodologias foram propostas na literatura. Um aspecto central em tal processo é a escolha de um formalismo matemático para representar o sistema de interesse. Quando o modelo a ser construído é linear, conta-se com a existência de formalismos gerais, como funções de transferência e respostas impulsivas, aos quais está associada uma teoria relativamente completa e bem compreendida. Contudo, o mesmo não é válido para formalismos não-lineares, o que se deve basicamente à diversidade deles e à complexidade por eles apresentada.

Historicamente, a adoção de formalismos lineares no modelamento de sistemas físicos sempre foi predominante. Contudo, são muitos os exemplos de sistemas para os quais a suposição de linearidade corresponde a uma aproximação ruim, o que impede modelos lineares de apresentarem um bom desempenho. Em tais situações, o ganho em desempenho obtido com métodos não-lineares muitas vezes compensa o aumento da dificuldade analítica envolvida no modelamento. Em face disto, nas últimas décadas é crescente o número de trabalhos que adotam uma abordagem não-linear [1]. Como exemplos recentes nas áreas de processamento

de sinais, telecomunicações e controle, podemos citar [2, 3, 4, 5, 6, 7, 8].

De modo geral, a escolha do formalismo é crítica para um modelamento bem-sucedido [9, 10]. No caso do modelamento não-linear, a obtenção de um modelo que seja preciso e tenha características desejáveis do ponto de vista prático é um problema particularmente desafiador, o qual está diretamente relacionado com tal escolha. De fato, este é um problema bastante estudado atualmente e cuja solução, em geral, é fortemente dependente do contexto específico em que surge.

Em diversas aplicações no contexto de processamento digital de sinais, a necessidade de modelar sistemas físicos surge naturalmente. Uma vez que muitos sistemas estudados nesse contexto comportam-se de forma significativamente não-linear, faz-se necessário que as não-linearidades sejam levadas em conta no modelamento para que os modelos construídos sejam capazes de representar seus análogos físicos com suficiente precisão. Além disso, para que um modelo matemático seja de fato útil neste contexto, ele deve ser, idealmente, eficientemente implementável e, em problemas de natureza variante no tempo, suscetível a uma atualização *online* de seus parâmetros (i.e., adaptação) de forma simples e eficiente. Ainda, o modelo deve ser também, evidentemente, estável, característica que pode ser bastante difícil de se verificar no caso de modelos não-lineares.

A construção de um modelo que atenda a todos requisitos mencionados no parágrafo anterior é um problema difícil, para o qual diversas abordagens têm sido propostas na literatura. Todavia, tal diversidade, associada à ausência de uma metodologia para a comparação sistemática dos desempenhos apresentados por diferentes formalismos, torna a escolha por um deles uma tarefa não-trivial. Adicionalmente, em muitos trabalhos o uso da abordagem adotada não é justificado de forma fundamentada e clara, mas sim com base em argumentos informais ou heurísticos.

1.1 Motivação

Nesta seção, serão descritas algumas aplicações de processamento digital de sinais¹ em que surge a necessidade de se representar matematicamente sistemas físicos através de estruturas eficientemente implementáveis para uso e/ou adaptação em tempo real. Além de fornecer uma motivação para o estudo realizado, essa descrição introduz alguns aspectos que serão úteis mais adiante.

1.1.1 Linearização de alto-falantes

Devido à atual tendência de miniaturização de dispositivos eletrônicos e de comunicação, é crescente o uso de alto-falantes de dimensões reduzidas. Em tais condições, para se manter um nível de potência suficiente nas baixas frequências, necessita-se que o cone do alto-falante sofra grandes excursões, o que introduz distorções não-lineares significativas na saída [11]. Evidentemente, trata-se de uma consequência indesejável, uma vez que, com isso, há um surgimento de componentes em frequências nas quais o sinal de entrada não possui energia.

Para atenuar tais distorções, uma estratégia que vêm sendo explorada na literatura, conhecida como linearização de alto-falantes (LA), consiste em aplicar um pré-processamento não-linear digital (ou pré-distorção) no sinal de entrada com a finalidade de compensar as não-linearidades do alto-falante, isto é, de modo que sua saída seja aproximadamente linear na entrada original. Uma das formas de se realizar uma linearização efetiva é através do uso de uma estrutura derivada de um modelo não-linear do alto-falante [11, 12, 13]. Observe que, como tal estrutura deve ser implementada para aplicação em tempo real (muitas vezes, em uma taxa de amostragem alta), o custo computacional associado a seu uso, o qual está diretamente relacionado às características do modelo usado na derivação, não pode ser demasiadamente elevado.

Além do uso para fins de reprodução sonora, a linearização de alto-falantes pode ser útil também na aplicação de controle ativo de ruído acústico [14]. Nesse

¹Devido à maior familiaridade dos autores com aplicações de áudio, elas constituem os exemplos tratados mais detalhadamente nesta seção.

caso, o objetivo é a atenuação do ruído incômodo produzido por alguma fonte, como um motor ou um sistema de exaustão, em um local onde é percebido após ser transmitido acusticamente. Para tanto, utiliza-se a estratégia que é ilustrada na Figura 1.1. O ruído produzido pela fonte primária, que é denotado no diagrama por $u(n)$, é captado por um microfone de referência, cujo sinal é fornecido como entrada para um filtro adaptativo. Esse filtro é usado com a finalidade de modelar o caminho acústico, denotado por $h(n)$, para que seja produzida uma estimativa do ruído percebido no local em que se deseja atenuá-lo, representada por $\hat{y}(n)$. Com isso, um sinal de mesma amplitude e fase oposta à estimativa computada, chamado de “anti-ruído”, é submetido a um alto-falante. Assim, ao se superpor ao ruído original, a saída do alto-falante surte o efeito de reduzi-lo, sendo que o ruído residual, $e(n)$, é captado por um microfone posicionado no local de atenuação para a adaptação do filtro. Deve-se observar que, quando o nível de pressão sonora do ruído é elevado, as não-linearidades do alto-falante podem introduzir componentes indesejáveis no anti-ruído emitido, o que compromete a eficácia da redução. Neste caso, a degradação do desempenho poderia ser evitada com o uso de técnicas de linearização como aquela descrita no parágrafo anterior [11].

Figura 1.1: Diagrama de blocos representando a configuração típica do problema de controle ativo de ruído acústico.

1.1.2 Cancelamento de eco acústico

Um problema bastante estudado na literatura de processamento de sinais de áudio é o de cancelamento de eco acústico (CEA). Em linhas gerais, o objetivo

no CEA é impedir que o som emitido pelo alto-falante de um equipamento de telecomunicação no modo viva-voz (que inevitavelmente é captado por seu microfone) seja retransmitido pelo canal, o que geralmente surte um efeito incômodo na forma de um eco que atrapalha a comunicação. Para isso, é necessário um modelo que permita estimar a porção da entrada captada que corresponde ao som emitido pelo alto-falante e possivelmente filtrado pela acústica do ambiente.

O diagrama de blocos da Figura 1.2 ilustra a configuração típica do problema de CEA. Nele, $u(n)$ corresponde ao sinal recebido pelo aparelho através do canal de comunicação e fornecido como entrada para que o alto-falante produza a saída $s(n)$. Em seguida, normalmente uma superposição de diversas cópias atenuadas e atrasadas de $s(n)$ são recebidas pelo microfone, devido à ação da resposta acústica do ambiente. Assim, esse sinal, denotado por $y(n)$, se mistura àquele produzido pelo usuário do equipamento, $d(n)$. Então, para que não haja interferência nociva à comunicação em andamento, o modelo indicado na figura é empregado para produzir estimativas das amostras de $y(n)$, denotadas por $\hat{y}(n)$, de tal forma que o sinal de interferência possa ser removido antes da transmissão para o usuário situado na outra extremidade do canal.

Figura 1.2: Diagrama de blocos representando a configuração típica do problema de cancelamento de eco acústico.

Para se modelar adequadamente a porção do sistema que se deve à acústica do ambiente, pode-se utilizar um filtro linear (conforme indicado no diagrama pela convolução $h(n) * s(n)$), o qual normalmente é adaptado continuamente de-

vido à variação no tempo de $h(n)$ e das estatísticas da entrada. Porém, o mesmo não é válido para o alto-falante, que geralmente apresenta fortes não-linearidades, especialmente quando suas dimensões são limitadas e um nível de potência elevado é desejado na saída [4], o que costuma ocorrer para equipamentos desse tipo. Sendo assim, a utilização de um modelo puramente linear impossibilita a correta predição do comportamento do alto-falante e, conseqüentemente, impede um cancelamento eficaz. Observe ainda que, assim como no caso da linearização de alto-falantes, um modelo eficientemente implementável é necessário, uma vez que a predição deve ser realizada em tempo real.

Deve-se notar que, conforme apontado por [11], uma alternativa interessante seria simplesmente aplicar técnicas de linearização de alto-falantes no dispositivo de telecomunicação, o que permitiria utilizar um modelo linear para estimar $y(n)$.

1.1.3 Simulação digital de dispositivos analógicos

Outra aplicação que vem recebendo atenção nos últimos anos é a simulação digital de dispositivos analógicos não-lineares usados em produções musicais, tais como pedais de efeitos, amplificadores de potência valvulados e alto-falantes [15, 16, 17, 18, 19]. Isso tem sido motivado pelas vantagens tipicamente apresentadas pelos dispositivos digitais em relação aos equipamentos analógicos modelados, como menor custo e maior flexibilidade. De fato, recentemente diversos produtos baseados nessa ideia têm sido desenvolvidos por fabricantes da indústria de equipamentos de áudio [15].

Nesta aplicação, os modelos podem ser utilizados tanto de modo *offline*, para o processamento de sinais previamente obtidos, quanto de modo *online*, em apresentações musicais. Para que sejam úteis neste último caso, é crucial que os modelos sejam eficientemente implementáveis, uma vez que normalmente é necessário computar amostras a uma taxa bastante elevada para se preservar a qualidade sonora da simulação. Além disso, o nível de tolerância para quaisquer atrasos introduzidos pelo processamento digital é baixíssimo nessa situação, pois um atraso alto pode atrapalhar o desempenho do usuário.

1.1.4 Cancelamento de vazamento de instrumentos musicais

Um problema considerado no início da pesquisa que resultou neste trabalho foi o de cancelamento de vazamento de instrumentos musicais (CVIM), o qual surge no contexto de produções musicais e guarda algumas semelhanças com o de CEA. Nesse caso, o objetivo é eliminar a interferência produzida pelo alto-falante ligado a cada instrumento nos sinais captados para os demais quando gravados simultaneamente em um estúdio. Na Figura 1.3, ilustra-se a configuração típica do problema. Observe que o microfone à esquerda capta o sinal do conjunto amplificador/alto-falante posicionado próximo a ele, $d_1(n)$, e também o sinal emitido pelo outro conjunto e filtrado pela resposta acústica da sala, $y_2(n)$. Assim, para eliminar essa interferência da gravação, um sinal de referência é extraído do conjunto amplificador/alto-falante à direita e fornecido a um modelo que deve estimar as amostras de $y_2(n)$.

Figura 1.3: Diagrama de blocos representando a configuração típica do problema de cancelamento de vazamento de instrumentos musicais.

No contexto do CVIM, em geral também é necessário que os modelos utilizados sejam não-lineares, uma vez que em gravações desta natureza frequentemente o conjunto amplificador de potência/alto-falante é ajustado de forma que distorções não-lineares sejam intencionalmente introduzidas.

1.1.5 Controle

Em aplicações de controle, em que se deseja que as trajetórias das saídas (variáveis controladas) minimizem uma determinada função de custo, modelos matemáticos são empregados de diversas formas na construção de controladores. No problema de controle preditivo, por exemplo, o controlador explora a capacidade preditiva de um modelo matemático em um procedimento de otimização para computar entradas (variáveis manipuladas) que causem o comportamento desejado. Na literatura, esta estratégia foi empregada em alguns trabalhos com o uso de modelos não-lineares, como em [20, 21].

Observe que, mesmo que a taxa de amostragem mínima necessária para a geração das predições possa ser relativamente baixa nesse caso, como costuma ocorrer no controle de plantas industriais, os sistemas modelados frequentemente possuem múltiplas saídas e múltiplas entradas, de modo que o modelo completo na realidade é constituído de diversos “submodelos”, o que por si só já implica em um aumento do custo computacional. Assim sendo, a eficiência computacional dos modelos obtidos também é um aspecto importante nesse contexto.

1.1.6 Outras aplicações

No problema de equalização de canais de comunicação, o objetivo é a eliminação de componentes não-lineares que são introduzidos por amplificadores de potência antes da transmissão através do canal, os quais deterioram o desempenho do sistema de comunicação. Para tanto, pode-se pré-processar a entrada de modo que o sistema resultante seja linear, de forma semelhante à linearização de alto-falantes, ou pode-se aplicar uma filtragem na saída de modo que os componentes não-lineares significativos sejam eliminados [1].

Outra potencial aplicação que surge em sistemas de comunicação é o cancelamento de eco elétrico, o qual é causado por um acoplamento entre componentes da rede de comunicação [1] e resulta em inconvenientes semelhantes aos que ocorrem no caso do eco acústico. Visto que não-linearidades estão muitas vezes presentes na cadeia de transmissão dos sinais, também se faz necessário o uso de modelos

não-lineares.

1.2 Justificativa

Conforme observado anteriormente, a escolha de um formalismo adequado é crítica para a obtenção de um modelo útil para fins práticos. Em aplicações de processamento digital de sinais, isso requer que o modelo seja capaz de representar com precisão as não-linearidades tipicamente encontradas nos sistemas de interesse e não demande um custo computacional excessivo para implementação e adaptação, estando este último requisito diretamente ligado à compacidade do modelo (a qual será tanto maior quanto for menor a complexidade de sua estrutura). Contudo, na literatura, com frequência um formalismo é escolhido sem que seja fornecida uma justificativa satisfatória sobre o atendimento às condições mencionadas.

Além do problema acima mencionado, a determinação de algumas características estruturais dos modelos usados como, por exemplo, a ordem de filtros de Volterra, frequentemente não é realizada de forma sistemática ou não é suficientemente justificada. Isso impede uma compreensão precisa do potencial e das limitações de cada formalismo, uma vez que tais características exercem grande influência no desempenho dos modelos.

Por fim, parecem não existir metodologias para a análise comparativa de diferentes formalismos e estruturas com base em considerações físicas e visando esclarecer suas capacidades em atender aos requisitos acima mencionados. Tais fatos, aliados à grande diversidade de formalismos, métodos e técnicas existentes para o modelamento de sistemas não-lineares, dificultam a escolha de uma abordagem adequada para a obtenção de um modelo de fato útil para uma determinada aplicação.

Para ilustrar de forma mais específica os problemas citados, a seguir enumeram-se alguns comentários sobre trabalhos que foram consultados. Os itens discutidos servirão para embasar os objetivos do presente trabalho, os quais serão descritos

na próxima seção.

- 1) Em [22], uma estrutura de Hammerstein polinomial generalizada, que possui capacidade de representação bastante limitada, é empregada para modelar um alto-falante. Porém, trata-se de uma escolha inadequada, conforme será mostrado em um estudo de caso no Capítulo 6. Já em [23, 24], a mesma estrutura é utilizada no modelamento de um limitador de áudio, embora o erro de modelamento mostrado graficamente em [24] seja significativamente alto em alguns instantes de tempo.
- 2) Os modelos baseados em séries de Volterra são usados frequentemente em aplicações de PDS, por se tratar de um formalismo que possui ampla capacidade de representação. Entretanto, em muitos casos empregam-se filtros de Volterra convencionais [2, 4, 13, 25, 26], os quais normalmente necessitam de um grande número de parâmetros para um modelamento de alta ordem. Por esse motivo, costuma-se utilizar estruturas de segunda ou, no máximo, terceira ordem, o que pode limitar o desempenho do modelo, e mesmo assim muitas vezes não soluciona satisfatoriamente o problema da alta complexidade estrutural. Assim, em geral a determinação da ordem do modelo não é feita de modo que se leve em conta características físicas do sistema modelado, a potência dos sinais de entrada tipicamente encontrados na prática e o nível de precisão desejado.
- 3) Com a finalidade de atenuar a elevada complexidade computacional associada ao uso de filtros de Volterra convencionais, diversas estruturas têm sido propostas como alternativas a eles [12, 27, 28, 29]. Basicamente, a ideia consiste em introduzir alguma aproximação visando reduzir a complexidade estrutural do filtro sem deteriorar significativamente seu desempenho na representação do sistema. Porém, apesar de se tratar de uma ideia promissora, há certas limitações à sua aplicabilidade, uma vez que a cada uma das aproximações propostas estão associadas condições que o sistema modelado deve satisfazer para que bons resultados sejam obtidos. Além disso, parece não haver um trabalho em que se propõe uma forma de avaliar o

desempenho de tais estruturas com base em considerações físicas a respeito de um determinado sistema. Note que, em [12, 27], embora seja discutido o potencial da aplicação das estruturas propostas para o modelamento de alto-falantes, por exemplo, nos experimentos realizados os núcleos de Volterra utilizados como referência foram estimados a partir de medidas, e portanto incorporam elementos de incerteza, fatores ambientais (como reverberação acústica) e, possivelmente, polarização.

1.3 Objetivos

Em linhas gerais, este trabalho tem por objetivo principal fornecer meios para que sejam esclarecidas as questões levantadas na seção anterior, relacionadas às capacidades de representação e às complexidades estruturais de formalismos não-lineares quando adotados para modelar um determinado sistema físico de interesse. Para tanto, propõe-se uma metodologia que consiste em fazer uso de um modelo físico como referência, o que possibilita a obtenção de conclusões sólidas baseadas em propriedades físicas deste sistema (contanto, evidentemente, que tal modelo represente adequadamente o sistema em questão). Adicionalmente, isto permite isolar a análise realizada de aspectos práticos inconvenientes como fatores ambientais (e.g., reverberação), ruído e incertezas de medição.

Como forma de se demonstrar o potencial dessa abordagem, apresenta-se um estudo de caso em que um modelo físico para alto-falantes, que se mostrou preciso no modelamento de um sistema real [30], é utilizado na avaliação do desempenho de alguns formalismos não-lineares estudados.

Mais especificamente, os principais pontos considerados neste estudo são como descritos a seguir.

- (i) Em primeiro lugar, estudam-se alguns formalismos não-lineares tipicamente empregados em aplicações de PDS, como filtros de Volterra, modelos NARX e modelos orientados a blocos.
- (ii) Estruturas alternativas de Volterra propostas recentemente na literatura

são descritas e analisadas com a finalidade de se buscar modelos capazes de oferecer um bom compromisso entre complexidade estrutural e precisão.

- (iii) Investigam-se alguns dos métodos dados na literatura para a derivação de núcleos de Volterra a partir de equações de espaço de estados, como aqueles descritos em [31, 32], com o intuito de se determinar uma forma eficiente de derivar os núcleos associados a um modelo físico de referência.
- (iv) Fazendo-se uso de um algoritmo eficiente proposto para o método de Carleman [33], o qual resultou da investigação feita no item (iii), expressões para os núcleos associados a um sistema físico são fornecidas, considerando-se os filtros dos conversores digital/analógico e analógico/digital tipicamente envolvidos em aplicações de processamento digital de sinais.
- (v) De posse dos núcleos associados a um modelo físico, obtidos com as expressões mencionadas no item anterior, mostra-se como se pode determinar uma ordem e uma extensão de memória suficientes para que um filtro de Volterra atinja o nível de precisão estabelecido. Com isso, tem-se uma estimativa do número de parâmetros exigido por um filtro de Volterra para o modelamento do sistema.
- (vi) Com o uso dos núcleos de referência calculados, mostra-se também como se pode avaliar quantitativamente o desempenho das estruturas de Volterra alternativas. Isso permite determinar as vantagens e desvantagens de cada uma delas em relação ao filtro de Volterra convencional composto pelos núcleos de referência.
- (vii) Ainda a partir de tais núcleos, demonstra-se também como se pode esclarecer se certas estruturas modulares simples existentes na literatura, como o modelo de Hammerstein e o modelo L-N, são adequadas para o modelamento do sistema de interesse. Isso é feito com base em expressões que caracterizam os núcleos de modelos de Volterra equivalentes a tais estruturas.

Por fim, é importante mencionar que o pressuposto de disponibilidade de um

modelo físico do sistema de interesse é razoável em muitas aplicações de engenharia, existindo diversos trabalhos em que são fornecidos ou utilizados modelos desta natureza [7, 8, 11, 30, 34, 35].

1.4 Contribuições

Como contribuições principais deste trabalho, podemos citar:

- Uma apresentação contendo os principais aspectos e resultados básicos envolvidos no uso de modelos de Volterra para a representação de sistemas não-lineares em aplicações de PDS.
- Uma revisão bibliográfica crítica de diversos formalismos e estruturas não-lineares, incluindo modelos baseados na série de Volterra, modelos modulares frequentemente utilizados e algumas estruturas de Volterra alternativas propostas recentemente na literatura [9, 27, 28, 29].
- A proposta de um algoritmo eficiente [33] para o método de bilinearização de Carleman [32, 36], o qual se mostrou eficaz para o cálculo dos núcleos de Volterra associados a equações de espaço de estados.
- A derivação de expressões para núcleos de referência do sistema de tempo discreto que resulta da conexão em série de um sistema físico de tempo contínuo e conversores digital/analógico e analógico/digital, com base nas equações de um modelo físico do sistema [33].
- A proposta de uma metodologia para a avaliação do desempenho de diferentes estruturas quando empregadas no modelamento de um sistema de interesse, com base nas informações fornecidas por um modelo físico. Essa avaliação leva em conta aspectos importantes do ponto de vista prático, como a capacidade de representação e a complexidade estrutural de cada estrutura, as características espectrais e de potência das entradas tipicamente encontradas na aplicação de interesse e o nível de precisão desejado.

Além daquelas mencionadas acima, contribuições adicionais incluem:

- Uma descrição concreta do uso do método da exponencial crescente [32] para o cálculo de funções de resposta em frequência generalizadas associadas a sistemas de equações de estado lineares-analíticos invariantes no tempo.
- O apontamento de problemas práticos que surgem quando os princípios dados em [31] são empregados para a obtenção numérica de núcleos associados a um conjunto de equações de espaço de estados.

1.5 Organização

Esta dissertação está organizada como descrito a seguir.

No Capítulo 2, apresenta-se um conteúdo introdutório a respeito do modelamento de sistemas não-lineares no contexto de aplicações de PDS, sendo estabelecidos conceitos básicos para o entendimento do restante do texto.

Uma revisão bibliográfica dos formalismos não-lineares estudados neste trabalho é feita no Capítulo 3. Em particular, discutem-se as principais características e os resultados mais importantes relacionados aos modelos baseados na série de Volterra, os quais constituem uma peça chave na metodologia proposta.

O Capítulo 4 descreve estruturas alternativas de Volterra propostas recentemente na literatura, as quais são motivadas pela busca de um compromisso entre eficiência (i.e., compacidade) e precisão da representação. As condições sob as quais tais estruturas apresentam bom desempenho são discutidas. Adicionalmente, faz-se uma comparação entre a complexidade estrutural apresentada por cada uma delas e aquela que caracteriza um filtro de Volterra convencional.

No Capítulo 5, apresenta-se uma estratégia para a derivação eficiente dos núcleos de Volterra associados a um modelo físico, a qual constitui o cerne da metodologia proposta. Alguns métodos de cálculo de núcleos a partir de equações de estado propostos na literatura são discutidos, com atenção especial ao método de bilinearização de Carleman, o qual foi escolhido para uso neste trabalho. Por

fim, um algoritmo eficiente que implementa tal método, proposto recentemente pelos autores, é descrito.

Um estudo de caso é conduzido no Capítulo 6, com o intuito de se demonstrar a utilidade da metodologia proposta na avaliação do desempenho de diferentes estruturas quando empregadas no modelamento de um sistema físico. Para tanto, utiliza-se um modelo físico para alto-falantes que é dado na literatura.

Finalmente, o Capítulo 7 encerra este documento, apresentando as conclusões obtidas e as perspectivas de trabalhos futuros.

Capítulo 2

Aspectos básicos sobre o modelamento de sistemas não-lineares

Antes de procedermos a uma descrição mais detalhada dos formalismos estudados neste trabalho, é necessário estabelecer alguns conceitos básicos, os quais serão fundamentais para a compreensão do restante do texto. Neste capítulo, uma visão geral sobre o modelamento de sistemas não-lineares é fornecida, introduzindo-se aspectos e definições importantes sobre o tema, especialmente aqueles mais relevantes para aplicações de processamento digital de sinais.

2.1 Conceitos básicos

2.1.1 Sistemas

Os sistemas considerados neste trabalho são, de modo geral, dinâmicos. Sendo assim, suas entradas e saídas são funções (sinais) dependentes do tempo que podem ser relacionadas através de equações diferenciais, em tempo contínuo, ou de diferenças, em tempo discreto. Assim, em termos matemáticos, um sistema pode ser visto como um operador, isto é, um mapeamento entre um conjunto de

funções de entrada e um conjunto de funções de saída [37], denotado por

$$\mathsf{T} : U \rightarrow Y \quad (2.1)$$

$$y(t) = \mathsf{T}[u(t)]. \quad (2.2)$$

Ou seja, T associa a cada entrada $u(t) \in U$ uma única saída, ou resposta, $y(t) \in Y$.

Como estamos nos restringindo a sistemas físicos, naturalmente, os operadores considerados satisfazem a condição de causalidade

$$\forall t \leq t_0, u_1(t) = u_2(t) \implies \forall t \leq t_0, y_1(t) = y_2(t) \quad (2.3)$$

onde $y_1(t) = \mathsf{T}[u_1(t)]$ e $y_2(t) = \mathsf{T}[u_2(t)]$. Em outras palavras, a saída do sistema em um dado instante de tempo t_0 pode depender apenas de valores da entrada relativos a instantes $t \leq t_0$.

Outro pressuposto admitido se refere à não-linearidade dos sistemas considerados. Isto é, os sistemas da classe estudada *não* satisfazem o princípio da superposição, segundo o qual

$$\mathsf{T}[u_i(t)] = y_i(t), \quad i = 1, \dots, n \quad (2.4)$$

implica em

$$\mathsf{T}[\alpha_1 u_1(t) + \dots + \alpha_n u_n(t)] = \alpha_1 y_1(t) + \dots + \alpha_n y_n(t). \quad (2.5)$$

Ou seja, a resposta de um sistema da classe estudada para uma combinação linear de entradas é diferente, em geral, da mesma combinação linear das respostas para tais entradas.

Já em relação à variação no tempo, ambas as situações são consideradas: alguns sistemas de interesse são invariantes no tempo, satisfazendo, para qualquer $\tau \in \mathbb{R}$, a condição

$$\mathsf{T}[u(t)] = y(t) \implies \mathsf{T}[u_\tau(t)] = y_\tau(t) \quad (2.6)$$

onde

$$u_\tau(t) = u(t + \tau) \quad \text{e} \quad y_\tau(t) = y(t + \tau) \quad (2.7)$$

enquanto, para outros, chamados de variantes no tempo, tal condição não se aplica. Em termos práticos, a variação no tempo de um sistema pode ocorrer, por exemplo, por desgaste ou devido à ação de algum mecanismo específico.

2.1.2 Modelos matemáticos

Para representar matematicamente um sistema como o de (2.2), um modelo deve explicar de forma quantitativa a relação entre a evolução temporal da entrada (excitação), $u(t) \in U$, e a trajetória da resposta (saída) produzida para tal entrada, $y(t) \in Y$. Assim, de maneira geral, um modelo matemático consiste em uma expressão (ou em um conjunto de expressões) através da(s) qual(is) é possível computar a resposta produzida por qualquer entrada do conjunto U considerado.

De modo abstrato, um modelo matemático $\mathcal{M}$ pode ser encarado formalmente como um operador que mapeia uma entrada $u(t) \in U$ em uma saída $y(t) \in Y$ através da expressão mencionada no parágrafo anterior, i.e.,

$$\mathcal{M} : U \rightarrow Y \quad (2.8)$$

$$y(t) = \mathcal{M}[u(t)]. \quad (2.9)$$

Observe que, diferentemente de (2.2), na relação acima o operador em questão foi denotado por uma letra maiúscula caligráfica, o que será adotado neste trabalho para distinguir modelos matemáticos de outros operadores que ocorrem no texto. Deve-se notar ainda que, devido à forma abstrata de (2.8)–(2.9), tal representação inclui tanto os casos em que a saída pode ser computada diretamente através de uma única expressão (ou seja, através de uma relação entrada/saída) como aqueles em que um sistema de equações diferenciais e algébricas constituem o modelo, situação em que a saída pode ser computada apenas indiretamente devido à necessidade de se integrarem as equações diferenciais (seja de forma analítica ou numérica).

Na realidade, como se espera que o modelo forneça apenas uma aproximação da saída real, a notação empregada em (2.9) não é apropriada. Denotando por $\hat{y}(t)$ a saída estimada pelo modelo em questão, uma formulação mais realista seria [9]

$$y(t) = \hat{y}(t) + \varepsilon(t) = \mathcal{M}[u(t)] + \varepsilon(t) \quad (2.10)$$

em que $y(t)$ é a saída do sistema real e o termo $\varepsilon(t)$, chamado de erro de predição ou resíduo, incorpora incertezas como erros de estimação, submodelamentos e ruído.

É importante ressaltar que, apesar da formalização abstrata apresentada acima, neste trabalho há interesse na forma concreta ou estrutura que cada modelo estudado assume. Como discutido anteriormente, esta estrutura está diretamente relacionada com o esforço computacional associado tanto à estimação quanto ao uso do modelo, aspectos que são de grande importância no contexto de aplicações de processamento digital de sinais.

Além do comentário feito do parágrafo anterior, outros três aspectos importantes para o restante do texto devem ser esclarecidos. Primeiro, como estamos interessados em modelos para implementação digital, na prática, os conjuntos U e Y contêm sequências de tempo discreto, sendo os modelos matemáticos definidos de acordo. Assim, quando estivermos nos referindo a modelos nesta forma, será adotada a convenção de representar os instantes de tempo por $n \in \mathbb{Z}$ (ou $k \in \mathbb{Z}$, em alguns casos), estando implícito nesse caso que

$$u(n) = u_c(nT_s) \quad \text{e} \quad y(n) = y_c(nT_s) \quad (2.11)$$

em que $u_c(t)$ e $y_c(t)$ são, respectivamente, a entrada e a saída em tempo contínuo, $u(n)$ e $y(n)$ são suas contrapartes de tempo discreto e T_s é o período de amostragem adotado.

Segundo, embora (2.9) e (2.10) contemplem a possibilidade de o modelo possuir uma memória infinita em relação à entrada, as expressões concretas de $\mathcal{M}$ empregadas na prática não podem, obviamente, possuir infinitos termos. Ao invés disso, os formalismos que exploram tal possibilidade (como os modelos

NARX, por exemplo, os quais serão apresentados na Seção 3.3) são definidos com o uso de alguma forma de recursão. Em outras palavras, eles são tais que a saída estimada, $\hat{y}(n)$, depende tanto de valores da entrada pertencentes a $\{u(k) \mid n - N_u + 1 \leq k \leq n\}$ quanto de valores da própria saída pertencentes a $\{y(k) \mid n - N_y + 1 \leq k < n\}$, para certas constantes inteiras N_u e N_y .

Terceiro e último, os modelos adiante estudados serão considerados, por simplicidade, em suas formas invariantes no tempo. Note que tal simplificação não traz prejuízo quanto aos objetivos do trabalho, uma vez que não impede a avaliação da capacidade de representação e da compacidade dos formalismos estudados, aspectos que permanecem válidos (e são de especial importância) quando os modelos são empregados de forma adaptativa em contextos não-estacionários, dada a necessidade de atualização contínua dos parâmetros ou coeficientes do modelo.

2.2 Uso de informações prévias no modelamento

Conforme mencionado no capítulo inicial, o modelamento matemático de um sistema físico é um problema fundamental para diversas aplicações práticas. Contudo, em muitas situações, tem-se pouco ou nenhum conhecimento a respeito do sistema de interesse. Nesse caso, para se obter o modelo deve-se recorrer a técnicas empíricas, as quais se utilizam de dados experimentais. Tais técnicas são o objeto central de estudo da área conhecida como identificação de sistemas [38, 39].

Quando disponíveis, informações *a priori* podem auxiliar significativamente na obtenção de um bom modelo. Por exemplo, um conhecimento profundo sobre o sistema estudado permite determinar uma estrutura para o modelo de forma a se obter um bom compromisso entre sua compacidade e sua capacidade de representá-lo precisamente. Desse modo, devido à grande importância desse aspecto, que impacta tanto o nível de dificuldade enfrentado no modelamento quanto a qualidade do modelo obtido, costuma-se classificar o modo como o modelamento é conduzido com relação o grau com que informações prévias são empregadas, conforme descrito a seguir.

Quando nenhum pressuposto sobre o sistema estudado é levado em conta, tem-se o que se chama de *modelamento caixa-preta*. Seguindo-se essa abordagem, todas as escolhas envolvidas no processo são realizadas com base apenas em medições experimentais realizadas e em critérios adotados. Sendo assim, obviamente, ela apresenta a vantagem de ser aplicável a sistemas bastante diferentes, sem requerer um estudo aprofundado de cada um deles. Por outro lado, os parâmetros ou coeficientes do modelo estimado em geral não têm uma interpretação física simples, o que é uma desvantagem em contextos em que se deseja analisar propriedades do sistema através de características do modelo obtido. Além disso, o modelamento caixa-preta normalmente está associado ao uso de formalismos de grande flexibilidade, os quais costumam exigir uma estrutura complexa para se obter uma representação precisa.

O outro extremo, i.e., a determinação completa de um modelo através do conhecimento prévio sobre o sistema, é denominado *modelamento caixa-branca*. Suas vantagens e desvantagens correspondem ao inverso do que foi mencionado para o modelamento caixa-preta. Isto é, os modelos caixa-branca são constituídos de parâmetros que possuem uma relação direta com propriedades físicas do sistema e são normalmente compactos, porém exigem um estudo aprofundado para sua formulação e restringem-se ao modelamento de algum tipo de dispositivo específico. Por fim, deve-se notar que alguns autores consideram como caixa branca tanto o caso em que os valores dos parâmetros do modelo são previamente conhecidos quanto aquele em que tais valores são estimados com base em medições [40].

Além destas duas formas, o modelamento pode ser feito seguindo-se uma estratégia mista, isto é, de modo que algum conhecimento sobre o sistema seja explorado em combinação com o uso de formalismos flexíveis para que se obtenha um modelo compacto e preciso em menor tempo do que o necessário para um modelamento caixa-branca. Tal estratégia, conhecida na literatura como *modelamento caixa-cinza*, constitui um tópico importante de pesquisa [40] que é bastante estudado atualmente [38].

2.3 Etapas típicas do processo de modelamento

Apesar da diversidade das metodologias propostas na literatura para o modelamento de sistemas não-lineares, tipicamente ele é realizado através de um processo constituído por quatro etapas principais:

- 1) *Escolha de um formalismo matemático*: nesta etapa, objetiva-se determinar um formalismo matemático adequado para a representação do sistema estudado. Trata-se da decisão mais importante envolvida no processo [9, 10] pois, além de impactar diretamente a qualidade do modelo obtido, determina que opções poderão ser consideradas, que técnicas poderão ser utilizadas e quais aspectos serão importantes nas etapas subsequentes.
- 2) *Definição da estrutura do modelo*: em seguida, é necessário definir uma topologia específica para o modelo. Em outras palavras, deve-se determinar sua forma concreta, considerando-se o conjunto de possibilidades oferecido pelo formalismo escolhido. Assim, para alguns formalismos, tal definição pode envolver diversos graus de liberdade – como na definição de uma função não-linear para representar a evolução de uma variável de estado em uma equação diferencial –, ao passo que para outros é necessário definir apenas algumas características – como no caso de filtros de Volterra, caracterizados completamente por uma ordem e uma extensão de memória. Quando não há conhecimento prévio disponível para se estabelecer uma estrutura adequada, pode-se lançar mão de algum método empírico, o que é conhecido na literatura como detecção de estrutura. Algumas técnicas de detecção de estrutura são discutidas, por exemplo, em [38, 39, 41, 42].
- 3) *Determinação dos parâmetros ou coeficientes do modelo*: uma vez definida uma estrutura concreta para o modelo, devem-se determinar os parâmetros ou coeficientes que a constituem, seja por conhecimento prévio do sistema modelado ou através de um método de estimação. Quando é preciso estimá-los, em geral, busca-se uma combinação deles que seja ótima segundo algum critério escolhido, o qual é avaliado com o uso de um conjunto de dados

experimentais (contendo entradas e saídas). Nesse caso, deve-se ter em mente que tanto eventuais submodelamentos quanto incertezas e erros de medição contidos nos dados afetarão as estimativas produzidas. Dentre os inúmeros trabalhos que discutem métodos de estimação de parâmetros, podem ser citados [9, 32, 37, 38, 39, 41, 43, 44, 45, 46].

- 4) *Validação do modelo*: nesta etapa, a finalidade é determinar se o modelo resultante é satisfatório para o contexto em que será aplicado ou se é necessário retornar a alguma etapa anterior [38, 39]. Em outras palavras, trata-se de um passo em que se decide se o modelo será aceito como produto final do processo de modelamento. Deve-se observar que, se o modelo é adaptativo (a adaptação de modelos será descrita sucintamente na Subseção 2.6.2), a validação deve considerar sua capacidade de representação e de rastreamento adequado das características que se deseja modelar, e não valores específicos de parâmetros. Sendo assim, nesse caso avalia-se também o desempenho do próprio algoritmo de adaptação empregado.

Além das etapas descritas acima, dependendo do contexto específico em que o modelamento é realizado, outra pode se necessária. Esta etapa adicional ocorre logo ao início e envolve o projeto e a realização de experimentos com o sistema de interesse, os quais são requeridos quando se faz uso de técnicas de identificação para a estimação *offline* de modelos matemáticos [38, 39]. No projeto, são feitas definições importantes para o experimento, como a característica espectral e a potência dos sinais de excitação usados, com o intuito de se extrair um conjunto de dados tão rico quanto possível em informações sobre o sistema.

2.4 Uso prático dos modelos

Embora representações matemáticas sejam exploradas de diversas maneiras em aplicações de processamento digital de sinais, neste estudo o foco reside sobre o uso de modelos de tempo discreto no cômputo preditivo de saídas de sistemas

físicos. Para tanto, independentemente do contexto específico, pressupõe-se que, no momento em que o modelo é empregado para o cálculo da amostra relativa a um determinado instante de tempo, há a disponibilidade de um conjunto finito de amostras da entrada medidas até tal instante. Em alguns casos, podem ser conhecidas também amostras da saída do sistema medidas em instantes anteriores.

Observe que a situação de uso típica considerada abrange tanto os casos em que as predições da saída são computadas de forma *online*, à medida em que novas amostras são disponibilizadas, quanto aqueles em que elas são calculadas de modo *offline* a partir de um conjunto completo e previamente conhecido de amostras da entrada. Como exemplos de uso *online*, podemos citar aqueles realizados nas aplicações de linearização de alto-falantes e de cancelamento de eco acústico, as quais são descritas nas Subseções 1.1.1 e 1.1.2, respectivamente; em relação ao uso *offline*, um exemplo de aplicação em que ele pode ocorrer é a simulação digital de dispositivos analógicos, apresentada na Subseção 1.1.3.

Quando o modelo matemático possui memória de duração infinita, que na prática sempre é realizada através de realimentação, o cômputo da amostra da saída para um determinado instante requer o uso de saídas de instantes anteriores. Neste caso, se não houver disponibilidade de saídas medidas, então saídas anteriores produzidas pelo próprio modelo devem ser empregadas para a realimentação, o que levanta a questão de estabilidade. Isto é, em tal situação, caso o modelo seja instável, a saída por ele produzida poderá ter magnitude ilimitada, o que compromete sua utilidade. Enquanto que para modelos lineares e invariantes no tempo a estabilidade é facilmente verificada observando-se os módulos de seus polos (no caso de modelos de tempo discreto), o mesmo não ocorre para modelos não-lineares, para os quais, em geral, a análise de estabilidade é uma tarefa bastante trabalhosa. De fato, a determinação de formalismos que ofereçam modelos não-lineares com realimentação suficientemente flexíveis e cuja estabilidade seja analisável em tempo razoável ainda é um problema em aberto, conforme apontado em [40].

Por outro lado, ainda considerando-se modelos com memória de duração infinita, o uso de saídas medidas nos termos de realimentação, quando elas são

disponíveis, pode ser desvantajoso devido a incertezas como ruído de medida e outras perturbações. Na presença destes elementos, as amostras computadas pelo modelo podem sofrer polarização, o que é especialmente problemático no caso de modelos não-lineares.

Um último comentário importante em relação ao uso típico dos modelos considerado neste trabalho é que, na maioria das vezes, as amostras computadas por um modelo são utilizadas de modo *online* em algum processamento subsequente. Ainda, muitas vezes, o modelo é continuamente estimado para que sejam rastreadas variações temporais em certos elementos envolvidos, como será descrito na Subseção 2.6.2. Por tais motivos é que os aspectos de eficiência computacional e estrutural dos modelos (os quais estão relacionados) são centrais.

2.5 Abordagens de modelamento

Conforme mencionado anteriormente, a escolha fundamental feita no processo de modelamento diz respeito ao formalismo matemático a ser empregado, de forma que deve ser feita criteriosamente e fazendo-se uso de todas as informações prévias disponíveis sobre o sistema a ser modelado. Na literatura, diversos formalismos não-lineares foram propostos, sendo possível agrupá-los em quatro abordagens principais: a não-paramétrica, a paramétrica, a modular e a conexionista [9]. No que segue, tais abordagens e algumas de suas características principais serão descritas.

2.5.1 Abordagem não-paramétrica

Na abordagem não-paramétrica, empregam-se formalismos que têm um caráter essencialmente geral. Isso pode constituir uma vantagem importante, principalmente quando se tem pouco ou nenhum conhecimento *a priori* a respeito do sistema sendo modelado, isto é, quando um modelamento caixa-preta é necessário. De fato, as escolhas envolvidas na etapa de definição da estrutura incluem, nessa abordagem, apenas algumas características do modelo, como, por exemplo,

seu grau de não-linearidade e a duração de sua memória.

Apesar desta adequação particular ao modelamento caixa-preta, os modelos baseados na série de Volterra, que são os mais comumente empregados dentre os não-paramétricos, podem ser derivados analiticamente a partir de um modelo físico do tipo caixa-branca. Nesse caso, como será visto adiante, a série de Volterra possui uma interpretação matemática muito importante como uma expansão funcional em série de potências da solução das equações diferenciais não-lineares associadas ao modelo físico em torno de um estado inicial de interesse.

A principal desvantagem de modelos não-lineares não-paramétricos, como é bem conhecido na literatura, é que eles costumam ser pouco compactos, especialmente quando o sistema de interesse é fortemente não-linear. Assim, em muitos casos, um imenso número de coeficientes se faz necessário para que um nível de precisão adequado seja atingido, o que constitui o principal obstáculo relacionado a seu uso prático. Além disso, modelos não-paramétricos são de difícil interpretabilidade, isto é, o estabelecimento de relações entre os elementos que compõem um modelo e o comportamento do sistema modelado não é uma tarefa fácil.

2.5.2 Abordagem paramétrica

Os modelos paramétricos baseiam-se em equações diferenciais ou de diferenças, possuindo, muitas vezes, alguma forma de realimentação. Por esse motivo, trata-se de uma abordagem que permite o uso de um número reduzido de parâmetros em comparação com os modelos não-paramétricos, de forma análoga à vantagem de compactidade que um modelo linear em espaço de estados apresenta em relação a uma representação através de resposta impulsiva.¹ Porém, ao contrário do que acontece com os modelos não-paramétricos, a estrutura paramétrica que deve ser estabelecida antes da estimação dos parâmetros, seja por conhecimento prévio sobre o sistema ou através da análise de um conjunto de dados disponível, envolve diversos graus de liberdade, que com frequência dão origem a um vasto conjunto combinatório de possibilidades.

¹Note que, na realidade, um modelo em equações de espaço de estados tem uma memória de duração infinita, de modo que estamos nos referindo a uma resposta impulsiva de duração finita que fornece apenas uma aproximação.

De fato, a compacidade oferecida por modelos paramétricos está associada a uma perda de generalidade, de forma que se deve definir uma estrutura específica e apropriada para representar o sistema de interesse, sob pena de se impossibilitar um bom modelamento do sistema, sejam quais forem os parâmetros escolhidos. Tipicamente, quanto mais compacto é um modelo, tanto maior é a necessidade de que sua estrutura seja coerente com a realidade para que ele produza bons resultados; tal compromisso é fundamental e deve ser levado em conta no modelamento. É importante notar que, em algumas situações pode-se até conseguir um bom desempenho em relação ao critério utilizado no procedimento de estimação dos parâmetros mesmo com uma estrutura inadequada, desde que ela seja suficientemente flexível para se ajustar aos dados. Todavia, isso não garante que o desempenho será bom também para qualquer outra entrada.

Outra desvantagem, que é consequência direta da presença de realimentação, é a possibilidade do modelo se comportar de forma instável na prática. Esse aspecto pode dificultar bastante o uso de um modelo para simulação, pois a tarefa de se analisar um modelo não-linear com realimentação para se certificar que ele não se comportará de forma instável em geral é bastante trabalhosa.

Exemplos de formalismos paramétricos incluem os modelos físicos, os quais são modelos caixa-branca constituídos por equações diferenciais, e modelos NARX (*nonlinear auto-regressive model with exogenous inputs*), dados por equações de diferenças não-lineares normalmente obtidas de forma caixa-preta. No próximo capítulo, tais formalismos serão tratados mais detalhadamente.

2.5.3 Abordagem modular

A abordagem modular consiste em construir modelos compostos pela interconexão em série ou em paralelo de blocos paramétricos ou não-paramétricos. Por exemplo, um bloco pode corresponder a um modelo linear ou uma não-linearidade sem memória, ambos comuns na prática. A definição de uma topologia específica de interconexão dos blocos pode ser feita buscando-se uma estrutura com caráter geral ou com base em informações (parciais ou completas) a respeito do comportamento ou estrutura interna do sistema modelado, isto é, de forma caixa-preta,

caixa-cinza ou até mesmo caixa-branca.

Um aspecto interessante desta abordagem é que a flexibilidade na definição de sua estrutura pode ser explorada para se obter um compromisso entre as propriedades de modelos paramétricos e de não-paramétricos. De fato, como será visto nos próximos capítulos, modelos modulares têm sido adotados com frequência para se obter representações precisas com um número reduzido de parâmetros.

Como exemplos de modelos modulares, temos o L-N, em que um bloco linear é sucedido por uma não-linearidade sem memória, o N-L (ou de Hammerstein), que possui uma configuração oposta à do L-N, e o modelo de Wiener, que consiste em uma não-linearidade sem memória multivariada precedida por um banco de filtros lineares paralelos. Estas e outras estruturas modulares serão descritas de forma mais completa no Capítulo 3.

2.5.4 Abordagem conexionista

Por fim, a abordagem conexionista compreende os modelos constituídos de redes neurais artificiais, as quais consistem em uma estrutura computacional em que elementos básicos de processamento, denominados neurônios, são organizados em camadas interconectadas. As redes neurais são conhecidas por sua ampla capacidade de aproximação de funções, conforme estabelecido pelo Teorema de Aproximação Universal [47], sendo portanto bastante utilizadas em problemas de regressão, inclusive quando o processo sob estudo tem natureza não-linear. Contudo, tratam-se de modelos que normalmente não são compactos e cuja estrutura é extremamente difícil de associar a aspectos do fenômeno físico de interesse [9].

Neste trabalho, como foi descrito no Capítulo 1, o interesse recai especialmente sobre modelos que apresentam uma relativa compacidade e/ou que são possíveis de se associar a aspectos físicos. Assim, devido às desvantagens mencionadas no parágrafo anterior, os modelos conexionistas parecem não se enquadrar nesta proposta, e não foram incluídos no escopo do estudo.

2.6 Aspectos sobre estimação dos parâmetros

Na etapa de determinação dos parâmetros ou coeficientes que compõem o modelo, frequentemente é necessário recorrer a algum método para estimá-los. Para tanto, emprega-se um procedimento de otimização segundo algum critério, como a minimização de uma função de custo.

Mais precisamente, seja $D \subset \mathbb{R}^d$ o espaço em que residem os possíveis vetores de parâmetros ou coeficientes, $\mathcal{M}_\theta$ o modelo composto pelo vetor de parâmetros $\theta \in D$ e Z o conjunto de dados disponível para estimação, contendo amostras de entradas e saídas medidas. Então, o problema de estimação de parâmetros pode ser formulado como a busca por um vetor $\hat{\theta} \in D$ que satisfaça

$$\hat{\theta} = \arg \min_{\theta \in D} J(\theta, \mathcal{M}_\theta, Z) \quad (2.12)$$

em que J é a função de custo escolhida para a estimação.

Em geral, o critério é escolhido de modo que a minimização de J corresponda à busca de um conjunto de parâmetros ótimo em um certo sentido estatístico. Admitindo que a saída do sistema que se deseja modelar, $y(n)$, e a entrada, $u(n)$, são produzidos por processos estocásticos conjuntamente estacionários no sentido amplo, uma escolha comum para J é o erro quadrático médio cometido pelo modelo, dado por

$$J(\theta, \mathcal{M}_\theta, Z) = E\{\varepsilon^2(n)\} = E\{[y(n) - \hat{y}(n)]^2\} \quad (2.13)$$

onde usou-se a relação (2.10), sendo que $\hat{y}(n)$ denota a saída estimada por $\mathcal{M}_\theta$. Supondo que Z contém N pares de amostras da entrada e da saída, i.e.,

$$Z = \{(u(n), y(n)) \mid 0 \leq n \leq N-1\} \quad (2.14)$$

e que o processo que produz $\varepsilon(n)$ é ergódico no sentido amplo, então a esperança matemática em (2.13) é substituída na prática pela média amostral,² resultando

²É importante observar que, na prática, condições iniciais adequadas devem ser estabelecidas para que tal estratégia produza estimativas precisas.

em

$$J(\theta, \mathcal{M}_\theta, Z) = \frac{1}{N} \sum_{n=0}^{N-1} \{y(n) - \hat{y}(n)\}^2. \quad (2.15)$$

Quando o modelo é linear nos parâmetros e (2.15) é adotada como a função de custo a ser minimizada, tem-se o importante caso em que a solução pode ser determinada analiticamente, através do estimador de mínimos quadrados. Tal estimador, o qual é fundamental na determinação de parâmetros de modelos matemáticos, será brevemente descrito na próxima subseção.

2.6.1 O estimador de mínimos quadrados

Suponha que, em termos concretos, o modelo possa ser escrito na forma

$$\hat{y}(n) = \varphi(n)^T \theta \quad (2.16)$$

onde os componentes de $\varphi(n)$ dependem (linearmente ou não-linearmente) de amostras da entrada e da saída³ medidas até os instantes n e $n - 1$, respectivamente. Então, diz-se que (2.16) é um modelo de regressão linear, e $\varphi(n)$ recebe o nome de vetor de regressores. Consequentemente, o erro de predição envolvido na equação (2.10) pode ser escrito para cada instante de tempo como

$$\varepsilon(n, \theta) = y(n) - \varphi(n)^T \theta. \quad (2.17)$$

Observe que a dependência de θ introduzida no erro de predição explicita o fato de que ele varia de acordo com o vetor de parâmetros escolhido para compor o modelo.

Sob tal condição, e escolhendo-se o critério (2.15) para a estimação, pode-se mostrar que o vetor θ^* ótimo corresponde à solução dada pelo clássico estimador

³Deve-se lembrar que, conforme discutido na Subseção 2.1.2, os modelos com memória de duração infinita são formulados, na prática, através da introdução de realimentação. Em outras palavras, as estimativas por eles fornecidas dependem de amostras de entradas e de saídas passadas.

de mínimos quadrados [38, 39]

$$\theta^* = \left[\frac{1}{N} \sum_{n=0}^{N-1} \varphi(n) \varphi^T(n) \right]^{-1} \frac{1}{N} \sum_{n=0}^{N-1} \varphi(n) y(n) \quad (2.18)$$

contanto que a inversa envolvida na equação (2.18) exista, o que basicamente depende de propriedades da entrada medida que é empregada na estimação [38, 39].

Assim, quando o modelo é tal que pode ser escrito na forma (2.16) e o erro quadrático médio é escolhido como critério de otimalidade para os parâmetros, tem-se o benefício de a função J possuir um mínimo global que pode ser computado através de algoritmos relativamente simples e com boas propriedades numéricas [41]. De fato, por tais motivos os estimadores da família de mínimos quadrados são muito empregados na estimação de parâmetros ou coeficientes tanto de modelos lineares quanto de não-lineares.

2.6.2 Adaptação de parâmetros

Em muitos problemas práticos no contexto de PDS, ocorrem variações temporais nas características de um sistema envolvido e/ou nas estatísticas das entradas a ele submetidas (as quais, em tal circunstância, são denominadas não-estacionárias). Nesse caso, se um modelo de tal sistema é utilizado para alguma finalidade, faz-se necessário empregar uma abordagem adaptativa, a qual basicamente consiste em reestimar os parâmetros ou coeficientes de forma *online*, à medida em que novas amostras de entradas e saídas são obtidas [45]. Desse modo, as variações inerentes à aplicação podem ser rastreadas pelo modelo (ou filtro) adaptativo para que seu desempenho não se degrade.

Assim como no caso de estimação *offline*, a adaptação é realizada buscando-se um vetor de parâmetros ótimo segundo um certo critério estabelecido. Contudo, no caso *online*, a busca deve ser realizada continuamente, de modo que os parâmetros sejam ajustados quando novas informações (amostras) são disponibilizadas, e não com base em um conjunto de dados completo como no caso *offline*. Para que esse ajuste iterativo seja viável em um contexto prático, é necessário que ele

não exija recursos computacionais em demasia, o que depende do algoritmo de adaptação empregado e também de características do modelo, como sua compacidade.

Na literatura, diversos algoritmos foram propostos para a implementação de filtros adaptativos. Porém, eles não serão descritos no presente trabalho por não se incluírem em seu escopo. Ao leitor interessado no assunto, recomenda-se como referência básica [45].

2.6.3 Aspectos estatísticos relacionados à estimação

Quando se emprega um método de estimação de parâmetros a partir de medições, o modelo resultante está sujeito a incorporar características das estatísticas dos sinais medidos e também de outros elementos que ocorrem na prática, como ruído de medição ou submodelamentos (i.e., porções das saídas medidas que compõem a resposta do sistema e não são explicadas pela estrutura adotada). Consideremos, por exemplo, o caso em que as saídas do sistema podem ser escritas na forma

$$y(n) = \varphi(n)^T \theta_0 + \varepsilon(n) \quad (2.19)$$

em que $\varphi(n)$ é o vetor de regressores definido pela estrutura escolhida para o modelamento, θ_0 é o vetor de parâmetros associados a tais regressores e $\varepsilon(n)$ é uma sequência de erro que pode representar tanto ruído de medida quanto o produto de eventuais submodelamentos. Quando há correlação entre $\varepsilon(n)$ e $\varphi(n)$, o que certamente será o caso para submodelamentos e pode ocorrer também na presença de ruído de medida, haverá um viés nas estimativas dadas por (2.18) [38, 39]. Ainda, se o suporte do espectro de potência da entrada for limitado a apenas uma determinada região de frequências, então os parâmetros resultantes podem estar relacionados a apenas algumas características dinâmicas do sistema (no caso de sistemas lineares, por exemplo, a alguns modos), as quais são efetivamente excitadas por tal entrada.

Naturalmente, esta sensibilidade das estimativas ao ruído e às estatísticas da entrada ocorre também no contexto adaptativo. Nesse caso, tal efeito é ainda

mais intenso, uma vez que, como a estimação dos parâmetros é feita continuamente à medida em que mais medições são obtidas, o modelo tende a se ajustar a características locais (em relação ao tempo) dos sinais envolvidos. No entanto, nos problemas atacados com uma abordagem adaptativa, isto não costuma ser algo indesejável, uma vez que o objetivo é que o filtro produza saídas que apresentem bom desempenho em relação ao critério escolhido para adaptação, mesmo que para isso as propriedades do modelo decorram também de características estatísticas das entradas, das saídas desejadas e até mesmo do ruído que afeta as medições.

Já em situações em que se deseja que o modelo obtido constitua uma representação “fiel”, i.e., de forma que os valores dos parâmetros (nesse caso, estimados *offline*) decorram direta e unicamente de propriedades físicas do sistema de interesse, os efeitos acima descritos são indesejáveis. Como exemplos, podemos citar aplicações de controle, aplicações nas quais se deseja fazer simulações e contextos nos quais o modelo é usado para a própria compreensão do sistema. Basicamente, trata-se de situações nas quais é importante que o modelo seja capaz de generalizar as relações inferidas a partir dos dados de estimação para quaisquer entradas, sem exibir dependências em relação às condições específicas em que tais dados foram obtidos. Note que, no caso adaptativo, este problema é contornado através do ajuste contínuo dos parâmetros, de modo que o modelo busca constantemente aprimorar seu desempenho em relação às propriedades “atuais” dos sinais de entrada e saída.

Quando se requer um modelo que represente o sistema com generalidade, como descrito no parágrafo anterior, então, em primeiro lugar, é necessário que sua estrutura seja estabelecida de forma adequada. Admitindo-se que esse é o caso, então, em segundo lugar, é necessário buscar meios para a estimação dos parâmetros da estrutura que impeçam ou reduzam a introdução de viés causado pelos elementos supracitados. Para tanto, diversas técnicas foram propostas na literatura, como o método de variáveis instrumentais e o estimador de mínimos quadrados estendido.

Apesar de serem fundamentais, métodos de estimação de parâmetros a partir

de medições não se incluem no escopo deste trabalho e, portanto, não serão aqui discutidos. Caso o leitor tenha interesse neste tópico, referências importantes que os discutem detalhadamente incluem [9, 32, 37, 38, 39, 41, 43, 45].

Capítulo 3

Representações não-lineares

Sistemas dinâmicos não-lineares têm sido objeto de intenso estudo desde o século XIX. Para representá-los, foram desenvolvidos vários formalismos matemáticos com propriedades diferentes. Alguns deles possuem caráter relativamente geral, enquanto outros possuem aplicabilidade mais restrita, sendo válidos somente para sistemas com características específicas. Tal diversidade é consequência da dificuldade de se elaborar uma representação que tenha ao mesmo tempo utilidade prática e capacidade de caracterizar uniformemente toda a gama de comportamentos complexos apresentados por sistemas não-lineares.

O modelamento de um sistema dinâmico não-linear, como foi visto, requer que se opte por um dentre tais formalismos ou representações. Para que a decisão tomada seja adequada, é imprescindível que se tenha uma compreensão das propriedades centrais de cada um, bem como suas vantagens e desvantagens em relação aos demais. Neste capítulo, objetiva-se apresentar algumas representações não-lineares usualmente empregadas em aplicações de processamento digital de sinais, expondo-se as características principais de cada uma delas.

O conteúdo que segue está organizado como descrito a seguir. Na Seção 3.1, apresentam-se os conceitos, as definições e os resultados mais importantes relacionados ao uso de modelos baseados na série de Volterra em aplicações de PDS, incluindo-se aspectos analíticos, como questões sobre a convergência da série, e aspectos relacionados à análise de sistemas não-lineares no domínio da frequência. Em seguida, a Seção 3.2 fornece uma breve descrição de modelos físicos, os quais constituem uma peça chave na metodologia proposta neste trabalho, e sua relação com modelos de Volterra. Embora os modelos NARX não sejam explorados neste

trabalho, por questões de completude tal formalismo é apresentado na Seção 3.3, discutindo-se também suas principais características e sua relação com modelos de Volterra. Por fim, os modelos modulares, cuja importância no contexto deste trabalho ficará evidente mais adiante, são apresentados de forma geral na Seção 3.4, sendo algumas estruturas específicas descritas com mais detalhes.

3.1 Modelos baseados na série de Volterra

Dentre as notáveis contribuições dadas pelo matemático italiano Vito Volterra, destaca-se a introdução do conceito de um funcional,¹ definido como uma função que depende dos valores de outra função sobre um intervalo contínuo [9]. Adicionalmente, Volterra desenvolveu uma expansão em série de potências para um funcional análoga à expansão em série de Taylor para funções contínuas multivariadas, o que veio a ser conhecido como série de Volterra.

Anos mais tarde, Norbert Wiener empregou este formalismo para a representação de operadores não-lineares associados a sistemas dinâmicos,² expandindo-os em séries funcionais de potências na entrada. Daí em diante, tal ideia passou a ser uma das mais exploradas no estudo de sistemas não-lineares, adquirindo grande importância tanto teórica quanto prática. Atualmente, o uso de expansões funcionais na caracterização de sistemas não-lineares é um conceito bem estabelecido na teoria de sistemas [48], sendo que, embora outras expansões tenham sido aplicadas com essa finalidade, o uso da série de Volterra é predominante.

Na apresentação da série de Volterra e de suas propriedades feita a seguir, por questões de simplicidade, consideraremos apenas operadores não-lineares cujas entradas e saídas são unidimensionais.

¹Na realidade, Volterra utilizava a expressão “função de uma linha”, sendo Hadamard quem empregou pela primeira vez o termo “funcional” [9].

²Muito embora, curiosamente, o trabalho de Volterra sequer foi citado por Wiener.

3.1.1 A série de Volterra

Definição 3.1 Seja $T : U \rightarrow Y$ um operador não-linear e $\{h_p\}_{p \in \mathbb{N}}$ uma sequência de funções limitadas $h_p : \mathbb{R} \times \mathbb{R}^p \rightarrow \mathbb{R}$. Então, dizemos que tais funções são núcleos de Volterra (do inglês *Volterra kernels*) que descrevem a relação entrada/saída de T se podemos escrevê-la através da série

$$y(t) = h_0(t) + \sum_{p=1}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_p(t, \tau_1, \dots, \tau_p) u(t - \tau_1) \dots u(t - \tau_p) d\tau_1 \dots d\tau_p \quad (3.1)$$

em que $u(t) \in U$, $y(t) \in Y$, $h_0(t) : \mathbb{R} \rightarrow \mathbb{R}$ é o componente de ordem zero e h_p é o núcleo de Volterra de ordem p .

Observe que (3.1) pode ser escrita na forma

$$y(t) = h_0(t) + \sum_{p=1}^{\infty} y_p(t) \quad (3.2)$$

onde a saída é decomposta em um somatório de componentes de ordens $p \in \mathbb{N}$ tais que, para $p > 0$, o componente $y_p(t)$ é dado por um operador de p -ésima ordem descrito pelo funcional

$$y_p(t) = H_p[t; u(\tau)] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_p(t, \tau_1, \dots, \tau_p) u(t - \tau_1) \dots u(t - \tau_p) d\tau_1 \dots d\tau_p \quad (3.3)$$

o qual é dito homogêneo de grau p , uma vez que satisfaz

$$H_p[t; \alpha u(\tau)] = \alpha^p H_p[t; u(\tau)] \quad (3.4)$$

sendo $\alpha \in \mathbb{R}$ uma constante. Dessa forma, (3.1) descreve o valor da saída y em um instante t através de uma série de funcionais na entrada u .

Note também que, para fins de análise teórica, sempre é possível eliminarmos o componente de ordem zero através de uma transformação $\tilde{y}(t) = y(t) - h_0(t)$, o que não implica em perda de generalidade. Assim, no restante do texto, de modo geral não incluiremos o núcleo de ordem zero na representação (como se costuma fazer na literatura [37, 49, 50]), exceto quando estritamente necessário.

O formalismo definido por (3.1) possui duas interpretações importantes. Em primeiro lugar, observa-se que ele pode ser visto como uma generalização da série de Taylor [37]. Como um funcional pode depender dos valores de uma função sobre todo um intervalo contínuo, então o termo de ordem p , no caso da série de Volterra, corresponde a uma integral em que o produto de p variáveis que assumem tais valores é ponderado pelo núcleo h_p , o qual possui papel análogo ao dos coeficientes de uma expansão de Taylor. No contexto do estudo de sistemas dinâmicos, como a série de Volterra é usada para representar uma relação dinâmica em função de variações temporais na entrada $u(t)$, ela é por vezes chamada informalmente de série de Taylor com memória [12].

Por outro lado, observe que a expressão (3.3) pode ser vista também como uma generalização da integral de convolução para outras ordens além da primeira (i.e., da convolução linear) [31, 49], a qual resulta de (3.3) quando $p = 1$. Assim, pode-se interpretar a série como uma extensão para o caso não-linear da representação de uma relação entrada/saída através da integral de convolução, o que requer que o somatório inclua, nesse caso, convoluções multidimensionais de todas as ordens na entrada. Por esse ponto de vista, portanto, os núcleos de Volterra correspondem a generalizações da resposta impulsiva.

Por fim, é importante observar que, evidentemente, condições de convergência devem ser estabelecidas para que (3.1) faça sentido. Esta questão será discutida posteriormente, na Subseção 3.1.8.

3.1.1.1 Unicidade da representação

A definição da série de Volterra na forma (3.1) apresenta um problema de ambiguidade, no sentido de que não garante a unicidade dos núcleos associados a um operador. Em outras palavras, através dela, um mesmo operador pode ser representado por infinitas séries com núcleos diferentes, o que se deve à equivalência de produtórios $u(t - \tau_1) \dots u(t - \tau_p)$ em que os índices τ_i são dados por permutações de uma sequência de valores $\tau_1^*, \dots, \tau_p^*$. Como exemplo, considere o

núcleo de segunda ordem

$$h_2(t, \tau_1, \tau_2) = \begin{cases} 1, & \tau_1 \leq \tau_2 \\ 0, & \text{caso contrário} \end{cases}. \quad (3.5)$$

Definindo

$$\tilde{h}_2(t, \tau_1, \tau_2) = \begin{cases} 1 - \alpha, & \tau_1 \leq \tau_2 \\ \alpha, & \tau_1 > \tau_2 \end{cases} \quad (3.6)$$

para uma constante $\alpha \in \mathbb{R}$ qualquer, temos que, para qualquer entrada $u(t)$

$$y_2(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{h}_2(t, \tau_1, \tau_2) u(t - \tau_1) u(t - \tau_2) d\tau_1 d\tau_2 \quad (3.7)$$

$$= \alpha \int_{-\infty}^{\infty} \int_{\tau_2}^{\infty} u(t - \tau_1) u(t - \tau_2) d\tau_1 d\tau_2 + (1 - \alpha) \int_{-\infty}^{\infty} \int_{-\infty}^{\tau_2} u(t - \tau_1) u(t - \tau_2) d\tau_1 d\tau_2 \quad (3.8)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\tau_2} u(t - \tau_1) u(t - \tau_2) d\tau_1 d\tau_2 + \alpha \int_{-\infty}^{\infty} \int_{\tau_2}^{\infty} u(t - \tau_1) u(t - \tau_2) d\tau_1 d\tau_2 - \alpha \int_{-\infty}^{\infty} \int_{-\infty}^{\tau_2} u(t - \tau_1) u(t - \tau_2) d\tau_1 d\tau_2. \quad (3.9)$$

Mas, devido à simetria de $u(t - \tau_1)u(t - \tau_2)$ nos argumentos τ_1 e τ_2 ,

$$\int_{-\infty}^{\infty} \int_{\tau_2}^{\infty} u(t - \tau_1) u(t - \tau_2) d\tau_1 d\tau_2 = \int_{-\infty}^{\infty} \int_{\tau_1}^{\infty} u(t - \tau_1) u(t - \tau_2) d\tau_2 d\tau_1 \quad (3.10)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\tau_2} u(t - \tau_1) u(t - \tau_2) d\tau_1 d\tau_2. \quad (3.11)$$

Logo,

$$y_2(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{h}_2(t, \tau_1, \tau_2) u(t - \tau_1) u(t - \tau_2) d\tau_1 d\tau_2 \quad (3.12)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\tau_2} u(t - \tau_1) u(t - \tau_2) d\tau_1 d\tau_2 \quad (3.13)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h_2(t, \tau_1, \tau_2) u(t - \tau_1) u(t - \tau_2) d\tau_1 d\tau_2 \quad (3.14)$$

muito embora $h_2 \neq \tilde{h}_2$.

Para eliminar essa ambiguidade, costuma-se adotar alguma restrição sobre

os núcleos que garanta a unicidade da representação. Duas formas comumente usadas para escrevê-los sem ambiguidade são a triangular e a simétrica [32]. Diz-se que um núcleo $h_p(t, \tau_1, \dots, \tau_p)$ está na forma triangular se, $\forall i, j \in \mathbb{N}^*$ tais que $i + j \leq p$, ele satisfaz a propriedade³

$$\tau_{i+j} < \tau_j \implies h_p(t, \tau_1, \dots, \tau_j, \dots, \tau_{i+j}, \dots, \tau_p) = 0. \quad (3.15)$$

Já a forma simétrica, como sugerido pelo nome, é aquela em que os valores de um núcleo sobre todas as permutações possíveis de uma sequência de índices são iguais. Em outras palavras, para quaisquer $t, \tau_1, \dots, \tau_p \in \mathbb{R}$, o núcleo h_p deve satisfazer

$$h_p(t, \tau_1, \dots, \tau_p) = h_p(t, \pi_1(\tau), \dots, \pi_p(\tau)) \quad (3.16)$$

em que $(\pi_1(\tau), \dots, \pi_p(\tau))$ é uma permutação qualquer do multi-índice $\tau = (\tau_1, \dots, \tau_p)$.

Observe que cada condição descrita acima induz uma partição em classes de equivalência do conjunto de todos os núcleos possíveis, estabelecendo como a forma canônica de cada classe um único núcleo (triangular ou simétrico) h_p . Quando se adota a representação triangular, h_p pode ser obtido a partir de um outro núcleo não-triangular equivalente, $\tilde{h}_p$, através da fórmula

$$h_p(t, \tau_1, \dots, \tau_p) = \begin{cases} \sum_{\pi} \tilde{h}_p(t, \pi_1(\tau), \dots, \pi_p(\tau)), & \tau_1 \leq \tau_2 \leq \dots \leq \tau_p \\ 0, & \text{caso contrário} \end{cases} \quad (3.17)$$

em que o somatório é sobre todas as permutações $\pi = (\pi_1(\tau), \dots, \pi_p(\tau))$ possíveis de $\tau = (\tau_1, \dots, \tau_p)$ que produzem multi-índices distintos. Note que, aplicando-se a fórmula ao exemplo (3.6), o resultado é precisamente igual a (3.5). Já no caso da representação simétrica, uma fórmula análoga seria

$$h_p(t, \tau_1, \dots, \tau_p) = \frac{1}{p!} \sum_{\pi} \tilde{h}_p(t, \pi_1(\tau), \dots, \pi_p(\tau)) \quad (3.18)$$

³Observe que esta não é a única forma de se definir a triangularidade dos núcleos. Na realidade, como observado em [32], existem $p!$ escolhas diferentes para o domínio triangular de um núcleo de ordem p (i.e., o domínio de seus índices sobre o qual ele pode assumir valores não-nulos), cada uma correspondendo a uma diferente permutação dos índices $\tau_1, \dots, \tau_p$.

onde, novamente, o somatório é sobre todas as permutações $\pi = (\pi_1(\tau), \dots, \pi_p(\tau))$ de $\tau = (\tau_1, \dots, \tau_p)$ que resultam em multi-índices distintos, sendo Π o número total de tais permutações. Como exemplo, aplicando esta fórmula ao núcleo dado por (3.6), teríamos como resultado

$$h_2(t, \tau_1, \tau_2) = \begin{cases} 1, & \tau_1 = \tau_2 \\ \frac{1}{2}, & \text{caso contrário} \end{cases}. \quad (3.19)$$

3.1.1.2 Causalidade e variação no tempo

No que se refere ao uso para a representação de sistemas físicos, a série na forma (3.1) é mais geral que o necessário uma vez que, como os índices τ_i podem assumir valores negativos, permite-se que a saída em um determinado instante dependa de valores da entrada de instantes posteriores. Assim, no restante do texto, nos restringiremos às representações de Volterra causais, isto é, cujos núcleos satisfazem a condição de causalidade

$$h_p(t, \tau_1, \dots, \tau_p) \neq 0 \implies \tau_i \geq 0, i = 1, \dots, p. \quad (3.20)$$

Admitindo-se que os núcleos de uma série de Volterra são causais, podemos escrevê-la na forma triangular, por exemplo, como

$$y(t) = \sum_{p=1}^{\infty} H_p[t; u(\tau), \tau \leq t] \quad (3.21)$$

$$H_p[t; u(\tau), \tau \leq t] = \int_0^{\infty} \int_0^{\tau_p} \dots \int_0^{\tau_2} h_p(t, \tau_1, \dots, \tau_p) \prod_{i=1}^p u(t - \tau_i) d\tau_1 \dots d\tau_{p-1} d\tau_p. \quad (3.22)$$

Já em relação ao aspecto de variação no tempo, a forma (3.1) também tem um caráter geral, uma vez que contempla a possibilidade de os núcleos serem variantes no tempo, como indicado pela dependência direta de t . Com frequência, a representação de Volterra é simplificada admitindo-se que os núcleos são invariantes no tempo, o que ocorrerá muitas vezes ao longo deste texto. Nesse

caso, elimina-se tal dependência do instante de tempo t , de forma que se pode escrever (3.21)–(3.22), por exemplo, como

$$y(t) = \sum_{p=1}^{\infty} H_p[u(\tau), \tau \leq t] \quad (3.23)$$

$$H_p[u(\tau), \tau \leq t] = \int_0^{\infty} \int_0^{\tau_p} \dots \int_0^{\tau_2} h_p(\tau_1, \dots, \tau_p) \prod_{i=1}^p u(t - \tau_i) d\tau_1 \dots d\tau_{p-1} d\tau_p \quad (3.24)$$

De fato, pode-se mostrar que o operador definido por (3.23)–(3.24) satisfaz a condição (2.6).

3.1.1.3 Estabilidade

Como a representação de Volterra dada pela definição (3.1) consiste em uma relação entrada/saída, a caracterização de sua estabilidade é feita no sentido BIBO (*bounded input, bounded output*). Diz-se que um sistema é estável no sentido BIBO se e somente se ele é tal que, ao ser excitado por entradas limitadas, suas saídas também são limitadas.

Uma condição necessária para que uma série de Volterra seja estável no sentido BIBO é que cada operador de ordem p que a compõe o seja. Isto é, os componentes $y_p(t)$ de sua saída devem ser limitados quando a entrada é limitada. Contudo, tal condição não é suficiente, o que se deve ao fato de que, mesmo quando cada operador de ordem p é estável no sentido BIBO, um somatório infinito de seus valores pode resultar em um valor não limitado. Esta questão está relacionada ao aspecto de convergência da série, o qual será discutido na Subseção 3.1.8.

Para a estabilidade BIBO do operador de ordem p , uma condição suficiente é que seus núcleos sejam absolutamente integráveis, i.e.,⁴

$$\|h_p\|_1 = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} |h_p(\tau_1, \dots, \tau_p)| d\tau_1 \dots d\tau_p < \infty. \quad (3.25)$$

Claramente, quando (3.25) é satisfeita e a entrada $u(t)$ é limitada, ou seja, $\|u(t)\|_{\infty} = M < \infty$, então o componente de ordem $y_p(t)$ também o será, dado

⁴Mais formalmente, a expressão $x < \infty$ pode ser escrita como: $\exists K \in \mathbb{R}$ tal que $x < K$.

que [1, 37]

$$|y_p(t)| = \left| \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_p(\tau_1, \dots, \tau_p) u(t - \tau_1) \dots u(t - \tau_p) d\tau_1 \dots d\tau_p \right| \quad (3.26)$$

$$\leq \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} |h_p(\tau_1, \dots, \tau_p)| |u(t - \tau_1) \dots u(t - \tau_p)| d\tau_1 \dots d\tau_p \quad (3.27)$$

$$\leq M^p \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} |h_p(\tau_1, \dots, \tau_p)| d\tau_1 \dots d\tau_p \quad (3.28)$$

$$= M^p \|h_p\|_1 < \infty. \quad (3.29)$$

É importante observar que, conforme é mencionado em [1, 37], (3.25) é uma condição suficiente porém não necessária para que um operador de ordem $p > 1$ seja estável. Por exemplo, em [37] demonstra-se que um operador de Volterra de segunda ordem invariante no tempo cujo núcleo é dado por

$$h_2(\tau_1, \tau_2) = \sum_{n=1}^{\infty} \frac{1}{(2n)^2} \delta(\tau_1 - 2n) \left[- \sum_{m=0}^{n-1} \delta(\tau_2 - m) + \sum_{m=n}^{2n-1} \delta(\tau_2 - m) \right] \quad (3.30)$$

em que $\delta(t)$ é a função delta de Kronecker,⁵ produz apenas saídas limitadas para quaisquer entradas limitadas, muito embora [37]

$$\|h_2(\tau_1, \tau_2)\|_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |h_2(\tau_1, \tau_2)| d\tau_1 d\tau_2 \quad (3.31)$$

$$= \sum_{n=1}^{\infty} (2n) \frac{1}{(2n)^2} \quad (3.32)$$

$$= \frac{1}{2} \sum_{n=1}^{\infty} \frac{1}{n} = \infty. \quad (3.33)$$

Até onde vai o conhecimento dos autores, condições que sejam simultaneamente suficientes e necessárias não são conhecidas, exceto para classes particulares de núcleos como os separáveis [1].

⁵Deve-se observar que, ao se considerar um núcleo nessa forma, estamos implicitamente admitindo uma versão estendida da Definição 3.1 em que os núcleos podem ser funções generalizadas.

3.1.2 Cálculo da série para sistemas bilineares

Sistemas bilineares constituem uma classe particularmente importante de sistemas não-lineares, tendo recebido atenção especial em diversos trabalhos [43, 48, 51]. Tal importância pode ser atribuída ao fato de que sistemas bilineares são tratáveis analiticamente, existindo uma teoria bem desenvolvida que os descreve, e são úteis para a abordagem de problemas práticos por serem capazes, em princípio, de aproximar qualquer relação entrada/saída em intervalos de tempo finitos [51].

Assim sendo, a classe dos sistemas bilineares foi a primeira para a qual expressões gerais dos núcleos de Volterra foram derivadas [32]. No que segue, mostraremos como tais expressões podem ser obtidas, uma vez que elas serão importantes para o método de cálculo dos núcleos discutido no Capítulo 5. Além disso, tal derivação será interessante também para ilustrar de forma concreta a relação entre as equações de espaço de estados que definem um sistema dinâmico não-linear – nesse caso um sistema bilinear – e sua representação em série de Volterra.

A derivação das expressões para os núcleos pode ser feita através de uma técnica de re-substituição [32] ou através da fórmula de Peano-Baker [36]. A seguir, utilizaremos a primeira abordagem para calcular os núcleos de um sistema bilinear. Por simplicidade, nos concentraremos no caso em que a entrada e a saída são escalares.

Consideremos um sistema bilinear representado pelas equações de espaço de estados

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{x}(t)u(t) + \mathbf{b}(t)u(t) \quad (3.34)$$

$$y(t) = \mathbf{c}^T(t)\mathbf{x}(t) \quad (3.35)$$

onde $\mathbf{x}(0) = \mathbf{x}_0$, $\mathbf{A}(t)$ e $\mathbf{D}(t)$ são matrizes $n \times n$, $\mathbf{c}(t)$, $\mathbf{x}(t)$ e $\mathbf{b}(t)$ são vetores-coluna n -dimensionais. Para que exista uma solução garantidamente única para (3.34)–(3.35) em um intervalo finito de tempo $[0, T]$, vamos admitir que neste intervalo $\mathbf{A}(t)$, $\mathbf{b}(t)$, $\mathbf{c}(t)$ e $\mathbf{D}(t)$ são contínuos e a entrada $u(t)$ é limitada (ou seja, $\forall t, |u(t)| < \infty$) e contínua por trechos [32]. Agora, suponha que a equação

diferencial que descreve a trajetória não-forçada do sistema, $\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t)$, com $\mathbf{x}(0) = \mathbf{x}_0$, admita como solução fundamental a função $\Phi(t, 0)$. Pode-se mostrar que tal função, conhecida como *matriz de transição*, é dada pela série de Peano-Baker [32]

$$\Phi(t, t_0) = \mathbf{I} + \int_{t_0}^t \mathbf{A}(\tau_1) d\tau_1 + \int_{t_0}^t \mathbf{A}(\tau_1) \int_{t_0}^{\tau_1} \mathbf{A}(\tau_2) d\tau_2 d\tau_1 + \dots \quad (3.36)$$

onde $\mathbf{I}$ é a matriz identidade. Em essência, trata-se de uma generalização da solução para o caso em que $\mathbf{A}$ é uma constante, o que pode ser constatado na expressão acima, que nesse caso se reduz a

$$\Phi(t, t_0) = e^{\mathbf{A}(t-t_0)}. \quad (3.37)$$

A matriz de transição $\Phi(t, t_0)$ possui ainda duas propriedades bastante importantes: ela satisfaz

$$\Phi(t, \tau_1)\Phi(\tau_1, \tau_2) = \Phi(t, \tau_2) \quad (3.38)$$

e admite uma inversa para quaisquer t e t_0 , a qual é dada por $\Phi^{-1}(t, t_0) = \Phi(t_0, t)$.

As propriedades acima podem ser usadas para obter a série de Volterra do sistema bilinear da seguinte maneira. Definindo-se $\mathbf{z}(t) = \Phi^{-1}(t, 0)\mathbf{x}(t)$, podemos reescrever as equações de espaço de estados na forma

$$\dot{\mathbf{z}}(t) = \tilde{\mathbf{D}}(t)\mathbf{z}(t)u(t) + \tilde{\mathbf{b}}(t)u(t) \quad (3.39)$$

$$y(t) = \tilde{\mathbf{c}}(t)\mathbf{z}(t) \quad (3.40)$$

em que

$$\tilde{\mathbf{b}}(t) = \Phi^{-1}(t, 0)\mathbf{b}(t) \quad (3.41)$$

$$\tilde{\mathbf{D}}(t) = \Phi^{-1}(t, 0)\mathbf{D}(t)\Phi(t, 0) \quad (3.42)$$

$$\tilde{\mathbf{c}}(t) = \mathbf{c}^T(t)\Phi(t, 0) \quad (3.43)$$

com $t \geq 0$, $\mathbf{z}(0) = \mathbf{z}_0$.

Agora, supondo que o sistema bilinear em questão é causal e que $\forall t < 0$,

$u(t) = 0$, integrando ambos os lados de (3.39) obtemos

$$\mathbf{z}(t) = \mathbf{z}_0 + \int_0^t \tilde{\mathbf{D}}(\tau_1) \mathbf{z}(\tau_1) u(\tau_1) d\tau_1 + \int_0^t \tilde{\mathbf{b}}(\tau_1) u(\tau_1) d\tau_1. \quad (3.44)$$

Substituindo $\mathbf{z}(\tau_1)$ no lado direito de (3.44) por uma expressão com a mesma forma de (3.44), porém usando outra variável de integração τ_2 , resulta

$$\begin{aligned} \mathbf{z}(t) = \mathbf{z}_0 &+ \int_0^t \tilde{\mathbf{D}}(\tau_1) \mathbf{z}_0 u(\tau_1) d\tau_1 + \int_0^t \int_0^{\tau_1} \tilde{\mathbf{D}}(\tau_1) \tilde{\mathbf{D}}(\tau_2) \mathbf{z}(\tau_2) u(\tau_1) u(\tau_2) d\tau_2 d\tau_1 \\ &+ \int_0^t \int_0^{\tau_1} \tilde{\mathbf{D}}(\tau_1) \tilde{\mathbf{b}}(\tau_2) u(\tau_1) u(\tau_2) d\tau_2 d\tau_1 + \int_0^t \tilde{\mathbf{b}}(\tau_1) u(\tau_1) d\tau_1. \end{aligned} \quad (3.45)$$

Empregando esse método de substituição $P - 1$ vezes, temos

$$\begin{aligned} \mathbf{z}(t) = \mathbf{z}_0 &+ \sum_{p=1}^P \int_0^t \int_0^{\tau_1} \dots \int_0^{\tau_{p-1}} \tilde{\mathbf{D}}(\tau_1) \dots \tilde{\mathbf{D}}(\tau_p) \mathbf{z}_0 u(\tau_1) \dots u(\tau_p) d\tau_p \dots d\tau_1 \\ &+ \sum_{p=1}^P \int_0^t \int_0^{\tau_1} \dots \int_0^{\tau_{p-1}} \tilde{\mathbf{D}}(\tau_1) \dots \tilde{\mathbf{D}}(\tau_{p-1}) \tilde{\mathbf{b}}(\tau_p) u(\tau_1) \dots u(\tau_p) d\tau_p \dots d\tau_1 \\ &+ \int_0^t \int_0^{\tau_1} \dots \int_0^{\tau_{P-1}} \tilde{\mathbf{D}}(\tau_1) \dots \tilde{\mathbf{D}}(\tau_P) \mathbf{z}(\tau_P) u(\tau_1) \dots u(\tau_P) d\tau_P \dots d\tau_1. \end{aligned} \quad (3.46)$$

Fazendo $P \rightarrow \infty$, o último termo de (3.46) tende a zero uniformemente em qualquer intervalo de tempo finito [32]. Assim, reescrevendo (3.46) em função das variáveis originais e utilizando (3.35), obtemos a série uniformemente convergente [32, 36]

$$y(t) = \mathbf{c}^T(t) \Phi(t, 0) \mathbf{x}_0 + \sum_{p=1}^P \int_0^t \int_0^{\tau_1} \dots \int_0^{\tau_{p-1}} \tilde{h}_p(t, \tau_1, \dots, \tau_p) u(\tau_1) \dots u(\tau_p) d\tau_p \dots d\tau_1 \quad (3.47)$$

onde

$$\begin{aligned} \tilde{h}_p(t, \tau_1, \dots, \tau_p) &= \mathbf{c}^T(t) \Phi(t, \tau_1) \mathbf{D}(\tau_1) \Phi(\tau_1, \tau_2) \mathbf{D}(\tau_2) \dots \mathbf{D}(\tau_p) \Phi(\tau_p, 0) \mathbf{x}_0 \\ &+ \mathbf{c}^T(t) \Phi(t, \tau_1) \mathbf{D}(\tau_1) \Phi(\tau_1, \tau_2) \mathbf{D}(\tau_2) \dots \mathbf{D}(\tau_{p-1}) \Phi(\tau_{p-1}, \tau_p) \mathbf{b}(\tau_p). \end{aligned} \quad (3.48)$$

Agora, definindo

$$h_p(t, \tau_1, \dots, \tau_p) \triangleq \tilde{h}_p(t, t - \tau_1, \dots, t - \tau_p) \quad (3.49)$$

temos que

$$h_p(t, t - \tau_1, \dots, t - \tau_p) = \tilde{h}_p(t, \tau_1, \dots, \tau_p) \quad (3.50)$$

e, portanto, a série (3.47) pode ser reescrita na forma

$$y(t) = \mathbf{c}^T(t) \Phi(t, 0) x_0 + \sum_{p=1}^P \int_0^t \int_0^{\tau_1} \dots \int_0^{\tau_{p-1}} h_p(t, t - \tau_1, \dots, t - \tau_p) u(\tau_1) \dots u(\tau_p) d\tau_p \dots d\tau_1 \quad (3.51)$$

$$= \mathbf{c}^T(t) \Phi(t, 0) \mathbf{x}_0 + \sum_{p=1}^P \int_0^t \int_{\tau_1}^t \dots \int_{\tau_{p-1}}^t h_p(t, \tau_1, \dots, \tau_p) u(t - \tau_1) \dots u(t - \tau_p) d\tau_p \dots d\tau_1 \quad (3.52)$$

$$= \mathbf{c}^T(t) \Phi(t, 0) \mathbf{x}_0 + \sum_{p=1}^P \int_0^t \int_0^{\tau_p} \dots \int_0^{\tau_2} h_p(t, \tau_1, \dots, \tau_p) u(t - \tau_1) \dots u(t - \tau_p) d\tau_1 \dots d\tau_p \quad (3.53)$$

que corresponde a uma versão da série de Volterra causal (3.21)–(3.22) (com núcleos triangulares) em que a memória considerada leva em conta apenas contribuições da entrada a partir do instante $t_0 = 0$, devido à suposição de que $\forall t < 0, u(t) = 0$. Substituindo (3.48) em (3.49), vê-se que os núcleos de tal série são dados por

$$\begin{aligned} h_p(t, \tau_1, \dots, \tau_p) = & \mathbf{c}^T(t) \Phi(t, t - \tau_1) \mathbf{D}(t - \tau_1) \Phi(t - \tau_1, t - \tau_2) \mathbf{D}(t - \tau_2) \dots \mathbf{D}(t - \tau_p) \Phi(t - \tau_p, 0) \mathbf{x}_0 \\ & + \mathbf{c}^T(t) \Phi(t, t - \tau_1) \mathbf{D}(t - \tau_1) \Phi(t - \tau_1, t - \tau_2) \mathbf{D}(t - \tau_2) \dots \mathbf{D}(t - \tau_{p-1}) \\ & \times \Phi(t - \tau_{p-1}, t - \tau_p) \mathbf{b}(t - \tau_p). \end{aligned} \quad (3.54)$$

A equação (3.54) fornece a expressão geral dos núcleos para um sistema bilinear qualquer, variante no tempo ou não. Para um sistema invariante no tempo, podemos reescrevê-la mais convenientemente procedendo da seguinte forma. Usando a identidade (3.37) e observando que $\mathbf{b}$, $\mathbf{c}$ e $\mathbf{D}$ não dependem do instante de tempo,

obtemos

$$h_p(t, \tau_1, \dots, \tau_p) = \mathbf{c}^T e^{\mathbf{A}\tau_1} \mathbf{D} e^{\mathbf{A}(\tau_2 - \tau_1)} \mathbf{D} \dots \mathbf{D} e^{\mathbf{A}(\tau_p - \tau_{p-1})} [\mathbf{D} e^{\mathbf{A}(t - \tau_p)} \mathbf{x}_0 + \mathbf{b}] \quad (3.55)$$

Ainda, se além de tal característica o sistema em questão for tal que $\mathbf{x}_0 = 0$, então é possível eliminar o primeiro termo que aparece entre colchetes na expressão (3.55) e, conseqüentemente, eliminar também o argumento extra t . Como resultado, temos que um sistema bilinear invariante no tempo com estado inicial nulo pode ser caracterizado por núcleos triangulares na forma

$$h_p(\tau_1, \tau_2, \dots, \tau_p) = \mathbf{c}^T e^{\mathbf{A}\tau_1} \mathbf{D} e^{\mathbf{A}(\tau_2 - \tau_1)} \mathbf{D} \dots \mathbf{D} e^{\mathbf{A}(\tau_p - \tau_{p-1})} \mathbf{b}. \quad (3.56)$$

3.1.3 Análise no domínio da frequência

Como foi visto, no domínio do tempo os núcleos de Volterra possuem uma interpretação importante do ponto de vista matemático como generalizações da resposta impulsiva. Logo, como é de se esperar, quando levados ao domínio da frequência os núcleos podem ser vistos como generalizações da função de resposta em frequência associada a um sistema. Assim como no caso linear, tais funções, denominadas funções de resposta em frequência generalizadas, oferecem informações importantes sobre o comportamento do sistema no que se refere ao conteúdo espectral das saídas por ele produzidas.

Definição 3.2 Sejam $\{h_p\}_{p \in \mathbb{N}^*}$ os núcleos que compõem a representação em série de Volterra de um operador não-linear causal e invariante no tempo $\mathbf{T} : U \rightarrow Y$. Então, analogamente ao caso linear, a função de resposta em frequência generalizada (FRFG) de ordem p de $\mathbf{T}$ é definida como a transformada de Fourier de h_p , a qual é dada por

$$H_p(j\Omega_1, \dots, j\Omega_p) = \int_0^\infty \dots \int_0^\infty h_p(\tau_1, \dots, \tau_p) e^{-j(\Omega_1\tau_1 + \dots + \Omega_p\tau_p)} d\tau_1 \dots d\tau_p. \quad (3.57)$$

De forma dual a (3.57), o núcleo de ordem p pode ser obtido a partir de H_p

através da transformada inversa

$$h_p(\tau_1, \dots, \tau_p) = \frac{1}{(2\pi)^p} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} H_p(j\Omega_1, \dots, j\Omega_p) e^{j(\Omega_1\tau_1 + \dots + \Omega_p\tau_p)} d\Omega_1 \dots d\Omega_p. \quad (3.58)$$

Uma condição suficiente para que a FRFG do núcleo de ordem p exista, o que requer que a integral de (3.57) seja convergente, pode ser obtida estendendo-se diretamente aquela empregada no caso linear. Isso é, tal condição corresponde precisamente à mesma estabelecida para a estabilidade do operador de ordem p , dada por (3.25). Quando ela é satisfeita, temos

$$\begin{aligned} |H_p(j\Omega_1, \dots, \Omega_p)| &= \left| \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_p(\tau_1, \dots, \tau_p) e^{-j(\Omega_1\tau_1 + \dots + \Omega_p\tau_p)} d\tau_1 \dots d\tau_p \right| \\ &\leq \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} |h_p(\tau_1, \dots, \tau_p)| |e^{-j(\Omega_1\tau_1 + \dots + \Omega_p\tau_p)}| d\tau_1 \dots d\tau_p \\ &\leq \|h_p\|_1 < \infty \end{aligned} \quad (3.59)$$

o que garante a existência da FRFG de ordem p .

Com base na Definição 3.2, pode-se ver que, em decorrência da ambiguidade inerente à definição dos núcleos discutida na Subseção 3.1.1.1, as FRFGs que caracterizam um sistema também não são únicas. No entanto, impondo-se restrições sobre a forma que uma FRFG pode assumir, é possível garantir sua unicidade, assim como foi feito em relação aos núcleos. Em particular, é fácil ver a partir de (3.57) que a FRFG associada a um núcleo simétrico em $\tau_1, \dots, \tau_p$, o qual é o único sob a restrição de simetria, é também simétrica em $j\Omega_1, \dots, j\Omega_p$ e única [52]. Por outro lado, a transformada de um núcleo triangular, embora única, não possui uma característica que permita distingui-la facilmente de outras transformadas assimétricas por simples inspeção visual [32].

Em geral, costuma-se admitir que as FRFGs possuem forma simétrica, o que normalmente facilita a análise. Esta suposição será adotada também neste trabalho, exceto quando indicado o contrário. Note que não há perda de generalidade uma vez que, dada uma FRFG de ordem p qualquer, $\tilde{H}_p$, pode-se obter sua equi-

valente simétrica, H_p , pela seguinte fórmula análoga à de (3.18) [52]

$$H_p(j\Omega_1, \dots, j\Omega_p) = \frac{1}{\Pi} \sum_{\pi} \tilde{H}_p(j\pi_1(\Omega), \dots, j\pi_p(\Omega)) \quad (3.60)$$

em que o somatório é sobre todas as Π possíveis permutações $\pi = (\pi_1(\Omega), \dots, \pi_p(\Omega))$ de $\Omega = (\Omega_1, \dots, \Omega_p)$ que produzem multi-índices distintos. Por fim, deve-se enfatizar que, independentemente da forma particular admitida para o núcleo, todos os conceitos e resultados discutidos a seguir permanecem válidos.

Com o uso das FRFGs definidas em (3.57), a relação entrada/saída de $\mathbf{T}$ pode ser escrita no domínio da frequência através de

$$Y(j\Omega) = \sum_{p=1}^{\infty} Y_p(j\Omega) \quad (3.61)$$

$$Y_p(j\Omega) = \mathbf{C} \left[\tilde{Y}_p(j\Omega_1, \dots, j\Omega_p) \right] \quad (3.62)$$

em que $\tilde{Y}_p$ é dada por⁶

$$\tilde{Y}_p(j\Omega_1, \dots, j\Omega_p) = H_p(j\Omega_1, \dots, j\Omega_p) U(j\Omega_1) \dots U(j\Omega_p) \quad (3.63)$$

e $\mathbf{C}$ é o operador de contração definido como [9]

$$\begin{aligned} \mathbf{C} \left[\tilde{Y}_p(j\Omega_1, \dots, j\Omega_p) \right] (j\Omega) &= \frac{1}{(2\pi)^{p-1}} \\ &\times \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \tilde{Y}_p(j\theta_1, \dots, j\theta_{p-1}, j\Omega - j\theta_1 - \dots - j\theta_{p-1}) d\theta_1 \dots d\theta_{p-1}. \end{aligned} \quad (3.64)$$

Note que, substituindo (3.63) em (3.64) e usando a definição de U e de H_p , temos

$$\begin{aligned} \mathbf{C} \left[\tilde{Y}_p(j\Omega_1, \dots, j\Omega_p) \right] (j\Omega) &= \frac{1}{(2\pi)^{p-1}} \\ &\times \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_p(\tau_1, \dots, \tau_p) e^{-j\theta_1\tau_1} \dots e^{-j\theta_{p-1}\tau_{p-1}} e^{-j\tau_p(\Omega - \theta_1 - \dots - \theta_{p-1})} d\tau_1 \dots d\tau_p \\ &\times U(j\theta_1) \dots U(j\theta_{p-1}) \int_{-\infty}^{\infty} u(\sigma) e^{-j\sigma(\Omega - \theta_1 - \dots - \theta_{p-1})} d\sigma d\theta_1 \dots d\theta_{p-1}. \end{aligned} \quad (3.65)$$

⁶Em [52], esta forma é chamada no domínio de Laplace de “transformada de ordem p da saída”.

Com certa manipulação, chega-se à forma

$$\begin{aligned} \mathbb{C} \left[\tilde{Y}_p(j\Omega_1, \dots, j\Omega_p) \right] (j\Omega) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_p(\tau_1, \dots, \tau_p) \\ &\times \prod_{i=1}^{p-1} \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} U(j\theta_i) e^{j\theta_i(\sigma + \tau_p - \tau_i)} d\theta_i \right] u(\sigma) e^{-j\Omega(\sigma + \tau_p)} d\tau_1 \dots d\tau_p d\sigma \quad (3.66) \end{aligned}$$

através da qual, tomando-se as transformadas entre colchetes e definindo-se $t \triangleq \sigma + \tau_p$, obtém-se finalmente

$$\begin{aligned} \mathbb{C} \left[\tilde{Y}_p(j\Omega_1, \dots, j\Omega_p) \right] (j\Omega) &= \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h_p(\tau_1, \dots, \tau_p) u(t - \tau_1) \dots u(t - \tau_p) e^{-j\Omega t} d\tau_1 \dots d\tau_p dt \quad (3.67) \end{aligned}$$

ou seja,

$$\mathbb{C} \left[\tilde{Y}_p(j\Omega_1, \dots, j\Omega_p) \right] (j\Omega) = Y_p(j\Omega) = \text{TF} [y_p(t)] \quad (3.68)$$

onde o operador TF denota a transformada de Fourier e $y_p(t)$ é o componente de ordem p da saída produzida por T. Logo, de forma dual a (3.2), as equações (3.61)–(3.64) fornecem a saída no domínio da frequência em termos de seus componentes de todas as ordens.

Por fim, é importante observar que, admitindo-se que H_p é simétrica, então $\tilde{Y}_p$ também será. Logo, a escolha de qual dos argumentos de $\tilde{Y}_p$ será dado por $j(\Omega - \sum \theta_i)$ na integral do operador de contração é irrelevante.

3.1.4 Caracterização de distorções não-lineares no domínio da frequência

Um aspecto chave da teoria de sistemas lineares é o fato de que a resposta produzida por um sistema linear pode conter energia apenas em componentes com as mesmas frequências daqueles presentes na entrada. O mesmo não se aplica, contudo, a sistemas não-lineares, os quais introduzem novos componentes de frequência devido à ocorrência de fenômenos não-lineares como a intermodulação e a produção de harmônicos. As FRFGs associadas a um sistema não-linear caracterizam quantitativamente este comportamento, possibilitando prever como a

ocorrência de tais fenômenos afeta o espectro da saída.

Para vermos como as FRFGs fornecem informações a respeito dos harmônicos na saída, por exemplo, consideremos a análise em regime permanente da saída de um sistema não-linear com FRFGs $H_p(j\Omega_1, \dots, j\Omega_p)$ quando ele é excitado pela entrada

$$u(t) = \alpha e^{j\Omega_0 t} \quad (3.69)$$

cuja transformada é

$$U(j\Omega) = 2\pi\alpha\delta(\Omega - \Omega_0). \quad (3.70)$$

Empregando-se as equações (3.62)–(3.64) e (3.70), verifica-se que o componente de ordem p da saída no domínio da frequência tem a forma

$$\begin{aligned} Y_p(j\Omega) = 2\pi\alpha^p \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} H_p(j\theta_1, \dots, j\theta_{p-1}, j\Omega - j\theta_1 - \dots - j\theta_{p-1}) \\ \times \delta(\theta_1 - \Omega_0) \dots \delta(\theta_{p-1} - \Omega_0) \delta(\Omega - \theta_1 - \dots - \theta_{p-1} - \Omega_0) d\theta_1 \dots d\theta_{p-1} \end{aligned} \quad (3.71)$$

cujo integrando é não-nulo somente para

$$\theta_i = \Omega_0, \quad i = 1, \dots, p-1 \quad (3.72)$$

$$\Omega = \theta_1 + \dots + \theta_{p-1} + \Omega_0 = p\Omega_0. \quad (3.73)$$

Logo, tal componente corresponde à transformada de Fourier de um harmônico em $p\Omega_0$ dada por [9]

$$Y_p(j\Omega) = 2\pi\alpha^p H_p(j\Omega_0, \dots, j\Omega_0) \delta(\Omega - p\Omega_0). \quad (3.74)$$

Assim, de modo geral, o p -ésimo harmônico produzido para uma senóide com frequência Ω_0 pode ser determinado observando-se o valor de $H_p(j\Omega_1, \dots, j\Omega_p)$ para $\Omega_1 = \dots = \Omega_p = \Omega_0$. Em outras palavras, a distorção harmônica produzida por um sistema é caracterizada pelas diagonais de suas FRFGs.

Analogamente, os valores das FRFGs fora de suas diagonais principais caracterizam a distorção introduzida devido à intermodulação entre componentes de frequência. Como exemplo, consideremos agora a entrada $u(t)$ e sua transformada

$U(j\Omega)$ dadas por

$$u(t) = \alpha_1 e^{j\Omega_1 t} + \alpha_2 e^{j\Omega_2 t} \quad (3.75)$$

$$U(j\Omega) = 2\pi\alpha_1\delta(\Omega - \Omega_1) + 2\pi\alpha_2\delta(\Omega - \Omega_2). \quad (3.76)$$

Antes de empregarmos as fórmulas (3.62)–(3.64), observe que a expressão

$$U(j\Omega - j\theta_1 - \dots - j\theta_{p-1}) \prod_{i=1}^{p-1} U(j\theta_i) \quad (3.77)$$

é não-nula somente quando $\theta_i \in \{\Omega_1, \Omega_2\}$, $i = 1, \dots, p$, e

$$\Omega = \Omega_1 + \sum_i^{p-1} \theta_i \quad \text{ou} \quad \Omega = \Omega_2 + \sum_i^{p-1} \theta_i. \quad (3.78)$$

Portanto, tem-se que $\Omega = q\Omega_1 + (p - q)\Omega_2$, onde $q \in \mathbb{N}$ tal que $q \leq p$. Com isso, pode-se concluir que, ao se empregar tais fórmulas, a expressão resultante para Y_p será composta apenas por componentes cujas frequências Ω satisfazem $\Omega = q\Omega_1 + (p - q)\Omega_2$, tendo a forma

$$Y_p(j\Omega) = 2\pi \sum_{q=0}^p C_q^p \alpha_1^q \alpha_2^{p-q} H_p(j\Omega_1, \dots, j\Omega_1, j\Omega_2, \dots, j\Omega_2) \delta(\Omega - q\Omega_1 + (q - p)\Omega_2) \quad (3.79)$$

em que $C_q^p = p! [q!(p - q)!]^{-1}$, os q primeiros argumentos de H_p são iguais a $j\Omega_1$, e os $p - q$ restantes iguais a $j\Omega_2$. Os termos do somatório acima correspondem aos componentes produzidos por intermodulação de ordem p , sendo que a presença dos coeficientes C_q^p se deve ao fato de que cada componente é produzido por C_q^p formas diferentes de se combinarem q argumentos $j\Omega_1$ e $p - q$ argumentos $j\Omega_2$ em H_p (admitindo-se, novamente, a simetria de H_p).

Para uma análise mais detalhada dos fenômenos não-lineares discutidos, pode-se consultar [53], em que se mostra também como as FRFGs podem ser usadas na predição de variações não-lineares do ganho aparente em uma frequência causadas pela presença de não-linearidades de ordem ímpar. Ao leitor interessado em uma análise geral de operadores de Volterra em regime permanente no domínio da

frequência, recomenda-se [52].

3.1.5 A transformada de Laplace multidimensional

A transformada de Laplace desempenha um papel central na teoria de sistemas lineares. Quando generalizada para o caso multidimensional, ela constitui uma ferramenta matemática importante também no estudo de sistemas não-lineares. Através de seu uso, por exemplo, pode-se obter expressões simples para algumas estruturas formadas pela interconexão de sistemas em cascata ou em paralelo, o que se deve ao fato de certas formas convolutivas poderem ser representadas como produtos de transformadas de Laplace multidimensionais, assim como ocorre no caso unidimensional [32]. Adicionalmente, como será visto adiante, alguns métodos de cálculo da série de Volterra associada a um conjunto de equações diferenciais baseiam-se na representação dos núcleos no domínio de Laplace [9, 32].

Definição 3.3 Seja $h_p(\tau_1, \dots, \tau_p)$ um núcleo de Volterra de ordem p , causal e invariante no tempo. A transformada de Laplace de h_p , também chamada de função de transferência generalizada (FTG) de ordem p , é definida como

$$\begin{aligned} H_p(s_1, \dots, s_p) &= \mathcal{L}[h_p(\tau_1, \dots, \tau_p)] \\ &= \int_0^\infty \dots \int_0^\infty h_p(\tau_1, \dots, \tau_p) e^{-s_1\tau_1 - \dots - s_p\tau_p} d\tau_1 \dots d\tau_p. \end{aligned} \quad (3.80)$$

Observe que tal definição consiste, como no caso linear, em uma generalização da FRFG de h_p dada por (3.57), a qual é obtida quando $s_i = j\Omega_i$, $i = 1, \dots, p$.

É evidente que, ao se especificar a transformada de Laplace de um núcleo, deve-se estabelecer uma região de convergência apropriada em termos das variáveis $s_1, \dots, s_p$, para que (3.80) faça sentido.

A transformada de Laplace multidimensional inversa é também definida de forma análoga ao caso linear, sendo dada por

$$\begin{aligned} h_p(\tau_1, \dots, \tau_p) &= \mathcal{L}^{-1}[H_p(s_1, \dots, s_p)] \\ &= \int_{\sigma_p - j\infty}^{\sigma_p + j\infty} \dots \int_{\sigma_1 - j\infty}^{\sigma_1 + j\infty} H_p(s_1, \dots, s_p) e^{s_1\tau_1 + \dots + s_p\tau_p} ds_1 \dots ds_p \end{aligned} \quad (3.81)$$

em que os valores de $\sigma_1, \dots, \sigma_p$ devem ser escolhidos apropriadamente para que não haja problemas de convergência [32].

Apesar das similaridades até aqui apontadas entre a transformada unidimensional e a multidimensional, não há um método de inversão análogo à decomposição em frações parciais para funções racionais multivariadas [32], de forma que, exceto para algumas funções particulares, a integral (3.81) deve ser calculada para se obter a transformada de Laplace multidimensional inversa. Desse modo, fica claro que, embora a extensão das definições para o caso multidimensional seja trivial, há um aumento considerável de dificuldade na manipulação matemática de algumas transformadas.

3.1.6 Caracterização da resposta de um sistema de Volterra para uma combinação de exponenciais

Uma propriedade importante da transformada multidimensional de Laplace é que ela pode ser empregada para se caracterizar a resposta de um sistema de Volterra causal homogêneo e estável quando ele é excitado por uma combinação linear de exponenciais crescentes, i.e., um sinal na forma

$$u(t) = \sum_{m=1}^M \alpha_m e^{\lambda_m t} \quad (3.82)$$

em que $\text{Re}\{\lambda_m\} > 0$, $m \in \{1, \dots, M\}$ [32]. Logo, por meio desta ferramenta matemática pode-se caracterizar a resposta de um sistema de Volterra causal e estável qualquer, visto que ela é dada pelo somatório das saídas de subsistemas de Volterra homogêneos.

Tal caracterização é uma extensão daquela empregada para sistemas lineares,

a qual será aqui recordada. Para um sistema linear estável causal e invariante no tempo com resposta impulsiva $h(t)$, um fato bem conhecido é que a resposta produzida para a entrada (3.82) é dada por

$$y(t) = \int_0^\infty h(\tau) \left[\sum_{m=1}^M \alpha_m e^{\lambda_m(t-\tau)} \right] d\tau \quad (3.83)$$

$$= \sum_{m=1}^M \alpha_m \left[\int_0^\infty h(\tau) e^{-\lambda_m \tau} d\tau \right] e^{\lambda_m t} \quad (3.84)$$

$$= \sum_{m=1}^M \alpha_m H(\lambda_m) e^{\lambda_m t} \quad (3.85)$$

em que $H(s)$ é a transformada de Laplace de $h(t)$. Como a resposta do sistema para uma exponencial $e^{\lambda_m t}$ pode ser escrita como um produto entre tal exponencial e um fator que depende de λ_m , diz-se que $e^{\lambda_m t}$ é uma autofunção de sistemas lineares, em analogia com o conceito dos autovetores de uma transformação linear. Observe que, como se supõe que o sistema é estável, então não haverá problemas de convergência nas integrais de (3.84) visto que, por hipótese, $\text{Re}\{\lambda_m\} > 0$.

Consideremos, agora, um sistema de Volterra homogêneo de segundo grau estável e causal com núcleo h_2 simétrico, cuja relação entrada/saída é

$$y(t) = \int_0^\infty \int_0^\infty h_2(\tau_1, \tau_2) u(t - \tau_1) u(t - \tau_2) d\tau_1 d\tau_2. \quad (3.86)$$

Quando excitado por uma entrada na forma (3.82), tal sistema produz a saída

$$y(t) = \int_0^\infty \int_0^\infty h_2(\tau_1, \tau_2) \left[\sum_{m_1=1}^M \alpha_{m_1} e^{\lambda_{m_1}(t-\tau_1)} \right] \left[\sum_{m_2=1}^M \alpha_{m_2} e^{\lambda_{m_2}(t-\tau_2)} \right] d\tau_1 d\tau_2 \quad (3.87)$$

$$= \int_0^\infty \int_0^\infty h_2(\tau_1, \tau_2) \sum_{m_1=1}^M \sum_{m_2=1}^M \alpha_{m_1} \alpha_{m_2} e^{\lambda_{m_1}(t-\tau_1)} e^{\lambda_{m_2}(t-\tau_2)} d\tau_1 d\tau_2 \quad (3.88)$$

$$= \sum_{m_1=1}^M \sum_{m_2=1}^M \alpha_{m_1} \alpha_{m_2} \left[\int_0^\infty \int_0^\infty h_2(\tau_1, \tau_2) e^{-\lambda_{m_1} \tau_1} e^{-\lambda_{m_2} \tau_2} d\tau_1 d\tau_2 \right] e^{(\lambda_{m_1} + \lambda_{m_2})t}. \quad (3.89)$$

Logo, vê-se facilmente que

$$y(t) = \sum_{m_1=1}^M \sum_{m_2=1}^M \alpha_{m_1} \alpha_{m_2} H_2(\lambda_{m_1}, \lambda_{m_2}) e^{(\lambda_{m_1} + \lambda_{m_2})t} \quad (3.90)$$

em que $H_2(s_1, s_2)$ é a transformada de Laplace do núcleo $h_2(\tau_1, \tau_2)$. Note que, novamente, as suposições de que $\text{Re}\{\lambda_{m_1}\}, \text{Re}\{\lambda_{m_2}\} > 0$ e de que o sistema é estável implicam na garantia de convergência da integral em (3.89).

Usando-se o mesmo raciocínio pode-se mostrar que, de modo geral, a resposta de um sistema de Volterra homogêneo de grau p estável e causal para a entrada (3.82) é dada por uma expressão na forma

$$y(t) = \sum_{m_1=1}^M \dots \sum_{m_p=1}^M \alpha_{m_1} \dots \alpha_{m_p} H_p(\lambda_{m_1}, \dots, \lambda_{m_p}) e^{(\lambda_{m_1} + \dots + \lambda_{m_p})t} \quad (3.91)$$

onde $H_p(s_1, \dots, s_p)$ denota a transformada de Laplace multidimensional simétrica do núcleo $h_p(\tau_1, \dots, \tau_p)$. Uma forma alternativa de se escrever a expressão (3.91) seria [32]

$$y(t) = \sum_{\substack{i_1, \dots, i_M=0 \\ i_1 + \dots + i_M = p}}^p p! \alpha_1^{i_1} \dots \alpha_M^{i_M} H_p(\lambda_1^{(i_1)}, \dots, \lambda_M^{(i_M)}) e^{(i_1 \lambda_1 + \dots + i_M \lambda_M)t} \quad (3.92)$$

em que a notação $\lambda_j^{(i_j)}$ denota que o argumento λ_j é repetido i_j vezes. Por exemplo, para $M = 4$ e $p = 3$, temos que

$$H_3(\lambda_1^{(2)}, \lambda_2^{(0)}, \lambda_3^{(1)}, \lambda_4^{(0)}) = H_3(\lambda_1, \lambda_1, \lambda_3) \quad (3.93)$$

$$H_3(\lambda_1^{(0)}, \lambda_2^{(0)}, \lambda_3^{(0)}, \lambda_4^{(3)}) = H_3(\lambda_4, \lambda_4, \lambda_4). \quad (3.94)$$

A expressão (3.92) será importante no Capítulo 5, em que servirá como base para um método de derivação analítica das FRFGs associadas a um modelo físico.

3.1.7 Núcleos resultantes de algumas interconexões entre operadores

Como mencionado na subseção anterior, o uso da transformada de Laplace multidimensional simplifica a representação de alguns operadores não-lineares resultantes de interconexões entre dois ou mais operadores. Esta estratégia será empregada a seguir na derivação dos núcleos de duas estruturas que serão importantes adiante. Antes de apresentarmos tais estruturas, vamos estabelecer dois teoremas que serão úteis.

Teorema 3.1 Seja h_p um núcleo causal dado por

$$h_p(\tau_1, \dots, \tau_p) = \int_0^\infty g_1(\sigma) g_p(\tau_1 - \sigma, \dots, \tau_p - \sigma) d\sigma. \quad (3.95)$$

Então, a transformada de Laplace de h_p pode ser escrita em termos das transformadas de g_1 e g_p através da expressão [32]

$$H_p(s_1, \dots, s_p) = G_1(s_1 + \dots + s_p) G_p(s_1, \dots, s_p). \quad (3.96)$$

Demonstração. Aplicando a definição da transformada ao núcleo h_p , temos

$$H_p(s_1, \dots, s_p) = \int_0^\infty \dots \int_0^\infty \int_0^\infty g_1(\sigma) g_p(\tau_1 - \sigma, \dots, \tau_p - \sigma) e^{-s_1 \tau_1 - \dots - s_p \tau_p} d\sigma d\tau_1 \dots d\tau_p. \quad (3.97)$$

Definindo-se $\sigma_i \triangleq \tau_i - \sigma$ para $i = 1, \dots, p$, obtém-se

$$H_p(s_1, \dots, s_p) = \int_0^\infty \dots \int_0^\infty \int_0^\infty g_1(\sigma) g_p(\sigma_1, \dots, \sigma_p) \times e^{-s_1 \sigma_1 - \dots - s_p \sigma_p} e^{-\sigma(s_1 + \dots + s_p)} d\sigma d\sigma_1 \dots d\sigma_p \quad (3.98)$$

$$= \int_0^\infty g_1(\sigma) e^{-\sigma(s_1 + \dots + s_p)} d\sigma \int_0^\infty \dots \int_0^\infty g_p(\sigma_1, \dots, \sigma_p) e^{-s_1 \sigma_1 - \dots - s_p \sigma_p} d\sigma_1 \dots d\sigma_p. \quad (3.99)$$

Então, pela definição da transformada de Laplace, temos que (3.99) resulta em

$$H_p(s_1, \dots, s_p) = G_1(s_1 + \dots + s_p) G_p(s_1, \dots, s_p). \quad (3.100)$$

■

Teorema 3.2 Se um núcleo de ordem p pode ser escrito na forma

$$h_p(\tau_1, \dots, \tau_p) = \int_0^\infty \dots \int_0^\infty g_1(\tau_1 - \sigma_1) \dots g_1(\tau_p - \sigma_p) g_p(\sigma_1, \dots, \sigma_p) d\sigma_1 \dots d\sigma_p \quad (3.101)$$

então sua transformada de Laplace é dada por

$$H_p(s_1, \dots, s_p) = G_1(s_1) \dots G_1(s_p) G_p(s_1, \dots, s_p). \quad (3.102)$$

Demonstração. Novamente, da definição (3.80) temos que

$$\begin{aligned} H_p(s_1, \dots, s_p) &= \int_0^\infty \dots \int_0^\infty \int_0^\infty \dots \int_0^\infty g_1(\tau_1 - \sigma_1) \dots g_1(\tau_p - \sigma_p) g_p(\sigma_1, \dots, \sigma_p) \\ &\quad \times e^{-s_1\tau_1 - \dots - s_p\tau_p} d\sigma_1 \dots d\sigma_p d\tau_1 \dots d\tau_p. \end{aligned} \quad (3.103)$$

Com a introdução das variáveis $\lambda_i \triangleq \tau_i - \sigma_i$, $i = 1, \dots, p$, temos

$$\begin{aligned} H_p(s_1, \dots, s_p) &= \int_0^\infty \dots \int_0^\infty \left[\int_0^\infty g_1(\lambda_1) e^{-s_1\lambda_1} d\lambda_1 \right] \dots \left[\int_0^\infty g_1(\lambda_p) e^{-s_p\lambda_p} d\lambda_p \right] \\ &\quad \times g_p(\sigma_1, \dots, \sigma_p) e^{-s_1\sigma_1 - \dots - s_p\sigma_p} d\sigma_1 \dots d\sigma_p \end{aligned} \quad (3.104)$$

$$H_p(s_1, \dots, s_p) = G_1(s_1) \dots G_1(s_p) G_p(s_1, \dots, s_p). \quad (3.105)$$

■

Com base nos resultados apresentados acima, vamos agora derivar os núcleos de duas estruturas: aquela em que um operador não-linear é precedido por um operador linear, e aquela em que os mesmos blocos são interconectados na ordem inversa. A justificativa para o interesse particular nessas estruturas é que elas surgem no problema de determinação de filtros de Volterra de referência em aplicações de PDS. Expressões para alguns outros tipos de estrutura, como aquelas formadas por operadores não-lineares em paralelo cujas saídas são somadas ou multiplicadas, são fornecidas, por exemplo, em [52].

Consideremos, inicialmente, um operador não-linear K causal cuja estrutura, que é ilustrada na Figura 3.1, consiste em um operador não-linear e invariante no tempo, T , sucedido por um operador linear e invariante no tempo com função de transferência $G(s)$. Suponha que T possui uma representação de Volterra composta pela sequência de núcleos $\{h_p\}_{p \in \mathbb{N}^*}$. Então, a saída do operador que resulta da interconexão é dada por

$$y(t) = \sum_{p=1}^{\infty} y_p(t) \quad (3.106)$$

em que

$$y_p(t) = \int_0^{\infty} g(\tau) w_p(t - \tau) d\tau \quad (3.107)$$

$$= \int_0^{\infty} g(\tau) \int_0^{\infty} \dots \int_0^{\infty} h_p(\tau_1, \dots, \tau_p) u(t - \tau - \tau_1) \dots u(t - \tau - \tau_p) d\tau_1 \dots d\tau_p d\tau. \quad (3.108)$$

Definindo $\sigma_i \triangleq \tau + \tau_i$, $i = 1, \dots, p$, temos

$$y_p(t) = \int_0^{\infty} g(\tau) \int_0^{\infty} \dots \int_0^{\infty} h_p(\sigma_1 - \tau, \dots, \sigma_p - \tau) u(t - \sigma_1) \dots u(t - \sigma_p) d\sigma_1 \dots d\sigma_p d\tau \quad (3.109)$$

$$= \int_0^{\infty} \dots \int_0^{\infty} k_p(\sigma_1, \dots, \sigma_p) u(t - \sigma_1) \dots u(t - \sigma_p) d\sigma_1 \dots d\sigma_p \quad (3.110)$$

onde o núcleo resultante, k_p , é dado por

$$k_p(\sigma_1, \dots, \sigma_p) = \int_0^{\infty} g(\tau) h_p(\sigma_1 - \tau, \dots, \sigma_p - \tau) d\tau. \quad (3.111)$$

Por conseguinte, utilizando o resultado do Teorema 3.1, temos que os núcleos associados ao operador resultante podem ser escritos no domínio de Laplace como

$$K_p(s_1, \dots, s_p) = G(s_1 + \dots + s_p) H_p(s_1, \dots, s_p) \quad (3.112)$$

para $p \in \mathbb{N}^*$.

Figura 3.1: Estrutura composta pela interconexão de um operador não-linear seguido de um linear.

Empregando uma estratégia semelhante, pode-se também tratar o caso inverso, em que o bloco linear precede a não-linearidade. Consideremos, então, um operador não-linear $\mathbf{K}$ causal cuja estrutura, que é ilustrada na Figura 3.2, consiste em um operador linear e invariante no tempo com função de transferência $G(s)$ sucedido por um operador não-linear e invariante no tempo, $\mathbf{T}$. Novamente, suponha que $\mathbf{T}$ possui uma representação de Volterra composta pelos núcleos $\{h_p\}_{p \in \mathbb{N}^*}$. Então, a saída produzida pela estrutura considerada é tal que o componente de ordem p , $y_p(t)$, pode ser escrito como

$$y_p(t) = \int_0^\infty \dots \int_0^\infty h_p(\tau_1, \dots, \tau_p) w(t - \tau_1) \dots w(t - \tau_p) d\tau_1 \dots d\tau_p \quad (3.113)$$

$$= \int_0^\infty \dots \int_0^\infty h_p(\tau_1, \dots, \tau_p) \prod_{i=1}^p \left[\int_0^\infty g(\sigma_i) u(t - \tau_i - \sigma_i) d\sigma_i \right] d\tau_1 \dots d\tau_p. \quad (3.114)$$

Introduzindo-se as variáveis $\lambda_i \triangleq \tau_i + \sigma_i$, $i = 1, \dots, p$, (3.114) pode ser colocada na forma

$$y_p(t) = \int_0^\infty \dots \int_0^\infty h_p(\tau_1, \dots, \tau_p) \prod_{i=1}^p \left[\int_0^\infty g(\lambda_i - \tau_i) u(t - \lambda_i) d\lambda_i \right] d\tau_1 \dots d\tau_p \quad (3.115)$$

$$= \int_0^\infty \dots \int_0^\infty k_p(\lambda_1, \dots, \lambda_p) u(t - \lambda_1) \dots u(t - \lambda_p) d\lambda_1 \dots d\lambda_p \quad (3.116)$$

em que o núcleo resultante k_p é dado no domínio do tempo por

$$k_p(\lambda_1, \dots, \lambda_p) = \int_0^\infty \dots \int_0^\infty g(\lambda_1 - \tau_1) \dots g(\lambda_p - \tau_p) h_p(\tau_1, \dots, \tau_p) d\tau_1 \dots d\tau_p. \quad (3.117)$$

Logo, de acordo com o Teorema 3.2, sua transformada de Laplace é

$$K_p(s_1, \dots, s_p) = G(s_1) \dots G(s_p) H_p(s_1, \dots, s_p). \quad (3.118)$$

Figura 3.2: Estrutura composta pela interconexão de um operador linear seguido de um não-linear.

Com as derivações realizadas, fica claro que, de fato, o efeito combinado dos operadores conectados em série é muito mais fácil de se representar e de se interpretar no domínio de Laplace (ou da frequência).

3.1.8 Questões de convergência

Por se tratar de um somatório infinito de integrais impróprias, a representação de Volterra dada por (3.1) faz sentido apenas quando condições apropriadas de convergência são estabelecidas. Dessa forma, somente podemos afirmar que um operador T admite uma representação de Volterra quando existe uma série que seja convergente para um certo conjunto de entradas e forneça uma representação exata de T . Como os termos da série envolvem potências na entrada, uma forma simples de se abordar este problema é reduzi-lo ao da convergência de uma série de potências convencional [32]. Para tanto, é conveniente utilizar a definição que segue.

Definição 3.4 Sejam $\{h_p\}_{p \in \mathbb{N}^*}$ os núcleos de uma série de Volterra que representa um operador não-linear causal e invariante no tempo $\mathsf{T} : U \rightarrow Y$. Então, a função de limite de ganho de tal série é dada por [49, 50]

$$\phi : \mathbb{R}_+ \rightarrow \mathbb{R} \quad (3.119)$$

$$\phi(x) = \sum_{p=1}^{\infty} \|h_p\|_1 x^p \quad (3.120)$$

cujos raio de convergência, portanto, é [49]

$$\rho_\phi = \left(\limsup_{p \rightarrow \infty} \sqrt[p]{\|h_p\|_1} \right)^{-1}. \quad (3.121)$$

Com isso, podemos então estabelecer a condição de convergência descrita a seguir.

Teorema 3.3 [49] Sejam $\{h_p\}_{p \in \mathbb{N}}$ os núcleos que caracterizam uma série de Volterra causal e invariante no tempo cuja função de limite de ganho é $\phi(x)$. Então, uma condição suficiente para a sua convergência é que a entrada $u(t)$ satisfaça $\|u\|_\infty < \rho_\phi$, em que ρ_ϕ é o raio de convergência de $\phi(x)$.

Demonstração.

$$|y(t)| = \left| \sum_{p=1}^{\infty} \int_0^{\infty} \dots \int_0^{\infty} h_p(\tau_1, \dots, \tau_p) \prod_{i=1}^p u(t - \tau_i) d\tau_1 \dots d\tau_p \right| \quad (3.122)$$

$$\leq \sum_{p=1}^{\infty} \int_0^{\infty} \dots \int_0^{\infty} |h_p(\tau_1, \dots, \tau_p)| \prod_{i=1}^p |u(t - \tau_i)| d\tau_1 \dots d\tau_p \quad (3.123)$$

$$\leq \sum_{p=1}^{\infty} \|h_p\|_1 \|u\|_\infty^p \quad (3.124)$$

$$= \phi(\|u\|_\infty). \quad (3.125)$$

Como estamos admitindo que $\|u\|_\infty < \rho_\phi$, então $y(t)$ é limitada e converge normalmente de modo que [49, 50]

$$\|y\|_\infty \leq \phi(\|u\|_\infty) < \infty. \quad (3.126)$$

■

Observe que, evidentemente, a condição estabelecida pelo Teorema 3.3, embora seja suficiente, pode ser demasiadamente restritiva [32]. Em outras palavras, as desigualdades nele empregadas podem conduzir a um limite superior superestimado para $|y(t)|$. Ainda que se deseje recorrer a este resultado, seu uso prático direto é dificultado pela necessidade de se estimar o comportamento assintótico das normas dos núcleos da série de interesse, o que não é uma tarefa fácil, exceto para sistemas em certas formas particulares [50].

Recentemente, em [50] foi proposto um algoritmo para o cômputo de um limite inferior para o raio de convergência que é fornecido pelo resultado acima,

restringindo-se à classe dos sistemas lineares analíticos, os quais possuem a forma

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t) + \mathbf{g}(\mathbf{x}(t), t)u(t) \quad (3.127)$$

$$y(t) = \lambda(\mathbf{x}(t), t) \quad (3.128)$$

onde $u(t)$ e $y(t)$ são escalares e as funções reais $\mathbf{f}$, $\mathbf{g}$ e λ são analíticas.⁷ Para tanto, basicamente explora-se uma fórmula de construção recursiva dos núcleos de Volterra, de maneira que a estimação do comportamento assintótico de suas normas é evitada calculando-se o raio de convergência de uma série de potências relacionada, o qual é, por construção, menor ou igual ao da série de Volterra de interesse. Sendo assim, conforme apontado pelos autores, o limite calculado através do algoritmo pode ser uma estimativa subestimada do limite efetivo relativo ao Teorema 3.3.

Uma alternativa ao resultado deste teorema seria estabelecer condições envolvendo restrições não só à magnitude da entrada, mas também ao intervalo de tempo considerado, como é por vezes feito na literatura [32]. Contudo, a determinação precisa de limites associados a tais condições para uma série de Volterra qualquer não é uma tarefa simples [32, 54]. Ao invés disso, existem resultados relacionados à existência de limites de convergência para representações de Volterra associadas a classes específicas de sistemas.

Em [36] demonstra-se que, se um sistema na forma linear analítica admite uma solução no caso não-forçado (i.e., com $u(t) \equiv 0$), então existe uma representação de Volterra com núcleos $\{h_p\}_{p \in \mathbb{N}^*}$ contínuos por trechos tais que, para qualquer $T > 0$, existe $\varepsilon(T) > 0$ para o qual a série converge absoluta e uniformemente no intervalo $[0, T]$, desde que a entrada $u(t)$ seja uma função contínua por trechos e satisfaça $|u(t)| < \varepsilon(T)$ em $[0, T]$. Deve-se observar que um aspecto inconveniente deste resultado é que sua validade se restringe a um intervalo de tempo finito $[0, T]$. Adicionalmente, com o aumento do limite superior deste intervalo, T , o comportamento de ε possivelmente é tal que $\varepsilon \rightarrow 0$.

⁷Deve-se recordar que uma função real é dita analítica se e somente se ela é localmente igual à sua expansão em série de Taylor em torno de qualquer ponto. Note que, para tanto, é necessário que ela seja infinitamente diferenciável.

Já no caso de sistemas bilineares, a própria derivação dos núcleos feita na Subseção 3.1.2 consiste em uma demonstração construtiva de que tais sistemas sempre admitem uma representação de Volterra uniformemente convergente em qualquer intervalo de tempo finito [32, 36].

3.1.9 Aproximação pela série truncada

Apesar da importância teórica da questão de existência de uma representação de Volterra (exata) para um determinado operador não-linear T , na prática, é de maior interesse sabermos se T pode ser aproximado com um nível de precisão desejado por alguma série de Volterra truncada. Em outras palavras, o que se deseja saber é se existe uma série composta por núcleos $\{h_p\}_{p \in \{1, \dots, P\}}$, com $P < \infty$, tal que

$$\hat{y}(t) = \sum_{p=1}^P \int_0^\infty \int_0^{\tau_p} \dots \int_0^{\tau_2} h_p(\tau_1, \dots, \tau_p) u(t - \tau_1) \dots u(t - \tau_p) d\tau_1 \dots d\tau_{p-1} d\tau_p \quad (3.129)$$

aproxime com precisão a saída $y(t) = T[u(t)]$.

Em se tratando de uma função contínua, de acordo com o Teorema de Stone-Weierstrass, sabemos que se pode aproximá-la tão precisamente quanto desejado sobre um domínio finito de interesse através de um polinômio. Observe que a aproximação mencionada no parágrafo anterior pode ser vista como uma generalização daquela de uma função por um polinômio; logo, é pertinente nos questionarmos se o resultado que vale naquele caso pode ser estendido para este.

De fato, com o uso do Teorema de Stone-Weierstrass, alguns resultados de aproximação foram desenvolvidos na literatura, como o que é enunciado a seguir.

Teorema 3.4 [32, 55] Seja U um subconjunto compacto de $L^2[0, T]$, $T \in \mathbb{R}$, e suponha que $T : U \rightarrow Y$ é um operador causal, invariante no tempo e contínuo sobre o intervalo de tempo $[0, T]$. Seja $\epsilon \in \mathbb{R}$ tal que $\epsilon > 0$. Então, existe uma série de Volterra truncada tal que, $\forall u \in U$ e $0 \leq t \leq T$, sua saída no instante t , $\hat{y}(t)$, satisfaz

$$|T[u](t) - \hat{y}(t)| \leq \epsilon \quad (3.130)$$

em que $\mathsf{T}[u](t)$ denota a saída de T no instante t quando ele opera sobre a função $u(t)$.

Demonstração. A demonstração será aqui omitida, sendo delineada em [32] com base no conteúdo de [56].

O resultado expresso pelo teorema acima é de certa forma limitado, o que se deve a três restrições indesejáveis: a finitude do intervalo de aproximação considerado, $[0, T]$, o requisito de compacidade do conjunto de funções de entrada, U , e a finitude do intervalo em que as entradas podem ser não-nulas [55]. Tais condições excluem a possibilidade de se considerar alguns conjuntos de entradas típicos, como, por exemplo, aqueles que compreendem combinações lineares de exponenciais complexas (as quais assumem valores não-nulos sobre um intervalo infinito de tempo) e o conjunto de entradas $u(t)$ tais que $\|u(t)\|_2 \leq 1$, o qual não é compacto [32].

A origem das restrições mencionadas reside no fato de que, para estabelecer o Teorema 3.4, emprega-se justamente o Teorema de Stone-Weierstrass, em que suposições análogas são necessárias. Uma forma de se contorná-las é estabelecendo-se que os operadores não-lineares considerados possuam memória desvanescente, o que corresponde a uma condição mais forte que a de continuidade. Ainda assim, trata-se de um requisito que é satisfeito por uma ampla classe de sistemas físicos e que, embora tenha sido introduzido formalmente em [55], já era considerado de maneira informal por Volterra e Wiener em seus trabalhos, conforme é observado em [55].

Definição 3.5 Seja $\mathsf{T} : U \rightarrow Y$ um operador não-linear causal e invariante no tempo, com $U \subset \mathbf{C}(\mathbb{R})$. Diz-se que T possui memória desvanescente sobre U se existe uma função decrescente $w : \mathbb{R}_+ \rightarrow (0, 1]$, satisfazendo $\lim_{t \rightarrow \infty} w(t) = 0$, tal que $\forall u, v \in U$ e $\epsilon > 0$, $\exists \delta > 0$ para o qual

$$\sup_{\tau < t} |u(\tau) - v(\tau)| w(t - \tau) < \delta \implies |\mathsf{T}[u](t) - \mathsf{T}[v](t)| < \epsilon. \quad (3.131)$$

Informalmente, podemos interpretar a Definição 3.5 da seguinte maneira: para um sistema com memória desvanescente, a influência que o valor da entrada

$u(t_0)$ exerce sobre a saída no instante $y(t)$ decai gradualmente com o aumento de $t - t_0$. Evidentemente, trata-se de uma condição que pode ser admitida para inúmeros sistemas físicos considerados em aplicações práticas. Com isso, pode-se estabelecer o importante resultado descrito a seguir.

Teorema 3.5 [55] Seja $\epsilon \in \mathbb{R}$ tal que $\epsilon > 0$ e

$$U = \{u \in \mathbf{C}(\mathbb{R}) \mid \|u\|_\infty < M_1 \text{ e } |u(t) - u(\tau)| \leq M_2(t - \tau), \text{ com } \tau \leq t\}. \quad (3.132)$$

Seja $\mathbf{T} : U \rightarrow Y$ um operador não-linear causal e invariante no tempo com memória desvanescente em U . Então, existe uma série de Volterra truncada tal que, $\forall u \in U$, sua saída, $\hat{y}(t)$, satisfaz

$$\|\hat{y}(t) - \mathbf{T}[u(t)]\|_\infty \leq \epsilon. \quad (3.133)$$

Demonstração. A demonstração deste teorema, a qual é relativamente elaborada, também não será descrita aqui. O leitor interessado pode consultá-la em [55].

Observe que, enquanto o Teorema 3.4 também estabelece a existência de uma série truncada que aproxima $\mathbf{T}$ com precisão arbitrária, a introdução da condição relativa à memória desvanescente permite eliminar os requisitos de compacidade do conjunto de entradas e finitude do intervalo de aproximação. Ao invés disso, as condições admitidas para a entrada, que são sua continuidade, a existência de um limite superior M_1 para sua norma e também para o seu crescimento (i.e., as entradas possuem constante de Lipschitz M_2), são bastante naturais do ponto de vista prático. Logo, para a classe dos sistemas com memória desvanescente, trata-se de um resultado forte que fornece uma justificativa teórica para o uso de modelos baseados em séries de Volterra truncadas em aplicações práticas. Deve-se notar que, para uma determinada série de Volterra truncada, há uma relação entre os valores de ϵ e M_1 tal que, com o aumento de M_1 , em geral há uma redução no nível de precisão atingido, ϵ , e vice-versa.

Na realidade, o Teorema 3.4 também faz uso do Teorema de Stone-Weierstrass, sendo as restrições mais brandas estabelecidas ao conjunto das entradas (i.e., a

delimitação da norma máxima e do crescimento máximo que tais entradas podem apresentar) suficientes neste caso pois, uma vez admitida a propriedade de memória desvanescente, tais condições fazem com que este conjunto “pareça” compacto ao operador não-linear em questão, ainda que não seja por si só compacto [55]. Esta é, em linhas gerais, a ideia central de que os autores do teorema lançaram mão para contornar os requisitos de compacidade e de finitude do intervalo considerado.

Um outro resultado relacionado à aproximação pela série de Volterra truncada, apresentado em [49], estabelece um limite superior para o erro cometido nesta aproximação quando a série truncada é obtida a partir da representação de Volterra de um operador (note que este não é necessariamente o caso nos Teoremas 3.4 e 3.5). Basicamente, este limite é obtido ao se considerar a cauda da série que define a função de limite de ganho, incluindo-se apenas os termos que correspondem às ordens maiores que a ordem de truncamento P . Isto é, pode-se mostrar que, se a série de Volterra formada pelos núcleos $\{h_p\}_{p \in \{1, \dots, P\}}$ corresponde ao truncamento de ordem P da série composta pelos núcleos $\{h_p\}_{p \in \mathbb{N}^*}$, então sua saída, $\hat{y}(t)$, satisfaz

$$\|y(t) - \hat{y}(t)\|_\infty \leq \sum_{p=P+1}^{\infty} \|h_p\|_1 \|u\|_\infty^p \quad (3.134)$$

sendo $y(t)$ a saída da série não truncada e $u(t)$ a entrada.

3.1.10 O Filtro de Volterra

Embora os resultados da subseção anterior assegurem a ampla capacidade de aproximação das séries de Volterra truncadas, os modelos usados para fins de computação numérica consistem, na realidade, em uma versão de (3.129) em que o tempo é discreto e a memória também é truncada. Tais modelos, denominados filtros de Volterra (FV), são definidos como segue.

Definição 3.6 Um filtro de Volterra $\mathcal{F}$ causal e invariante no tempo consiste em um modelo matemático não-linear de tempo discreto cuja relação entrada/saída

é definida como

$$y(n) = \sum_{p=1}^P \sum_{k_1=0}^{N_p-1} \dots \sum_{k_p=k_{p-1}}^{N_p-1} h_p(k_1, \dots, k_p) u(n - k_1) \dots u(n - k_p) \quad (3.135)$$

em que as funções $h_p : \{0, 1, \dots, N_p - 1\}^p \rightarrow \mathbb{R}$ são chamadas de núcleos de Volterra de tempo discreto, sendo N_p a extensão da memória do núcleo de ordem p , $u : \mathbb{Z}_+ \rightarrow \mathbb{R}$ é a entrada e $y : \mathbb{Z}_+ \rightarrow \mathbb{R}$ é a saída.

Observe que, na Definição 3.6, admite-se uma forma triangular para os núcleos. Tal suposição é usualmente empregada na formulação de um FV uma vez que, fixadas a ordem e as extensões das memórias de seus núcleos, ela envolve o número mínimo de operações aritméticas necessárias para o cômputo da saída, sendo, portanto, adequada para implementação.⁸

Em [55], um resultado análogo ao do Teorema 3.5 é obtido para séries de Volterra truncadas de tempo discreto. Para tanto, a definição de desvanescência no tempo (Definição 3.5) é ajustada de modo que w seja uma sequência, i.e., $w : \mathbb{Z}_+ \rightarrow (0, 1]$. Com isso, pode-se mostrar o seguinte resultado.

Teorema 3.6 [55] Seja $\epsilon \in \mathbb{R}$ tal que $\epsilon > 0$ e

$$U = \{u \in \ell^\infty \mid \|u\|_\infty < M_1\}. \quad (3.136)$$

Seja $T : U \rightarrow Y$, em que $Y \subset \ell^\infty$, um operador de tempo discreto não-linear, causal e invariante no tempo com memória desvanescente em U . Então, existe uma série de Volterra de tempo discreto truncada tal que sua saída, $\hat{y}(n)$, satisfaz, $\forall u \in U$,

$$\|\hat{y}(n) - T[u(n)]\|_\infty \leq \epsilon. \quad (3.137)$$

Demonstração. Novamente, a demonstração está disponível em [55].

Note que, semelhantemente ao que foi discutido em relação ao Teorema 3.5, para uma dada série de Volterra de tempo discreto truncada, há uma relação de

⁸Deve-se ressaltar que, nesta discussão, considera-se o cálculo da saída no domínio do tempo. No Capítulo 4 será discutida uma forma de se implementar um FV no domínio da frequência que, sob determinadas condições, é mais eficiente.

inter-dependência entre M_1 e ϵ na forma de um compromisso entre a precisão da aproximação e a máxima norma admitida para a entrada.

Neste ponto, é importante fazermos também o seguinte comentário a respeito do Teorema 3.6. Na realidade, a série de Volterra de tempo discreto considerada em [55] é truncada apenas na ordem, e não na memória. Em outras palavras, os núcleos são definidos de modo a possuir memória de duração infinita. Na prática, evidentemente, isso não é realizável para modelos em que a saída de um instante se dá apenas em função de amostras da entrada, como é o caso dos filtros de Volterra convencionais. Porém, admitindo-se que a energia do núcleo de ordem p após um certo instante de tempo N_p é pouco significativa em relação à sua energia total, então a aproximação oferecida por um FV será, em geral, aproximadamente tão precisa quanto aquela oferecida pela série com núcleos não truncados na memória.

Quando um FV é empregado para fins de cômputo, a complexidade computacional associada ao seu uso está diretamente relacionada com o número de coeficientes que é necessário para descrever seus núcleos numericamente sobre os respectivos domínios de suporte. De fato, este número corresponde justamente ao total de termos envolvidos no somatório de (3.135), de modo que o cômputo de cada amostra da saída, quando realizado diretamente a partir da definição, envolve um número de operações aritméticas que é uma função de tal número.

Observando-se a equação (3.135), pode-se ver que tal número de coeficientes é $O(N_P^P)$, crescendo muito rapidamente, portanto, com P e com a duração da memória do núcleo de ordem P . Contudo, devido à redundância envolvida na definição do núcleo de ordem p quando $p > 1$ (semelhantemente ao que foi discutido para a série de tempo contínuo), o número de coeficientes necessários para caracterizá-lo é menor do que N_p^p . Mais precisamente, o número total de coeficientes de um FV $\mathcal{F}$ definido por (3.135) é

$$\Psi_{\text{FV}}(\mathcal{F}) = \sum_{p=1}^P \binom{N_p + p - 1}{p} \quad (3.138)$$

$$= \sum_{p=1}^P \frac{(N_p + p - 1)!}{p!(N_p - 1)!}. \quad (3.139)$$

A expressão (3.138) fornece o número das diferentes combinações dos índices k_i necessárias para descrever núcleos triangulares. Observe que se utiliza, para tanto, a fórmula de combinação com reposição, uma vez que apenas uma permutação de determinados índices $k_1, \dots, k_p$ conta para o resultado (i.e., aquela que satisfaz a condição de triangularidade) e pode haver $p - 1$ repetições de valores nos índices $k_1, \dots, k_p$.

Deve-se notar que, ainda que a representação simétrica de um núcleo seja adotada ao invés da triangular, por exemplo, o total de coeficientes que o caracteriza é essencialmente o mesmo, uma vez que, neste caso, os coeficientes associados a todas as permutações distintas de certos índices $k_1, \dots, k_p$ são necessariamente idênticos.

3.1.10.1 Funções de resposta em frequência generalizadas de tempo discreto

Analogamente à definição das FRFGs em tempo contínuo, as FRFGs de um FV são definidas como as transformadas de Fourier de tempo discreto (TFTD) multidimensionais dos núcleos do filtro, como é feito a seguir.

Definição 3.7 Seja $\mathcal{F}$ um filtro de Volterra de tempo discreto constituído de núcleos h_p , $p = 1, \dots, P$, com $P > 1$. A função de resposta em frequência generalizada (FRFG) de seu núcleo de ordem p é dada por

$$H_p(e^{j\omega_1}, \dots, e^{j\omega_p}) = \sum_{k_1=0}^{\infty} \dots \sum_{k_p=0}^{\infty} h_p(k_1, \dots, k_p) e^{-j\omega_1 k_1} \dots e^{-j\omega_p k_p} \quad (3.140)$$

sendo uma função periódica com período de 2π em todos os argumentos $\omega_1, \dots, \omega_p$. Naturalmente, o núcleo h_p também pode ser obtido a partir de sua FRFG com o uso da transformada inversa

$$h_p(k_1, \dots, k_p) = \frac{1}{(2\pi)^p} \int_{-\pi}^{\pi} \dots \int_{-\pi}^{\pi} H_p(e^{j\omega_1}, \dots, e^{j\omega_p}) e^{j\omega_1 k_1} \dots e^{j\omega_p k_p} d\omega_1 \dots d\omega_p. \quad (3.141)$$

Na prática, o cômputo numérico de transformadas de núcleos em tempo discreto é feito com o uso da transformada de Fourier discreta (TFD) multidimen-

sional, que consiste em uma extensão da TFD convencional, como definido a seguir.

Definição 3.8 Seja $h_p(k_1, \dots, k_p)$ um núcleo de ordem p com memória de N_p amostras. A transformada de Fourier discreta multidimensional de h_p é dada por

$$H_p(m_1, \dots, m_p) = \text{TFD}[h_p(k_1, \dots, k_p)] \quad (3.142)$$

$$= \sum_{k_1=0}^{N_p-1} \dots \sum_{k_p=0}^{N_p-1} h_p(k_1, \dots, k_p) e^{-j\frac{2\pi}{N_p}m_1k_1} \dots e^{-j\frac{2\pi}{N_p}m_pk_p} \quad (3.143)$$

para $m_i = 0, \dots, N_p - 1$, com $i = 1, \dots, p$. A relação inversa, nesse caso, é dada por

$$h_p(k_1, \dots, k_p) = \text{TFD}^{-1}[H_p(m_1, \dots, m_p)] \quad (3.144)$$

$$= \frac{1}{N_p^p} \sum_{m_1=0}^{N_p-1} \dots \sum_{m_p=0}^{N_p-1} H_p(m_1, \dots, m_p) e^{j\frac{2\pi}{N_p}m_1k_1} \dots e^{j\frac{2\pi}{N_p}m_pk_p}. \quad (3.145)$$

Assim como no caso unidimensional, vê-se que a TFD multidimensional de um núcleo fornece uma função discreta que corresponde à amostragem da TTFD do núcleo em pontos nos quais os índices de frequência são múltiplos inteiros de $2\pi/N_p$.

A transformada dada na Definição 3.8 pode ser empregada para se computar a convolução multidimensional circular $\bar{y}_p(n)$ entre uma sequência $u(n)$ e um núcleo de ordem p , $h_p(k_1, \dots, k_p)$, analogamente ao caso linear. Para tanto, usam-se as relações [1]

$$\tilde{Y}_p(m_1, \dots, m_p) = H_p(m_1, \dots, m_p)U(m_1)\dots U(m_p) \quad (3.146)$$

$$\tilde{y}_p(n_1, \dots, n_p) = \text{TFD}^{-1} \left[\tilde{Y}_p(m_1, \dots, m_p) \right] \quad (3.147)$$

$$\bar{y}_p(n) = \tilde{y}_p(n_1, \dots, n_p) \Big|_{n_1=\dots=n_p=n} \quad (3.148)$$

em que $m_1, \dots, m_p, n_1, \dots, n_p, n \in \{0, \dots, M - 1\}$, M é o comprimento escolhido para a convolução circular e $U(m)$ é a TFD unidimensional de comprimento M de $u(n)$. Deve-se notar que o resultado da convolução multidimensional regular

(i.e., linear)

$$y_p(n) = \sum_{k_1=0}^{N_p-1} \dots \sum_{k_p=0}^{N_p-1} h_p(k_1, \dots, k_p) u(n - k_1) \dots u(n - k_p) \quad (3.149)$$

pode ser obtido com o uso de (3.146)–(3.148) se for definido um comprimento M adequado para as transformadas. Mais especificamente, M deve ser maior ou igual ao comprimento da resposta de $y_p(n)$, $N_p + L - 1$, em que L é a duração da sequência $u(n)$. Logicamente, isso implica em computar as TFDs $U(m)$ e $H_p(m_1, \dots, m_p)$ considerando-se, respectivamente, uma sequência $\tilde{u}$ e um núcleo $\tilde{h}_p$ dados por

$$\tilde{u}(n) = \begin{cases} u(n), & n \in \{0, \dots, L - 1\} \\ 0, & n \in \{L, \dots, M - 1\} \end{cases} \quad (3.150)$$

$$\tilde{h}_p(k_1, \dots, k_p) = \begin{cases} h_p(k_1, \dots, k_p), & \forall i \in \{1, \dots, p\}, k_i \in \{0, \dots, N_p - 1\} \\ 0, & \exists i \in \{1, \dots, p\} \text{ tal que } k_i \in \{N_p, \dots, M - 1\} \end{cases} \quad (3.151)$$

Em outras palavras, a entrada e o núcleo são estendidos com zeros para que suas TFDs possuam o comprimento mínimo necessário para evitar um rebatimento causado pela convolução circular. Sob tais condições, tem-se então que $\forall n \in \{0, \dots, N_p + L - 1\}$, $\bar{y}_p(n) = y_p(n)$.

3.1.10.2 Discretização de núcleos de Volterra

Quando os núcleos de uma série de Volterra de tempo contínuo são conhecidos, pode-se aplicar um método de discretização para se obter núcleos correspondentes de tempo discreto. No que segue, descrevem-se dois métodos que foram propostos na literatura com essa finalidade.

O primeiro deles, chamado de invariância dos núcleos [9], é uma generalização do método de invariância do impulso, o qual é usado na discretização de respostas impulsivas. Trata-se de um método simples em que os coeficientes do núcleo de

tempo discreto correspondem simplesmente a amostras de suas contrapartes de tempo contínuo tomadas em uma grade regular. Isto é, o núcleo de tempo discreto, $h_p(k_1, \dots, k_p)$, é obtido a partir do de tempo contínuo, $\tilde{h}_p(\tau_1, \dots, \tau_p)$, através da relação

$$h_p(k_1, \dots, k_p) = (T_s)^p \tilde{h}_p(k_1 T_s, \dots, k_p T_s) \quad (3.152)$$

onde T_s é o intervalo de amostragem adotado. Note que o uso deste método corresponde a aproximar a integral

$$y_p(t) = \int_0^{N_p T_s} \dots \int_0^{N_p T_s} \tilde{h}_p(\tau_1, \dots, \tau_p) u(t - \tau_1) \dots u(t - \tau_p) d\tau_1 \dots d\tau_p \quad (3.153)$$

pelo somatório dos volumes de hiper-paralelepípedos dado por

$$y_p(nT_s) = \sum_{k_1=0}^{N_p-1} \dots \sum_{k_p=0}^{N_p-1} (T_s)^p \tilde{h}_p(k_1 T_s, \dots, k_p T_s) u((n - k_1)T_s) \dots u((n - k_p)T_s) \quad (3.154)$$

para se obter os valores da saída nos instantes de tempo nT_s , $n \in \mathbb{Z}$.

Uma característica do método de invariância dos núcleos é que ele normalmente introduz rebatimento no domínio da frequência [57], o que pode ser um problema em algumas situações. Tal característica se deve ao fato de que a amostragem no domínio do tempo é dada pelo produto

$$h_p(\tau_1, \dots, \tau_p) = \tilde{h}_p(\tau_1, \dots, \tau_p) s_1(\tau_1) \dots s_p(\tau_p) \quad (3.155)$$

em que

$$s_i(\tau_i) = \sum_{k=-\infty}^{\infty} \delta(\tau_i - kT_s) \quad (3.156)$$

e, portanto, no domínio da frequência, corresponde à convolução multidimensional entre a transformada de $\tilde{h}_p$ e as transformadas de s_i , $i \in \{1, \dots, p\}$, as quais possuem a forma [58]

$$S_i(j\Omega_i) = \Omega_s \sum_{k=-\infty}^{\infty} \delta(\Omega_i - k\Omega_s) \quad (3.157)$$

em que $\Omega_s = 2\pi/T_s$ é a frequência de amostragem. Logo, há a possibilidade de

superposição das “cópias” das transformadas em regiões próximas de múltiplos inteiros de $\Omega_N = \Omega_s/2$, uma vez que

$$H_p(e^{j\omega_1}, \dots, e^{j\omega_p}) = \sum_{m_1=-\infty}^{\infty} \dots \sum_{m_p=-\infty}^{\infty} \tilde{H}_p\left(j\frac{\omega_1 + 2\pi m_1}{T_s}, \dots, j\frac{\omega_p + 2\pi m_p}{T_s}\right) \quad (3.158)$$

para $-\pi \leq \omega_i \leq \pi$, $i \in \{1, \dots, p\}$, em que H_p é a FRFG de tempo discreto de h_p e $\tilde{H}_p$ a FRFG de $\tilde{h}_p$. Portanto, de modo geral, tem-se que

$$H_p(e^{j\omega_1}, \dots, e^{j\omega_p}) \neq \tilde{H}_p\left(j\frac{\omega_1}{T_s}, \dots, j\frac{\omega_p}{T_s}\right) \quad (3.159)$$

para $-\pi \leq \omega_i \leq \pi$, $i \in \{1, \dots, p\}$.

Evidentemente, a diminuição do intervalo de amostragem T_s – ou, equivalentemente, o aumento de Ω_s – tende a atenuar o efeito discutido acima, uma vez que, para sistemas físicos, espera-se que a energia de suas FRFGs decaia gradualmente com o aumento dos índices de frequência. Uma alternativa, contudo, seria derivar as FRFGs de tempo discreto a partir das FRFGs de tempo contínuo de tal forma que

$$H_p(e^{j\omega_1}, \dots, e^{j\omega_p}) = \tilde{H}_p\left(j\frac{\omega_1}{T_s}, \dots, j\frac{\omega_p}{T_s}\right) \quad (3.160)$$

para $-\pi \leq \omega_i \leq \pi$, $i = 1, \dots, p$. O segundo método de discretização dos núcleos aqui apresentado baseia-se justamente em tal princípio, consistindo, portanto, em definir os núcleos de tempo discreto de modo que seu conteúdo espectral coincida com aquele dos núcleos de tempo contínuo na banda de interesse, a qual é definida pela frequência de amostragem [57].

O método discutido no parágrafo anterior pode ser visto como um janelamento no domínio da frequência, de modo que sejam eliminados os componentes com frequências maiores que Ω_N , seguido de uma amostragem regular no domínio do tempo [57]. Isto é, definindo-se a janela retangular

$$W(j\Omega) = \begin{cases} 1, & \text{se } |\Omega| \leq \Omega_N \\ 0, & \text{caso contrário} \end{cases} \quad (3.161)$$

então, o núcleo de tempo discreto é dado por

$$h_p(k_1, \dots, k_p) = (T_s)^p \tilde{g}_p(k_1 T_s, \dots, k_p T_s) \quad (3.162)$$

onde $\tilde{g}_p$ é tal que sua FRFG satisfaz

$$\tilde{G}_p(j\Omega_1, \dots, j\Omega_p) = W(j\Omega_1) \dots W(j\Omega_p) \tilde{H}_p(j\Omega_1, \dots, j\Omega_p) \quad (3.163)$$

sendo $\tilde{H}_p$ a FRFG do núcleo de tempo contínuo que se deseja discretizar. Note que, para se determinar h_p desta forma, seria necessário calcular $\tilde{g}_p$ a partir da relação equivalente a (3.163) no domínio do tempo

$$\tilde{g}_p(\tau_1, \dots, \tau_p) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} w(\sigma_1) \dots w(\sigma_p) \tilde{h}_p(\tau_1 - \sigma_1, \dots, \tau_p - \sigma_p) d\sigma_1 \dots d\sigma_p \quad (3.164)$$

em que $w(t)$ é a anti-transformada de $W(j\Omega)$, ou seja [57]

$$w(t) = \frac{\Omega_s}{2\pi} \text{sinc}(\Omega_N t). \quad (3.165)$$

Na prática, admitindo que $\tilde{h}_p$ contém pouca energia após o instante de tempo T , então h_p pode ser aproximado computando-se a transformada de Fourier discreta inversa de

$$\tilde{V}_p(k_1, \dots, k_p) = \tilde{H}_p \left(j\Omega_s \frac{\bar{k}_1}{N_p}, \dots, j\Omega_s \frac{\bar{k}_p}{N_p} \right) \quad (3.166)$$

com $k_1, \dots, k_p \in \{0, \dots, N_p - 1\}$ e

$$\bar{k}_i = \begin{cases} k_i & 0 \leq k_i \leq \frac{N_p}{2} \\ k_i - N_p & \frac{N_p}{2} < k_i < N_p \end{cases} \quad (3.167)$$

sendo $N_p = \lceil T\Omega_s/2\pi \rceil$.

Entretanto, é importante observarmos que isso corresponde a amostrar a convolução circular dos núcleos com a função sinc no domínio do tempo, e portanto haverá uma porção não-causal dos núcleos aparecendo em amostras de h_p com índices próximos a N_p (i.e., um rebatimento no domínio do tempo), o que pode degradar a aproximação [57].

Uma forma de se atenuar este efeito seria, ao invés de se utilizar a janela retangular, empregar uma janela senoidal [57], como a de Hamming ou a de von Hann, definidas em tempo discreto como, respectivamente, [58]

$$W_H(k) = \begin{cases} 0,54 - 0,46 \cos(2\pi \frac{2k-N_p}{2N_p}), & 0 \leq k < N_p \\ 0, & \text{caso contrário} \end{cases} \quad (3.168)$$

$$W_{vH}(k) = \begin{cases} 0,5 - 0,5 \cos(2\pi \frac{2k-N_p}{2N_p}), & 0 \leq k < N_p \\ 0, & \text{caso contrário} \end{cases} . \quad (3.169)$$

Assim, (3.166) seria substituída por

$$\tilde{V}_p(k_1, \dots, k_p) = \left[\prod_{m=1}^p W_s(k_m) \right] \tilde{H}_p \left(j\Omega_s \frac{\bar{k}_1}{N_p}, \dots, j\Omega_s \frac{\bar{k}_p}{N_p} \right) \quad (3.170)$$

em que W_s é a janela senoidal escolhida. Com isso, podem-se reduzir as oscilações que aparecem em regiões de índices próximos a N_p e correspondem à porção não-causal que resulta da convolução circular com a função sinc.

Para finalizar, deve-se notar que, se for válida a condição

$$|\Omega_i| > \frac{\pi}{T_s} \text{ para algum } i \in \{1, \dots, p\} \implies \tilde{H}_p(j\Omega_1, \dots, j\Omega_p) = 0 \quad (3.171)$$

então as FRFGs de tempo discreto obtidas através das duas abordagens apresentadas, i.e., resultantes das expressões (3.158) e (3.160), são equivalentes. De fato, vê-se que, como tal condição é imposta a (3.163) com o janelamento no domínio da frequência, em (3.162) realiza-se a discretização de sua anti-transformada através da primeira abordagem sem que haja rebatimento de componentes de frequência.

3.2 Modelos físicos

3.2.1 Definição

Os modelos físicos (MF) são representações matemáticas paramétricas cuja topologia é estabelecida com base em conhecimento prévio sobre a estrutura interna do sistema de interesse ou sobre sua operação, sendo fundamentada em leis de campos da física como a mecânica clássica ou o eletromagnetismo. Logo, naturalmente, eles são formulados em tempo contínuo, uma vez que tais leis também o são.

Do ponto de vista adotado neste trabalho, em que os modelos são representações matemáticas de operadores, os modelos físicos podem ser definidos como segue.

Definição 3.9 Seja $T : U \rightarrow Y$ um operador não-linear causal. Um modelo físico de T consiste em um conjunto de equações diferenciais e algébricas conhecido como equações de espaço de estados, cuja forma geral é

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), u(t)) \quad (3.172)$$

$$y(t) = \lambda(t, \mathbf{x}(t), u(t)) \quad (3.173)$$

onde $\mathbf{x} \in \mathbb{R}^n$ é o vetor de estados, com componentes que caracterizam o estado interno do sistema, $\dot{\mathbf{x}}(t)$ denota a derivada de $\mathbf{x}$ em relação a t , $u(t) \in U$, $y(t) \in Y$ e as funções

$$\mathbf{f} : \mathbb{R} \times \mathbb{R}^n \times U \rightarrow \mathbb{R}^n \quad (3.174)$$

$$\lambda : \mathbb{R} \times \mathbb{R}^n \times U \rightarrow Y \quad (3.175)$$

são não-linearidades arbitrárias, possivelmente variantes no tempo. Normalmente, um estado inicial $\mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^n$ é especificado juntamente com o modelo.

Deve-se observar que (3.172) é uma equação diferencial ordinária com vetor de estados de dimensão finita. Portanto, estamos nos restringindo ao caso em que

o sistema tem parâmetros concentrados, ou seja, desprezam-se variações espaciais das variáveis de estado, o que corresponde a uma aproximação da realidade suficiente para muitas aplicações.

Por representar matematicamente as relações que governam o funcionamento interno do sistema, modelos físicos na forma (3.172)–(3.173) apresentam a vantagem de fornecer informações para uma análise detalhada sobre a própria operação dele, sendo úteis para outros fins além da implementação computacional visando simulação ou predição. A possibilidade de interpretação física direta das variáveis permite uma compreensão aprofundada dos fenômenos envolvidos, o que em geral não se aplica aos demais formalismos estudados neste trabalho.

3.2.2 Núcleos de Volterra de modelos físicos

Uma questão importante no estudo de sistemas não-lineares diz respeito a, dado que um modelo físico é conhecido, como se pode obter uma representação de Volterra equivalente. De fato, a derivação de núcleos de Volterra associados a um modelo na forma (3.172)–(3.173) é útil para diversas finalidades, tanto teóricas quanto práticas. No Capítulo 6, ela servirá como base para a avaliação da adequação de diferentes estruturas ao modelamento de um sistema de interesse, o que se enquadra na metodologia mencionada no Capítulo 1, cuja ideia central consiste no uso de um MF. Além disso, a partir dos núcleos de Volterra calculados, pode-se, por exemplo, calcular as FRFGs do sistema, as quais quantificam certas características tipicamente apresentadas por sistemas não-lineares no domínio da frequência, como discutido na Subseção 3.1.3.

É importante ressaltar que, excetuando-se alguns casos, como aquele em que a função $\mathbf{f}$ da equação (3.172) pode ser escrita na forma

$$\mathbf{f}(t, \mathbf{x}(t), u(t)) = \mathbf{A}(t)x(t) + \mathbf{q}(t, u(t)) \quad (3.176)$$

onde $\mathbf{A}(t) \in \mathbb{R} \rightarrow \mathbb{R}^{n \times n}$ e $\mathbf{q} : \mathbb{R} \times U \rightarrow \mathbb{R}^n$ é uma função analítica qualquer, a representação de Volterra associada a um sistema definido por (3.172)–(3.173) possui um número infinito de núcleos. Mais precisamente, as representações de

Volterra (quando elas existem, evidentemente) de sistemas cujas equações diferenciais são não-lineares no vetor de estados possuem um número infinito de núcleos. Logo, em geral apenas uma parte dos núcleos da série é derivada para os fins mencionados no parágrafo anterior. Este aspecto ficará mais claro no Capítulo 5, em que o método de bilinearização de Carleman será explicado em detalhes.

De maneira geral, os métodos desenvolvidos na literatura se aplicam a classes específicas de modelos físicos cujas não-linearidades (i.e., as funções $\mathbf{f}$ e λ da definição 3.9) possuem determinadas características, como a analiticidade. Devido à importância que a derivação dos núcleos de um MF possui no contexto deste trabalho, este assunto será tratado mais detalhadamente no Capítulo 5, em que quatro métodos que foram estudados serão discutidos.

3.2.3 Uso e estimação de modelos físicos

Do ponto de vista do uso prático, para se implementar computacionalmente um MF, é necessário aplicar algum método de integração numérica ou de discretização às equações (3.172) e (3.173). Com isso, pode-se calcular a evolução temporal do vetor de estado e da saída ao longo de uma sequência de instantes de tempo escolhidos. São muitas as referências que tratam de técnicas de discretização e integração numérica de modelos em espaço de estados de tempo contínuo; exemplos incluem [59, 60].

Devido à generalidade de (3.172)–(3.173), pouco pode ser afirmado em relação à estimação dos parâmetros envolvidos em tais equações. Em alguns casos, pode-se encontrar certa facilidade nesta etapa, caso as funções $\mathbf{f}$ e λ possuam forma relativamente simples. Em outros, pode ser necessário empregar um procedimento de otimização computacionalmente custoso e tal que a formulação de seus critérios seja bastante trabalhosa.

3.3 Modelos NARX

3.3.1 Definição

Na Seção 3.1, vimos que se pode pensar na representação de Volterra como uma generalização da integral de convolução que caracteriza a resposta de sistemas lineares. Restringindo-nos aos sistemas lineares de tempo discreto racionais, uma representação que oferece uma alternativa ao somatório de convolução é a equação de diferenças, cujas vantagens em alguns aspectos são bem conhecidas. Assim, é natural nos indagarmos se, a partir da generalização da equação de diferenças, incluindo-se termos não-lineares na entrada e na saída, pode-se obter uma representação para sistemas não-lineares que apresente as mesmas vantagens em relação à representação de Volterra.

A ideia discutida no parágrafo anterior fornece uma motivação para a representação paramétrica denominada NARX, a qual é definida a seguir.

Definição 3.10 Um modelo NARX para um operador não-linear de tempo discreto $T : U \rightarrow Y$ consiste em uma equação de diferenças na forma [38, 61, 62]

$$y(n) = F[y(n-1), \dots, y(n-N_y), u(n), \dots, u(n-N_u+1)] \quad (3.177)$$

em que F é uma função não-linear, N_u é a extensão da memória para as entradas e N_y é a extensão da memória para as saídas, ambas dadas em número de amostras.

O nome atribuído a esta representação se deve ao fato de que ela é uma extensão para o caso não-linear da representação ARX, a qual é empregada no modelamento de sistemas lineares, sendo uma sigla para *Nonlinear Auto-Regressive model with eXogenous inputs* (modelo não-linear auto-regressivo com entradas exógenas). O qualificador “auto-regressivo” se refere à possibilidade de realimentação (i.e., termos envolvendo saídas passadas), enquanto as entradas exógenas correspondem aos termos de u [9].

Como observado em [61], (3.177) é uma formulação bastante geral para a representação de sistemas não-lineares de tempo discreto finitamente realizáveis.

Porém, em termos práticos, o uso de um modelo desta natureza requer que seja escolhida uma forma particular para a não-linearidade F . Uma escolha um tanto óbvia seria simplesmente definí-la como uma função polinomial de ordem finita P , de modo que (3.177) passa a ter a forma

$$y(n) = \sum_{p=1}^P \sum_{\substack{l,k \\ l+k=p}} c_{l,k}(\alpha^{(l)}, \beta^{(k)}) u(n - \alpha_1) \dots u(n - \alpha_l) y(n - \beta_1) \dots y(n - \beta_k) \quad (3.178)$$

onde $c_{l,k}(\cdot)$ são os coeficientes do modelo, $\alpha^{(l)} = (\alpha_1, \dots, \alpha_l)$ e $\beta^{(k)} = (\beta_1, \dots, \beta_k)$ são multi-índices e o último somatório é sobre todos os multi-índices $\alpha^{(l)}$ e $\beta^{(k)}$ tais que seus componentes satisfaçam $\alpha_j \in \{0, \dots, N_u - 1\}$ e $\beta_j \in \{1, \dots, N_y\}$, sendo N_u a memória em relação à entrada e N_y a memória em relação à saída.

Para uma melhor compreensão da forma (3.178), é útil recorrermos a um exemplo concreto.

Exemplo 3.1

Consideremos um modelo NARX com $P = 3$, $N_y = 3$ e $N_u = 2$ tal que apenas os coeficientes $c_{2,1}(0, 1, 1) = 0,5$, $c_{0,2}(1, 1) = -1$ e $c_{2,1}(1, 1, 3) = 0,155$ sejam não-nulos. Então, a partir de (3.178) tem-se a equação de diferenças não-linear

$$y(n) = 0,5u(n)u(n-1)y(n-1) - y^2(n-1) + 0,155u^2(n-1)y(n-3).$$

□

Deve-se notar que outras escolhas são possíveis para a forma de F , como a racional [38]. Contudo, a forma polinomial é predominante na definição de modelos NARX, uma vez que, por ser linear nos parâmetros, ela possibilita o uso de técnicas como aquelas discutidas na Subseção 2.6.1 para a estimação deles. De fato, na literatura, quando a representação NARX é mencionada, em geral refere-se à forma (3.178).

Outra observação importante é que, em alguns trabalhos, acrescentam-se ao modelo NARX termos para o modelamento de ruído, de modo que a representação (estocástica) resultante é denominada de NARMAX, em que as letras adicionadas são as iniciais de *moving average*, ou média móvel. Trata-se de um recurso utilizado para prevenir a polarização das estimativas dos parâmetros do modelo [63] (admitindo-se, evidentemente, que sua estrutura corresponde à realidade), muito embora, na prática, o modelo efetivamente usado corresponda apenas à parte determinística estimada.

3.3.2 Definição da estrutura de um modelo NARX

No Capítulo 2 foi dito que um aspecto central na construção de modelos paramétricos é a definição de uma estrutura apropriada para a representação do sistema de interesse. Em se tratando de modelos NARX na forma (3.178), essa definição envolve a determinação de um grau de não-linearidade para o polinômio, P , das extensões da memória para as entradas e saídas, N_u e N_y , respectivamente, e de quais termos envolvidos no somatório serão não-nulos.

Se há alguma informação a respeito da não-linearidade que descreve a dinâmica do sistema, ela pode ser útil na especificação da estrutura do modelo NARX. Por exemplo, se a estrutura de um modelo físico é conhecida, então pode-se utilizar o método indireto descrito em [9, Cap. 3] para se obter uma equação de diferenças não-linear. Neste caso, o procedimento envolve a derivação analítica de núcleos de Volterra no domínio de Laplace a partir do modelo físico, os quais são discretizados para se obter uma forma concreta para o polinômio multivariável do modelo NARX. Deve-se notar, contudo, que tal método pode ser bastante trabalhoso na prática.

Quando não há disponibilidade de informações prévias, é necessário determinar a estrutura de forma empírica. Para tanto, frequentemente se emprega um procedimento que possibilita detectar a estrutura e estimar os parâmetros de forma iterativa [38, 41]. Basicamente, em cada passo um termo (ou regressor, de acordo com a terminologia da Subseção 2.6.1) do polinômio é escolhido para inclusão no modelo com base em um critério estatístico e seu respectivo coefi-

ciente é estimado. Tal critério, chamado de taxa de redução de erro (ERR, do inglês *error reduction ratio*), é definido como a correlação entre os resíduos (i.e., a parte da saída ainda não explicada pelo modelo) e um regressor candidato, de maneira que se escolhe em cada iteração aquele regressor que apresentar a maior ERR [38]. Assim, novos termos são incorporados no modelo até que ele seja considerado suficiente segundo alguma medida estabelecida, como o erro quadrático médio por ele cometido.

Para que o procedimento descrito no parágrafo anterior seja aplicado, deve-se escolher previamente o conjunto de regressores candidatos. Normalmente, isto é feito definindo-se os valores de P , N_u e N_y , de modo que o procedimento na realidade é utilizado apenas para determinar quais termos do polinômio serão não-nulos e estimar seus coeficientes. A partir da equação (3.178), pode-se ver que a cardinalidade deste conjunto de regressores candidatos cresce muito rapidamente com P , N_u ou N_y . Como em cada iteração do procedimento se computa a ERR de cada regressor ainda não incluído, o custo computacional envolvido para a determinação do modelo NARX por esta abordagem pode ser consideravelmente elevado [9].

Outro problema apresentado pelo procedimento acima descrito é que, na presença de ruído, ele pode conduzir a um modelo impreciso [9]. Para se reduzir a degradação das estimativas em tais condições, uma alternativa seria considerar também termos de ruído, obtendo-se, assim, um modelo NARMAX. Nesse caso, pode-se empregar uma versão estendida do procedimento discutido em que, após se obter um modelo inicial contendo apenas termos envolvendo entradas e saídas, termos para o modelamento de ruído são incluídos também iterativamente, utilizando-se os resíduos obtidos na determinação do modelo inicial para formar os regressores envolvendo amostras de ruído [41].

3.3.3 Relação com filtros de Volterra

No início desta seção, mencionou-se que uma motivação para considerarmos equações de diferenças não-lineares é a possibilidade de se obter uma representação que apresente as mesmas vantagens em relação à série de Volterra que

a equação de diferenças linear apresenta em relação à resposta impulsiva (ou função-peso) de um sistema linear e invariante no tempo (LIT). No que segue, apresentaremos um argumento informal que dá suporte a essa motivação.

Em se tratando de sistemas LIT de tempo discreto, um fato conhecido é que a representação através da equação de diferenças — ou, equivalentemente, através da função de transferência racional — é muito mais compacta do que a resposta impulsiva de um sistema. Aliás, no caso de sistemas físicos, os quais em geral apresentam resposta impulsiva com duração infinita (IIR, do inglês *infinite impulse response*), uma representação exata através da função-peso necessitaria que um número infinito de coeficientes fosse determinado, o que evidentemente não é factível. Ao invés disso, usa-se como aproximação uma resposta de duração finita (FIR, do inglês *finite impulse response*). Mesmo assim, geralmente tal aproximação é constituída por um número de coeficientes bem maior do que aqueles que compõem a equação de diferenças do sistema.

Um argumento intuitivo por vezes empregado para explicar esse fato é que na equação de diferenças amostras de entradas e saídas passadas são usadas para compor a expressão da saída em um dado instante, ao passo que o somatório de convolução faz uso apenas de amostras da entrada em instantes anteriores. Fazendo um paralelo, podemos ver que o mesmo ocorre com a relação entrada/saída de um FV, dada por (3.135), que envolve apenas amostras passadas do sinal de entrada. Por este motivo, a equação de diferenças não-linear de um modelo NARX normalmente é constituída por um número bem menor de parâmetros quando comparada com um modelo de Volterra convencional. Em outras palavras, assim como ocorre no caso linear, a introdução de realimentação no modelo faz com que a representação se torne muito mais compacta.

A rigor, a existência de amostras da saída no lado direito de (3.178) faz com que o sistema possua memória infinita (analogamente aos filtros LIT IIR), de modo que para se obter um FV equivalente seria necessário considerar uma versão de (3.135) em que os núcleos possuam duração infinita. Adicionalmente, por incluir produtos cruzados entre entradas e saídas, o modelo NARX pode apresentar também grau de não-linearidade infinito. Deve-se notar, contudo, que

os coeficientes que atuam sobre os termos de alta ordem introduzidos por tais produtos estão restritos a potências (ou produtórios) dos parâmetros do modelo.

Para exemplificar, consideremos um modelo NARX na forma

$$y(n) = c_1 u(n-1) + c_2 u(n-1)y(n-1).$$

Utilizando recursivamente a própria expressão para expandir o termo do lado direito, temos

$$\begin{aligned} y(n) &= c_1 u(n-1) + c_1 c_2 u(n-1)u(n-2) + c_2^2 u(n-1)u(n-2)y(n-2) \\ &= c_1 u(n-1) + c_1 c_2 u(n-1)u(n-2) + c_1 c_2^2 u(n-1)u(n-2)u(n-3) \\ &\quad + c_2^3 u(n-1)u(n-2)u(n-3)y(n-3) \end{aligned}$$

e assim por diante. Se procedermos dessa forma, pode-se perceber que a cada substituição a ordem do polinômio é aumentada, de modo que não há expressão com grau de não-linearidade truncada do tipo (3.135) que seja exatamente equivalente a este modelo.

Observe, porém, que os coeficientes que multiplicam os termos das equações se restringem a produtórios de c_1 e c_2 . Consequentemente, se considerarmos um modelo de Volterra de tempo discreto equivalente (o que requer uma versão de (3.135) infinita em memória e em grau de não-linearidade), então seus núcleos terão um número reduzido de graus de liberdade, o que faz com que a representação da série de Volterra de tempo discreto (com memória e ordem infinitas) compreenda a representação induzida por um modelo NARX como caso especial.

Do ponto de vista prático, os aspectos mencionados nos parágrafos anteriores constituem de fato uma vantagem dos modelos NARX em relação a filtros de Volterra convencionais, pois estes são sempre implementados com memória e número de núcleos finitos. Por outro lado, uma consequência negativa da inclusão de saídas passadas na expressão de um modelo é que torna-se possível que ele se comporte de forma instável caso saídas produzidas pelo próprio modelo sejam utilizadas, como discutido na Seção 2.4. Para que não haja problema nesse sentido, seria necessário, então, o uso de saídas medidas no lado direito de (3.178),

as quais podem conter incertezas e ruído de medida, afetando as previsões feitas pelo modelo. Além disso, em algumas aplicações, simplesmente pode não se ter acesso a saídas medidas, como ocorre no problema de CEA⁹ (ver Subseção 1.1.2), de modo que a garantia de estabilidade de um modelo NARX passa a ser necessária para que ele possa ser empregado na prática.

Observe que, para FVs, não há razão para preocupação em relação à questão de estabilidade, pois núcleos com memória finita são sempre absolutamente somáveis, satisfazendo uma condição de estabilidade no sentido BIBO análoga àquela descrita na Subseção 3.1.1.3 para núcleos de tempo contínuo. Já no caso da classe NARX, não há uma caracterização de estabilidade simples e diretamente verificável a partir da equação de diferenças. Desse modo, caso seja necessário utilizar as saídas produzidas pelo próprio modelo para realimentação, a escolha de um modelo NARX traz a necessidade de um estudo aprofundado para verificar suas propriedades de estabilidade sob as condições de operação consideradas.

3.4 Modelos modulares

A obtenção de modelos que capturem comportamentos dinâmicos não-lineares arbitrários e sejam ao mesmo tempo convenientes para uso prático é uma tarefa geralmente difícil, como foi discutido no Capítulo 1. Enquanto formalismos gerais como os modelos de Volterra convencionais frequentemente conduzem a representações excessivamente complexas, o uso de modelos paramétricos como os NARX ou os modelos físicos pode ser dificultado ou até mesmo inviabilizado pela necessidade de se determinar uma estrutura adequada antes da estimação dos parâmetros ou por questões de estabilidade.

Numa tentativa de se obter um compromisso entre a abordagem paramétrica e a não-paramétrica, foram propostos diversos modelos orientados a bloco (MOB) que se enquadram na chamada abordagem modular [9]. Trata-se de modelos

⁹Embora o sinal captado pelo microfone mostrado na Figura 1.2 contenha a saída de interesse, a presença de outros elementos nele, como o sinal local $d(n)$ e ruído de medição, inviabiliza seu uso na realimentação de um modelo não-linear.

cujas estruturas são compostas pela interconexão de blocos geralmente simples e relativamente compactos, como modelos lineares e não-linearidades sem memória [64].¹⁰ Com isso, busca-se, em linhas gerais, uma alternativa que permita construir modelos mais compactos que os não-paramétricos sem que sejam necessários o uso de realimentação e/ou o estabelecimento de uma estrutura especializada para cada sistema, como tipicamente ocorre na modelagem paramétrica.

As estruturas modulares empregadas na literatura variam bastante em complexidade e em capacidade de representação. Uma maneira de caracterizar cada uma delas segundo este último aspecto é derivando-se um modelo de Volterra equivalente [9], uma vez que a forma dos núcleos obtidos e o número de graus de liberdade neles envolvidos fornecem um indicativo da generalidade da estrutura em questão. Esta estratégia é baseada na capacidade apresentada pela representação de Volterra de aproximar uma ampla classe de sistemas, como discutido na Seção 3.1.

Normalmente, a adoção de MOBs com capacidade de representação limitada é motivada pela simplicidade apresentada por tais modelos e pela relativa facilidade envolvida na estimação de seus parâmetros. Contudo, é evidente que, se uma estrutura deste tipo é adotada para o modelamento de um sistema de interesse, a dinâmica dele deve possuir algumas características específicas, ao menos aproximadamente, para que não haja prejuízo em relação à precisão do modelo obtido. Embora em muitos trabalhos se mencione que algumas de tais estruturas, como os modelos N-L e L-N, são adequadas para o modelamento de alguns sistemas reais – especialmente nas áreas de sistemas biológicos, processos químicos industriais e processamento de sinais [64, 65] –, são poucas as referências em que isso é de fato demonstrado, seja de forma prática (exemplos, nesse caso, incluem [66]) ou teórica.

Já o uso de MOBs mais gerais, como o de Wiener, é motivado pelo fato de que, com frequência, eles são capazes de aproximar precisamente operadores

¹⁰Deve-se lembrar que uma não-linearidade é dita sem memória quando o resultado de sua aplicação sobre o sinal de entrada para um instante de tempo depende apenas do valor desse sinal naquele mesmo instante, enquanto não-linearidades com memória dependem de amostras da entrada no instante atual e também em instantes anteriores (admitindo-se que elas são causais).

não-lineares através de estruturas muito mais compactas que filtros de Volterra convencionais, sobretudo quando se modela um sistema físico. De fato, o modelo de Wiener apresenta capacidade de representação equivalente aos filtros de Volterra, como foi demonstrado em [55], desde que sua estrutura possua complexidade suficiente. Porém, em se tratando do modelamento de sistemas físicos, em geral pode-se abrir mão de alguns graus de liberdade em favor da compacidade do modelo sem que haja perda significativa de precisão. Esta ideia será discutida com mais detalhes no Capítulo 4.

No restante desta seção, descreveremos algumas estruturas comumente usadas na literatura, especialmente em aplicações de PDS. Primeiramente, serão apresentados os MOBs com capacidade de representação mais restrita, característica que será demonstrada através da derivação de núcleos de modelos de Volterra equivalentes. Deve-se notar que estes resultados já foram mencionados [67] e demonstrados na literatura [9], à exceção daquele relativo ao modelo de Hammerstein generalizado. Contudo, optou-se por demonstrá-los aqui também para se fornecer uma justificativa concreta em relação às limitações mencionadas e também para possibilitar uma comparação mais clara entre tais estruturas e aquelas que serão apresentadas no próximo capítulo. Por fim, o modelo de Wiener será descrito, sendo fornecidos argumentos em relação à generalidade de tal formalismo.

3.4.1 Modelo N-L ou de Hammerstein

O modelo N-L ou de Hammerstein, que corresponde a um dos MOBs mais simples, é constituído por um bloco não-linear sem memória em cascata com um sistema LIT [38, 65, 68], como representado na Figura 3.3. Observe que a saída $w(n)$ da não-linearidade F é fornecida como entrada para o bloco LIT, representado pela função de transferência racional $G(z)$. Embora esteja representado no diagrama, o sinal $w(n)$ não é acessível para fins de estimação do modelo, sendo apenas conhecidas a entrada e a saída.

Matematicamente, a relação entrada/saída de um modelo de Hammerstein

Figura 3.3: Estrutura de um modelo N-L ou de Hammerstein.

causal é dada por

$$y(n) = \sum_{k=0}^{\infty} g(k)w(n-k) \quad (3.179)$$

$$= \sum_{k=0}^{\infty} g(k)F[u(n-k)] \quad (3.180)$$

onde $g(n)$ é a resposta impulsiva que caracteriza o sistema LIT $G(z)$. Normalmente, costuma-se expressar $F(x)$ através de uma combinação linear de funções de base $f_i(x)$, resultando na forma

$$y(n) = \sum_{k=0}^{\infty} g(k) \left\{ \sum_{i=0}^P \alpha_i f_i[u(n-k)] \right\} \quad (3.181)$$

$$= \sum_{i=0}^P \alpha_i \sum_{k=0}^{\infty} g(k) f_i[u(n-k)] \quad (3.182)$$

em que α_i , $i = 0, \dots, P$ são os coeficientes da expansão. Uma escolha comum para as funções de base é $f_i(x) = x^i$, ou seja, considera-se uma expansão polinomial truncada da não-linearidade.

Admitindo-se que $G(z)$ é uma função racional que corresponde à equação de diferenças

$$y(n) = \sum_{k=0}^M b_k w(n-k) - \sum_{k=1}^N a_k y(n-k) \quad (3.183)$$

então a relação entrada/saída resultante tem a forma

$$y(n) = \sum_{i=0}^P \alpha_i \sum_{k=0}^M b_k f_i[u(n-k)] - \sum_{k=1}^N a_k y(n-k). \quad (3.184)$$

Quando tanto F quanto G são desconhecidos, como o modelo é não-linear nos parâmetros α_i , b_k e a_k , não há solução analítica geral para a estimação deles. Nesse caso, para a estimação *offline* a partir de um conjunto de dados, um método

bastante utilizado na literatura consiste em estimar os parâmetros de um dos blocos enquanto o outro é considerado constante, procedendo-se de forma iterativa até que as estimativas convirjam [69]. Em [68], mostra-se que tal método converge para valores que minimizam o erro quadrático médio cometido pelo modelo, desde que seja feita uma normalização dos parâmetros α_i ou dos parâmetros b_k em cada passo para evitar problemas de identificabilidade (causados, basicamente, pela ambiguidade dos valores de tais parâmetros em relação aos valores dos produtos $\alpha_i b_k$) e desde que a entrada $u(n)$ empregada para produzir o conjunto de dados seja suficientemente complexa.

Em relação à capacidade de representação, a estrutura de Hammerstein é severamente limitada, como será demonstrado a seguir.

Teorema 3.7 [9] Seja $\mathcal{H}$ um modelo de Hammerstein de tempo discreto com não-linearidade analítica $F(x)$ e bloco LIT estável $G(z)$. Então, os núcleos $\{h_p\}_{p \in \mathbb{N}}$ de uma série de Volterra equivalente a $\mathcal{H}$ devem satisfazer, para $p > 0$,

$$h_p(\tau_1, \dots, \tau_p) \neq 0 \implies \tau_1 = \tau_2 = \dots = \tau_p. \quad (3.185)$$

Em outras palavras, os núcleos do operador de Volterra equivalente são diagonais.

Demonstração. Como $F(x)$ é analítica, podemos representá-la usando a série de Taylor convergente em torno de $x = 0$

$$F(x) = \sum_{p=0}^{\infty} \alpha_p x^p \quad (3.186)$$

em que os coeficientes da expansão são dados por

$$\alpha_p = \frac{1}{p!} \left. \frac{d^p F}{dx^p} \right|_{x=0}. \quad (3.187)$$

Usando (3.182) sob as suposições de que $P \rightarrow \infty$ e $f_p(x) = x^p$, podemos escrever a saída do modelo como

$$y(n) = \sum_{p=0}^{\infty} \alpha_p \sum_{k=0}^{\infty} g(k) u^p(n-k). \quad (3.188)$$

Igualando-se tal relação entrada/saída com a expressão de uma série de Volterra de tempo discreto

$$y(n) = h_0 + \sum_{p=1}^{\infty} \sum_{k_1=0}^{\infty} \sum_{k_2=k_1}^{\infty} \dots \sum_{k_p=k_{p-1}}^{\infty} h_p(k_1, \dots, k_p) u(n - k_1) \dots u(n - k_p) \quad (3.189)$$

podemos facilmente ver que, fazendo

$$h_0 = \alpha_0 \sum_{n=0}^{\infty} g(n) \quad (3.190)$$

e definindo-se os núcleos diagonais

$$h_p(k_1, \dots, k_p) = \begin{cases} \alpha_p g(k), & \text{se } k = k_1 = k_2 = \dots = k_p \\ 0, & \text{caso contrário} \end{cases} \quad (3.191)$$

a relação (3.189) resulta em (3.188).

■

O resultado acima indica que há duas fortes limitações na capacidade de representação de um modelo de Hammerstein: seus núcleos podem ter coeficientes não-nulos apenas na diagonal principal e as diagonais principais dos núcleos de diferentes ordens são proporcionais, i.e.,

$$h_p(k, \dots, k) = \frac{\alpha_p}{\alpha_q} h_q(k, \dots, k) \quad (3.192)$$

para $0 \leq k \leq \min\{N_p - 1, N_q - 1\}$. Em relação à primeira delas, deve-se notar que, embora a energia de cada núcleo associado a um sistema físico tenda a se concentrar em uma região próxima à sua diagonal principal, conforme será discutido na Seção 4.4, a possibilidade de existência de coeficientes não-nulos apenas em tal diagonal implica na aplicabilidade de tal modelo a uma classe reduzida de sistemas.

3.4.2 Modelo de Hammerstein generalizado

Se generalizarmos (3.181) de modo que a cada função de base da expansão escolhida $F(x) = \sum \alpha_i f_i(x)$ seja associado um bloco LIT próprio, $G_i(z)$, então uma certa flexibilidade é acrescentada ao modelo, que passa a se chamar modelo de Hammerstein generalizado [67] ou modelo de Uryson [70]. Quando a expansão da não-linearidade é polinomial, tem-se então o modelo de Hammerstein polinomial generalizado [23], o qual é também denominado *power filter* no caso em que cada filtro linear da estrutura possui resposta impulsiva de duração finita [71].

A expressão geral de um modelo de Hammerstein generalizado é

$$y(n) = \sum_{i=0}^P \sum_{k=0}^{\infty} g_i(k) w_i(n-k) \quad (3.193)$$

$$= \sum_{i=0}^P \alpha_i \sum_{k=0}^{\infty} g_i(k) f_i[u(n-k)]. \quad (3.194)$$

Supondo novamente que as funções $G_i(z)$, $i = 0, \dots, P$, são dadas por

$$G_i(z) = \frac{\sum_{k=0}^{M_i} b_{i,k} z^{-k}}{1 + \sum_{k=1}^{N_i} a_{i,k} z^{-k}} \quad (3.195)$$

então temos um modelo resultante na forma

$$y(n) = \sum_{i=0}^P \alpha_i y_i(n) \quad (3.196)$$

em que

$$y_i(n) = \sum_{k=0}^{M_i} b_{i,k} f_i[u(n-k)] - \sum_{k=1}^{N_i} a_{i,k} y_i(n-k). \quad (3.197)$$

Um diagrama de blocos correspondente a essa definição é apresentado na Figura 3.4. Note que, analogamente ao sinal $w(n)$ da Figura 3.3, os sinais $w_i(n)$, $i = 0, \dots, P$ não são medidos durante a operação do sistema.

Em [23], descreve-se um método para a estimação de modelos de Hammerstein polinomial generalizados em que os coeficientes α_i são todos iguais a um (o que não introduz perda de generalidade) e cujos filtros $G_i(z)$ possuem resposta impulsiva de duração finita (ou seja, filtros FIR, do inglês *finite impulse response*). Em

Figura 3.4: Estrutura de um modelo de Hammerstein generalizado ou modelo de Uryson.

linhas gerais, o sistema de interesse é excitado com uma varredura senoidal exponencial (i.e., uma senóide cuja frequência varia exponencialmente com o tempo) e, subsequentemente, a saída produzida é filtrada por um filtro FIR construído a partir da própria sequência de excitação. Devido às propriedades da excitação e do filtro utilizados, o sinal de tempo discreto resultante contém “cópias” das respostas impulsivas g_i , $i = 0, \dots, P$, espalhadas ao longo de sua duração, a qual, se for suficientemente longa em relação ao decaimento das respostas g_i , garantirá que a superposição delas não será significativa.

De forma semelhante ao que foi feito para o modelo de Hammerstein simples, pode-se mostrar que a estrutura generalizada também possui capacidade de representação bastante limitada. Este resultado será demonstrado a seguir.

Teorema 3.8 Seja $\mathcal{H}_g$ um modelo de Hammerstein generalizado de tempo discreto tal que as funções base da expansão escolhida para a não-linearidade, que denotaremos por $f_i(x)$, com $i = 0, \dots, M$, são analíticas, e seus respectivos blocos LIT associados, $G_i(z)$, são estáveis. Então, a série de Volterra que estabelece uma relação entrada/saída equivalente à de $\mathcal{H}_g$ é tal que seus núcleos, $\{h_p\}_{p \in \mathbb{N}}$, são diagonais.

Demonstração. Inicialmente, lembremos que o modelo de Hammerstein generalizado utiliza uma expansão da não-linearidade na forma

$$F(x) = \sum_{i=0}^M \alpha_i f_i(x). \quad (3.198)$$

Então, seguindo o mesmo raciocínio do Teorema 3.7, podemos utilizar a analiticidade das funções f_i para expandí-las em torno de $x = 0$, obtendo

$$f_i(x) = \sum_{p=0}^{\infty} \beta_{i,p} x^p \quad (3.199)$$

em que

$$\beta_{i,p} = \frac{1}{p!} \left. \frac{d^p f_i}{dx^p} \right|_{x=0}. \quad (3.200)$$

A partir da expressão dada em (3.194), podemos então escrever

$$y(n) = \sum_{i=0}^M \alpha_i \sum_{p=0}^{\infty} \beta_{i,p} \sum_{k=0}^{\infty} g_i(k) u^p(n-k) \quad (3.201)$$

onde $g_i(k)$ é a resposta impulsiva do filtro $G_i(z)$. Novamente, comparando-se tal relação com a expressão de uma série de Volterra de tempo discreto dada por (3.189), temos que com

$$h_0 = \sum_{i=0}^M \alpha_i \beta_{i,0} \sum_{n=0}^{\infty} g_i(n) \quad (3.202)$$

$$h_p(k_1, \dots, k_p) = \begin{cases} \sum_{i=0}^M \alpha_i \beta_{i,p} g_i(k), & \text{se } k = k_1 = k_2 = \dots = k_n \\ 0, & \text{caso contrário} \end{cases} \quad (3.203)$$

obtemos uma representação de Volterra equivalente a $\mathcal{H}_g$ com núcleos diagonais.

■

É importante observar que, embora os núcleos de um modelo de Hammerstein generalizados também estejam sujeitos à condição de diagonalidade, nesse caso a diagonal de cada núcleo é dada por uma combinação linear das respostas impulsivas $g_i(n)$ com coeficientes próprios $\alpha_i \beta_{i,p}$. Logo, em conjunto, seus núcleos possuem mais graus de liberdade do que aqueles de um modelo de Hammerstein simples.

3.4.3 Modelo L-N

Invertendo-se a ordem dos blocos do modelo de Hammerstein, obtém-se outra estrutura denominada modelo L-N, o qual é bastante mencionado na literatura [9], sendo algumas vezes chamado de modelo de Wiener [38, 65, 72]. Contudo, o uso desta última denominação pode causar confusão, uma vez que, na realidade, o modelo em blocos formulado por Norbert Wiener possui uma forma mais geral do que aquela que com frequência é apresentada [9], como ficará claro na Subseção 3.4.5.

O diagrama de blocos de um modelo L-N é mostrado na Figura 3.5, onde os elementos exibidos são os mesmos da Figura 3.3, porém conectados de forma inversa. Empregando-se uma expansão finita para a não-linearidade, assim como feito em (3.181), obtém-se a relação entrada/saída

$$y(n) = F \left[\sum_{k=0}^{\infty} g(k)u(n-k) \right] \quad (3.204)$$

$$= \sum_{i=0}^P \alpha_i f_i \left[\sum_{k=0}^{\infty} g(k)u(n-k) \right]. \quad (3.205)$$

Note que, caso $G(z)$ possua forma racional como aquela da equação (3.195), então temos

$$y(n) = \sum_{i=0}^P \alpha_i f_i \left[\sum_{k=0}^M b_k u(n-k) - \sum_{k=1}^N a_k w(n-k) \right]. \quad (3.206)$$

Figura 3.5: Estrutura de um modelo L-N.

Para a estimação dos parâmetros de um modelo L-N, existem vários métodos disponíveis na literatura. Dentre eles, podem citados aqueles descritos em [65, 72, 73].

Assim como ocorre com as estruturas de Hammerstein descritas, os modelos L-N possuem capacidade de representação limitada. Este aspecto será esclarecido pelo teorema a seguir.

Teorema 3.9 [9] Seja $\mathcal{L}$ um modelo L-N de tempo discreto descrito por (3.205), com não-linearidade $F(x)$ analítica e filtro LIT $G(z)$ estável. Então, a relação entrada/saída associada a $\mathcal{L}$ é equivalente a uma representação de Volterra com núcleos $\{h_p\}_{p \in \mathbb{N}}$ tais que

$$h_p(k_1, \dots, k_p) = \beta_p \prod_{i=1}^p v(k_i) \quad (3.207)$$

para uma constante $\beta_p \in \mathbb{R}$ e uma sequência $v : \mathbb{Z}_+ \rightarrow \mathbb{R}$.

Demonstração. Expressando a não-linearidade em função de sua série de Taylor, assim como foi feito nos Teoremas 3.7 e 3.8, obtemos através de (3.205) a relação entrada/saída

$$y(n) = \sum_{p=0}^{\infty} \alpha_p \left[\sum_{k=0}^{\infty} g(k) u(n-k) \right]^p \quad (3.208)$$

em que α_p é o p -ésimo coeficiente da série de Taylor de F . De imediato, verifica-se que todos os termos de ordem p de tal expansão têm a forma

$$\alpha_p \prod_{i=1}^p g(k_i) u(n-k_i), \quad \text{com } k_i \geq 0. \quad (3.209)$$

A partir de (3.209), pode-se ver que uma série de Volterra equivalente a (3.208) é obtida definindo-se os núcleos simétricos

$$h_0 = \alpha_0 \quad (3.210)$$

$$h_p(k_1, \dots, k_p) = \alpha_p g(k_1) g(k_2) \dots g(k_p), \quad \forall k_1, \dots, k_p \in \mathbb{Z}_+. \quad (3.211)$$

Assim, a condição (3.207) se aplica a todos os núcleos se definirmos, $\forall p \in \mathbb{N}^*$, $\beta_p = \alpha_p$ e $v(k) = g(k)$, com $k \geq 0$.

■

De acordo com o resultado obtido, é nítido que há poucos graus de liberdade envolvidos na definição dos núcleos de um modelo de Volterra equivalente a um L-N. Uma interpretação interessante que dá uma ideia sobre esta limitação pode ser feita da seguinte forma. Considerando-se o caso em que o modelo tem me-

mória finita de M amostras, o núcleo de segunda ordem h_2 do filtro de Volterra equivalente pode ser representado pela matriz¹¹ $M \times M$

$$\mathbf{H}_2 = \alpha_p \begin{bmatrix} g^2(0) & g(0)g(1) & \dots & g(0)g(M-1) \\ g(0)g(1) & g^2(1) & \dots & g(1)g(M-1) \\ \vdots & \vdots & \ddots & \vdots \\ g(0)g(M-1) & g(1)g(M-1) & \dots & g^2(M-1) \end{bmatrix}. \quad (3.212)$$

Visivelmente, $\mathbf{H}_2$ tem posto apenas unitário, e portanto suas colunas (ou linhas) pertencem a um subespaço unidimensional do $\mathbb{R}^M$.

3.4.4 O modelo L-N-L

Outra estrutura modular, a qual foi proposta visando o modelamento de comportamentos um pouco mais complexos do que aqueles capturados pelos modelos L-N e N-L, consiste em uma combinação destes, sendo denominada modelo L-N-L ou, como também é (inadequadamente) chamada, modelo de Wiener-Hammerstein [74, 75]. No diagrama de blocos que a ilustra, mostrado na Figura 3.6, observa-se que a não-linearidade nesse caso é colocada entre dois filtros LIT.

Figura 3.6: Estrutura de um modelo L-N-L.

Matematicamente, a relação entrada/saída do modelo L-N-L é dada por

$$y(n) = \sum_{k=0}^{\infty} g_2(k) w_2(n-k) \quad (3.213)$$

$$= \sum_{k=0}^{\infty} g_2(k) F[w_1(n-k)] \quad (3.214)$$

$$= \sum_{k=0}^{\infty} g_2(k) F \left[\sum_{l=0}^{\infty} g_1(l) u(n-k-l) \right]. \quad (3.215)$$

¹¹Na Subseção 4.4.1, uma definição mais formal desta associação entre um núcleo de segunda ordem e uma matriz será dada.

Empregando-se uma expansão na forma $F(x) = \sum_{i=0}^P \alpha_i f_i(x)$ para a não-linearidade, (3.215) resulta em

$$y(n) = \sum_{i=0}^P \alpha_i \sum_{k=0}^{\infty} g_2(k) f_i \left[\sum_{l=0}^{\infty} g_1(l) u(n-k-l) \right]. \quad (3.216)$$

Admitindo-se que os filtros $G_1(z)$ e $G_2(z)$ são racionais, sendo dados por expressões na forma (3.195), então tem-se

$$y(n) = \sum_{i=0}^P \alpha_i \sum_{k=0}^{M_2} b_{2,k} f_i \left[\sum_{l=0}^{M_1} b_{1,l} u(n-k-l) - \sum_{l=1}^{N_1} a_{1,l} w_1(n-k-l) \right] - \sum_{k=1}^{N_2} a_{2,k} y(n-k). \quad (3.217)$$

Exemplos de trabalhos que tratam da estimação de modelos com estrutura L-N-L incluem [73, 75]. Embora tal estrutura possua uma generalidade maior quando comparada com as anteriormente apresentadas, ainda assim há restrições em relação aos operadores que podem ser representados por ela. Esse aspecto será tratado no teorema que segue.

Teorema 3.10 [9] Seja $\mathcal{L}_N$ um modelo L-N-L de tempo discreto composto por uma não-linearidade $F(x)$ analítica e blocos LIT $G_1(z)$ e $G_2(z)$ estáveis. Então, existe uma série de Volterra de tempo discreto equivalente à relação entrada/saída de $\mathcal{L}_N$ com núcleos $\{h_p\}_{p \in \mathbb{N}}$ na forma

$$h_p(k_1, \dots, k_p) = \sum_{k=0}^{\infty} s_{p,2}(k) \prod_{i=1}^p s_1(k_i - k) \quad (3.218)$$

para $k_1, \dots, k_p \in \mathbb{Z}_+$.

Demonstração. Novamente, vamos começar pela expansão de F em sua série de Taylor em torno de $x = 0$. Utilizando (3.216) com $P \rightarrow \infty$ e $f_p(x) = x^p$, temos

$$y(n) = \sum_{p=0}^{\infty} \alpha_p \sum_{k=0}^{\infty} g_2(k) \left[\sum_{l=0}^{\infty} g_1(l) u(n-k-l) \right]^p \quad (3.219)$$

em que α_p , $p \in \mathbb{Z}_+$, são os coeficientes da expansão. Observando-se a relação

(3.219), fica claro que os termos de ordem p da resposta possuem a forma

$$\alpha_p g_2(k) \prod_{i=1}^p g_1(l_i) u(n - k - l_i) \quad (3.220)$$

com $l_i \geq 0$, $i \in \{1, \dots, p\}$. Deve-se notar também que todas as combinações de k com os índices l_i que resultam nos mesmos valores $k_i = k + l_i$ multiplicam o mesmo produto de entradas $\prod_{i=1}^p u(n - k_i)$ na expressão da saída. Em outras palavras, o núcleo da representação de Volterra resultante pode ser escrito como

$$h_0 = \alpha_0 \sum_{k=0}^{\infty} g_2(k) \quad (3.221)$$

$$h_p(k_1, \dots, k_p) = \sum_{k=0}^{\infty} \alpha_p g_2(k) \prod_{i=1}^p g_1(k_i - k), \quad \forall k_1, k_2, \dots, k_p \in \mathbb{Z}_+. \quad (3.222)$$

Assim definindo $s_1(k) = g_1(k)$ e $s_{p,2}(k) = \alpha_p g_2(k)$, vê-se que de fato os núcleos têm a forma apresentada em (3.218).

■

Ao contrário do que ocorre para o modelo L-N, o núcleo quadrático de um modelo de L-N-L com memória finita pode, em sua forma mais geral, ser associado a uma matriz de posto completo. Isso decorre da possibilidade de se ponderarem os produtos dos termos $g_1(k_i - k)$ por valores da sequência g_2 . Assim, como esperado, a adição de outro bloco LIT em um modelo L-N após a não-linearidade sem memória aumenta a flexibilidade da estrutura resultante.

Por outro lado, em virtude de os núcleos resultarem da forma convolutiva (3.222), eles possuem regiões com coeficientes necessariamente nulos quando o comprimento de $g_1(n)$ é finito e o comprimento de $g_2(n)$ é maior que um. Mais precisamente, é necessariamente nulo todo coeficiente $h_p(k_1, \dots, k_p)$ cuja distância para a diagonal principal, definida como $\max\{k_1, \dots, k_p\} - \min\{k_1, \dots, k_p\}$, é maior ou igual ao comprimento de $g_1(n)$. Isto pode ser compreendido se notarmos que, de acordo com (3.222), $h_p(k_1, \dots, k_p)$ resulta de uma superposição ponderada de núcleos na forma $g_1(k_1) \dots g_1(k_p)$ deslocados ao longo da direção da diagonal principal [76].

Para exemplificarmos de forma concreta a propriedade discutida no parágrafo anterior, consideremos um modelo L-N-L em que $\alpha_2 = 1$ e as respostas impulsivas $g_1(n)$ e $g_2(n)$ são dadas na forma vetorial por

$$\mathbf{g}_1 = [g_1(0) \ g_1(1)]^T \quad (3.223)$$

$$\mathbf{g}_2 = [g_2(0) \ g_2(1)]^T. \quad (3.224)$$

Com base na expressão (3.222), temos que a representação matricial do núcleo de segunda ordem possui a forma

$$\begin{aligned} \mathbf{H}_2 &= \begin{bmatrix} g_2(0)g_1^2(0) & g_2(0)g_1(0)g_1(1) & 0 \\ g_2(0)g_1(0)g_1(1) & g_2(0)g_1^2(1) + g_2(1)g_1^2(0) & g_2(1)g_1(0)g_1(1) \\ 0 & g_2(1)g_1(0)g_1(1) & g_2(1)g_1^2(1) \end{bmatrix} \quad (3.225) \\ &= g_2(0) \begin{bmatrix} g_1^2(0) & g_1(0)g_1(1) & 0 \\ g_1(0)g_1(1) & g_1^2(1) & 0 \\ 0 & 0 & 0 \end{bmatrix} + g_2(1) \begin{bmatrix} 0 & 0 & 0 \\ 0 & g_1^2(0) & g_1(0)g_1(1) \\ 0 & g_1(0)g_1(1) & g_1^2(1) \end{bmatrix}. \end{aligned} \quad (3.226)$$

Desse modo, vê-se claramente que, devido à presença do filtro $G_2(z)$, $\mathbf{H}_2$ pode ser vista como uma soma de duas matrizes, cada uma delas contendo um bloco proporcional a $\mathbf{g}_1\mathbf{g}_1^T$, deslocado de acordo com o respectivo índice de k da forma convolutiva (3.222). Com isso, a memória do núcleo resultante é superior à duração de $g_1(n)$. No entanto, $h_2(0, 2)$ e $h_2(2, 0)$ são necessariamente nulos, visto que ambos possuem distância para a diagonal principal de $\mathbf{H}_2$ igual a dois, e portanto nenhuma das parcelas da soma (3.226) contribuirá com valores não-nulos nas posições de tais coeficientes devido ao comprimento de $g_1(n)$ ser também igual a dois.

Além da limitação acima descrita, no modelo L-N-L há uma relação de interdependência entre os coeficientes de todos os seus núcleos, uma vez que eles resultam de produtórios de coeficientes das mesmas respostas impulsivas g_1 e g_2 . Por exemplo, conforme apontado em [9], devido à causalidade dos filtros $G_1(z)$ e

$G_2(z)$, os valores do núcleo h_p sobre o plano em que $k_1 = 0$ satisfazem

$$h_p(0, k_2, \dots, k_p) = \alpha_p g_2(0) g_1(0) g_1(k_2) \dots g_1(k_p) \quad (3.227)$$

uma vez que $g_1(0 - k)$ é não-nulo apenas para $k = 0$. Mas, para qualquer q inteiro positivo, tem-se também que

$$h_{p+q}(0, \dots, 0, k_2, \dots, k_p) = \alpha_{p+q} g_2(0) [g_1(0)]^{q+1} g_1(k_2) \dots g_1(k_p) \quad (3.228)$$

de forma que os valores admitidos por h_{p+q} sobre este domínio são proporcionais àqueles de h_p sobre o plano $k_1 = 0$.

3.4.5 O modelo de Wiener

Se generalizarmos a estrutura L-N de modo que ela seja constituída por um número finito qualquer de filtros lineares em paralelo seguidos de uma não-linearidade sem memória multivariada, como representado na Figura 3.7, tem-se como resultado o que se chama de modelo de Wiener. O nome dado a esta estrutura decorre do fato de que, em seu estudo seminal sobre o modelamento caixa-preta de sistemas não-lineares, Wiener propôs um modelo geral de tempo contínuo que consistia em um banco de filtros ortonormais (como, por exemplo, filtros de Laguerre) cujas saídas eram fornecidas a uma função não-linear multivariada representada por uma expansão envolvendo produtos de polinômios de Hermite [9, 37]. Fazendo uso deste arranjo, Wiener reduziu o problema de identificação de um sistema não-linear à determinação dos coeficientes que caracterizam a expansão da não-linearidade.

Embora tal estratégia desenvolvida por Wiener não tenha sido muito utilizada para a identificação de sistemas não-lineares por ser pouco prática, a concepção do modelo descrito no parágrafo anterior foi uma das importantes contribuições dadas por este cientista [9]. Isto se deve ao fato de que sua estrutura possui poder de representação tão amplo quanto o das séries de Volterra truncadas.

Em [55], a generalidade do modelo de Wiener é estabelecida formalmente,

Figura 3.7: Estrutura de um modelo de Wiener.

estendendo-se os Teoremas 3.5 e 3.6, que tratam, respectivamente, do caso em tempo contínuo e do caso em tempo discreto. Em ambas extensões, demonstra-se que, se um operador não-linear tem memória desvanescente, então existe uma série de Volterra truncada realizada através de uma estrutura de Wiener com não-linearidade polinomial que é capaz de aproximá-lo com precisão arbitrária. Trata-se de um resultado muito importante e bastante explorado por diversos pesquisadores com o intuito de formular modelos compactos que aproximem filtros de Volterra convencionais com precisão, assunto que será tratado no próximo capítulo. Por ora, ao invés de reproduzirmos aqui as demonstrações dos teoremas mencionados, as quais podem ser encontradas em [55], forneceremos argumentos intuitivos sobre a validade deste resultado.

O primeiro argumento, válido para modelos em tempo contínuo, é baseado em [9, Cap. 2]. Nesse caso, se considerarmos uma estrutura análoga àquela da Figura 3.7, então os sinais $w_i(t)$, $i \in \{1, \dots, M\}$, são dados por

$$w_i(t) = \int_0^\infty g_i(\tau) u(t - \tau) d\tau. \quad (3.229)$$

Assim, definindo F como uma não-linearidade polinomial de grau P , temos

$$y(t) = \sum_{p=1}^P \sum_{j_1=1}^M \dots \sum_{j_p=1}^M a_p(j_1, \dots, j_p) w_{j_1}(t) \dots w_{j_p}(t) \quad (3.230)$$

Usando (3.229), resulta

$$y(t) = \sum_{p=1}^P \int_0^\infty \dots \int_0^\infty \sum_{j_1=1}^M \dots \sum_{j_p=1}^M a_p(j_1, \dots, j_p) \\ \times g_{j_1}(\tau_1) \dots g_{j_p}(\tau_p) u(t - \tau_1) \dots u(t - \tau_p) d\tau_1 \dots d\tau_p. \quad (3.231)$$

Então, ao definirmos

$$h_p(\tau_1, \dots, \tau_p) = \sum_{j_1=1}^M \dots \sum_{j_p=1}^M a_p(j_1, \dots, j_p) g_{j_1}(\tau_1) \dots g_{j_p}(\tau_p) \quad (3.232)$$

a relação entrada/saída resultante passa a consistir em uma série de Volterra truncada na ordem P com núcleos simétricos dados por (3.232). Assim, pode-se ver que se o conjunto de filtros g_i oferecer um nível adequado de graus de liberdade, então pode-se aproximar um núcleo qualquer com precisão arbitrária através da expansão (3.232).

No caso de tempo discreto, o argumento é ainda mais simples. Na Subseção 3.4.3, viu-se que o núcleo de segunda ordem da série de Volterra equivalente à relação entrada/saída de um modelo L-N pode ser representado pela matriz de posto unitário

$$\mathbf{H}_2 = \alpha \mathbf{g} \mathbf{g}^T \quad (3.233)$$

em que α é um coeficiente da expansão de Taylor da não-linearidade e $\mathbf{g}$ é a representação vetorial do filtro linear. Empregando a mesma ideia que foi usada na demonstração do Teorema 3.9, pode-se mostrar que o núcleo de segunda ordem associado a um filtro de Volterra equivalente a tal modelo quando são acrescentados mais $M - 1$ filtros lineares antes da não-linearidade sem memória possui a forma

$$h_2(k_1, k_2) = \sum_{i=1}^M \sum_{j=1}^M \alpha_{i,j} g_i(k_1) g_j(k_2) \quad (3.234)$$

para $k_1, k_2 \in \{0, 1, \dots, N - 1\}$, em que os coeficientes $\alpha_{i,j}$ são dados pela expansão de Taylor da não-linearidade do modelo. Consequentemente, tal núcleo pode ser

representado matricialmente por

$$\mathbf{H}_2 = \sum_{i=1}^M \sum_{j=1}^M \alpha_{i,j} \mathbf{g}_i \mathbf{g}_j^T \quad (3.235)$$

em que $\mathbf{g}_i$ é a representação vetorial do filtro $g_i(n)$, $i \in \{1, \dots, M\}$. Observando a expressão (3.235), conclui-se que qualquer núcleo simétrico de segunda ordem de posto M pode ser por ela representado, desde que o conjunto $\{\mathbf{g}_1, \dots, \mathbf{g}_M\}$ seja linearmente independente, cada vetor $\mathbf{g}_i$ possua ao menos um coeficiente associado $\alpha_{i,j}$ não-nulo e $\forall i, j \in \{1, \dots, M\}, \alpha_{i,j} = \alpha_{j,i}$. Em outras palavras, a inclusão de outros filtros lineares permite que o núcleo resultante possua uma forma mais geral.¹² Para núcleos de ordem $p > 2$, um argumento semelhante pode ser utilizado, ou seja, graus de liberdade podem ser adicionados com a inclusão de filtros lineares na estrutura de Wiener. Assim, fica evidente que uma série de Volterra truncada de tempo discreto pode ser aproximada com precisão arbitrária por um modelo deste tipo.

Se nos restringirmos aos filtros de Volterra com memória finita de M amostras, uma estrutura de Wiener que fornece a saída de um FV de forma *exata* pode ser obtida definindo-se os filtros lineares simplesmente como

$$G_i(z) = z^{-i+1}, \quad i = 1, \dots, M \quad (3.236)$$

e a não-linearidade F como um polinômio em $u(n), u(n-1), \dots, u(n-M+1)$ com coeficientes extraídos dos núcleos de tempo discreto de tal FV. De fato, no teorema sobre a aproximação de operadores de tempo discreto através de modelos de Wiener demonstrado em [55], a estrutura considerada possui esta última forma descrita.

Deve-se ressaltar que os argumentos acima visam fornecer uma ideia intuitiva sobre a validade da equivalência entre séries de Volterra truncadas e modelos de Wiener em termos de aproximação da relação entrada/saída de modelos físicos com memória desvanescente. Em particular, observe que esta última forma dis-

¹²Deve-se observar que o argumento fornecido é válido ainda que se considere uma matriz com número infinito de linhas e colunas correspondente a um núcleo com duração infinita. Neste caso, evidentemente, os filtros lineares teriam também dimensão infinita.

cutida, em que os filtros de tempo discreto são simplesmente elementos de atraso, corresponde a uma realização direta de um FV convencional com memória finita, apresentando complexidade estrutural semelhante a um filtro deste tipo.

Na realidade, o interesse prático no uso do modelo de Wiener se deve ao fato de que ele possibilita, muitas vezes, que boas aproximações sejam obtidas mesmo com um número reduzido de parâmetros em comparação com um filtro de Volterra convencional [9, 29], desde que os filtros lineares e a não-linearidade sem memória sejam determinados apropriadamente (de fato, a escolha (3.236) não é adequada para tal finalidade). Se este for o caso, abre-se mão de alguns graus de liberdade em favor da compacidade da estrutura sem que a perda de precisão traga prejuízo em relação ao objetivo com que o modelo será empregado. Por exemplo, uma aproximação do núcleo cuja representação matricial é dada em (3.235) pode ser obtida simplesmente reduzindo-se o posto de tal matriz através do uso de um número de filtros lineares $M' < M$, o que pode ser suficiente para o desempenho desejado.

Finalmente, no caso em tempo contínuo, um outro aspecto do modelo de Wiener que pode ser vantajoso na prática é a própria facilidade de realização. De fato, trata-se de uma estrutura mais facilmente implementável do que uma série de Volterra truncada, cujos núcleos variam continuamente em função de diversos índices.

Capítulo 4

Estruturas de Volterra alternativas

Conforme discutido no capítulo anterior, os filtros de Volterra são capazes de aproximar uma ampla classe de sistemas físicos. No entanto, o número de coeficientes envolvidos em um modelo deste tipo cresce muito rapidamente tanto com as extensões de memória de seus núcleos quanto com sua ordem ou grau de não-linearidade, como se pode verificar através da seguinte expressão dada na Subseção 3.1.10

$$\Psi_{\text{FV}}(\mathcal{H}) = \sum_{p=1}^P \frac{(N_p + p - 1)!}{p!(N_p - 1)!} \quad (4.1)$$

onde P é a ordem do FV $\mathcal{H}$ e N_p é a extensão da memória do núcleo de ordem p , $p \in \{1, \dots, P\}$. Logo, em situações nas quais é necessário um modelo com ordem elevada e/ou com memória extensa para que o nível de precisão desejado seja atingido, o uso de um FV passa a ser desvantajoso ou até mesmo inviável.

Esta característica de baixa compacidade dos FVs é amplamente conhecida, sendo mencionada em muitos trabalhos [27, 28, 29, 77]. De forma análoga ao que ocorre com a aproximação de funções através de polinômios, ela constitui um obstáculo particularmente severo para o modelamento de sistemas fortemente não-lineares, fato que também é bastante comentado na literatura. Contudo, quando se requer um alto nível de precisão mediante a presença de entradas com potência elevada, o uso de FVs pode ser inviável mesmo para sistemas fracamente não-lineares.

Devido às dificuldades acima discutidas, diversos pesquisadores têm buscado estruturas derivadas do filtro de Volterra que forneçam um compromisso entre precisão e compacidade. De modo geral, tais estruturas de Volterra alternativas (EVA) são obtidas introduzindo-se alguma aproximação nos núcleos que permita

reduzir consideravelmente o número de coeficientes envolvidos na representação. A motivação básica dessa ideia é que, embora um núcleo de alta ordem resida em um espaço vetorial de dimensão bastante elevada, frequentemente ele contém algum tipo de redundância, especialmente quando está relacionado a um sistema físico, e/ou regiões de coeficientes que não contribuem significativamente para a saída. Sendo assim, em princípio, é possível representá-lo, ou ao menos aproximá-lo com bastante precisão, através de estruturas de dimensão menor que capturem sua informação essencial.

Neste capítulo, descreveremos de forma breve algumas EVAs propostas recentemente, as quais podem ser úteis na busca por uma abordagem apropriada para o modelamento de um sistema não-linear no contexto de aplicações de PDS. O foco desta apresentação residirá na estratégia subjacente a cada uma das estruturas propostas, nas condições sob as quais tais estratégias são adequadas e na redução de complexidade que cada uma delas proporciona. Dessa forma, embora sejam importantes na prática, métodos de estimação dos parâmetros das EVAs a partir de medições e estratégias de adaptação não fazem parte do escopo deste capítulo, cujo propósito é somente apontar algumas estruturas propostas que, sob determinadas condições, se mostraram promissoras para a obtenção de um bom compromisso entre precisão e complexidade computacional.

Antes de apresentarmos tais EVAs, descreveremos, para fins de comparação, duas formas de implementação usualmente empregadas para um filtro de Volterra convencional: a direta e a baseada na TFD.

4.1 Implementação de filtros de Volterra convencionais

4.1.1 Implementação direta

No Capítulo 3 (mais precisamente, na equação (3.135) da Subseção 3.1.10), viu-se que a relação entrada/saída de um FV convencional de ordem P com

núcleos triangulares tem a forma

$$y(n) = \sum_{p=1}^P \sum_{k_1=0}^{N_p-1} \dots \sum_{k_p=k_{p-1}}^{N_p-1} h_p(k_1, \dots, k_p) u(n - k_1) \dots u(n - k_p). \quad (4.2)$$

Uma maneira direta de se implementar um filtro dado por (4.2), frequentemente citada na literatura [2, 25, 28], consiste em computar, para cada instante de tempo n , o somatório

$$y(n) = \sum_{p=1}^P \mathbf{h}_p^T \mathbf{u}_p(n) \quad (4.3)$$

em que cada vetor $\mathbf{u}_p(n)$ contém como componentes todos os possíveis produtos de ordem p envolvendo as amostras $u(n), u(n-1), \dots, u(n - N_p + 1)$ e os vetores $\mathbf{h}_p(n)$ contém os coeficientes do núcleo de p -ésima ordem, ordenados analogamente aos componentes de $\mathbf{u}_p(n)$ que multiplicam. Assim, para $p = 1$ tem-se

$$\mathbf{u}_1(n) = [u(n) \ u(n-1) \ \dots \ u(n - N_1 + 1)]^T \quad (4.4)$$

$$\mathbf{h}_1 = [h_1(0) \ h_1(1) \ \dots \ h_1(N_1 - 1)]^T \quad (4.5)$$

enquanto, para $p = 2$,

$$\mathbf{u}_2(n) = \begin{bmatrix} u^2(n) \\ u(n)u(n-1) \\ \vdots \\ u(n)u(n - N_2 + 1) \\ u^2(n-1) \\ u(n-1)u(n-2) \\ \vdots \\ u^2(n - N_2 + 1) \end{bmatrix}, \quad \mathbf{h}_2 = \begin{bmatrix} h_2(0,0) \\ h_2(0,1) \\ \vdots \\ h_2(0, N_2 - 1) \\ h_2(1,1) \\ h_2(1,2) \\ \vdots \\ h_2(N_2 - 1, N_2 - 1) \end{bmatrix} \quad (4.6)$$

e assim por diante. Observe que estamos considerando vetores $\mathbf{u}_p(n)$ não-redundantes, isto é, tais que há exatamente um componente para cada possível produto de ordem p envolvendo as amostras $u(n), u(n-1), \dots, u(n - N_p + 1)$. No vetor $\mathbf{u}_2(n)$ mostrado em (4.6), por exemplo, não há um componente igual a $u(n-1)u(n)$

após aquele dado por $u(n)u(n - N_2 + 1)$, uma vez que ele seria igual ao segundo elemento do vetor. Correspondentemente, os vetores $\mathbf{h}_p$ possuem exatamente um componente para cada coeficiente de $h_p(k_1, \dots, k_p)$ tal que seus índices satisfazem a condição de triangularidade $k_i \leq k_{i+1}$, $i \in \{1, \dots, p-1\}$, em consonância com a formulação (4.2) e com o que foi discutido na Subseção 3.1.10.

O esquema descrito acima é uma generalização da forma direta de se implementar um filtro linear FIR, a qual corresponde ao caso particular de (4.2) em que $P = 1$. Assim como ocorre neste caso, o uso de (4.2) na implementação de um FV de ordem $P > 1$ é custoso computacionalmente quando a extensão da memória do modelo é relativamente longa. Na realidade, tal ineficiência é ainda mais problemática no caso não-linear, uma vez que a complexidade computacional associada ao cálculo do componente de ordem p da saída é $O(N_p^p)$ por amostra, crescendo, portanto, exponencialmente com p .

4.1.2 Implementação baseada na TFD

Como alternativa para o esquema de implementação direto descrito na subseção anterior, podem-se computar as convoluções multidimensionais envolvidas na expressão (4.2) com o uso da transformada de Fourier discreta dos núcleos, a qual foi descrita na Subseção 3.1.10.1. Novamente, trata-se de uma extensão para o caso não-linear de uma técnica bastante empregada na implementação de filtros lineares, a qual consiste em decompor uma convolução linear em diversas convoluções circulares entre sequências obtidas a partir da resposta impulsiva do filtro e de blocos sucessivos disjuntos de amostras da entrada. Assim, os resultados de tais convoluções, que podem ser eficientemente computadas no domínio da TFD, são combinados com o uso de um esquema *overlap-add* ou *overlap-save* [58].

Em se tratando de filtros de Volterra, a implementação baseada na TFD consiste em computar, para cada componente de ordem p da saída, uma sequência de convoluções circulares multidimensionais que, ao serem somadas, o compõem. Inicialmente, define-se um tamanho para os blocos de entradas, M . Em seguida, admitindo-se que $u(n)$ possui comprimento dado por LM , para algum $L \in \mathbb{N}^*$,

consideram-se L seqüências de comprimento $M + N_p - 1$ definidas como

$$\tilde{u}_l(n) = \begin{cases} u(lM + n), & n \in \{0, \dots, M - 1\} \\ 0, & n \in \{M, \dots, M + N_p - 2\} \end{cases} \quad (4.7)$$

com $l \in \{0, \dots, L - 1\}$. Assim, a partir da TFD,

$$\tilde{H}_p(m_1, \dots, m_p) = \sum_{k_1=0}^{M+N_p-2} \dots \sum_{k_p=0}^{M+N_p-2} \tilde{h}_p(k_1, \dots, k_p) e^{-j \frac{2\pi}{M+N_p-1} (m_1 k_1 + \dots + m_p k_p)} \quad (4.8)$$

em que

$$\tilde{h}_p(k_1, \dots, k_p) = \begin{cases} h_p(k_1, \dots, k_p), & \forall i \in \{1, \dots, p\}, k_i \in \{0, \dots, N_p - 1\} \\ 0, & \exists i \in \{1, \dots, p\} \text{ tal que } k_i \in \{N_p, \dots, M + N_p - 2\} \end{cases} \quad (4.9)$$

o componente $y_p(n)$ pode ser reescrito com o uso das relações

$$y_p(n) = \sum_{l=0}^{L-1} s_l(n) \quad (4.10)$$

$$s_l(lM + n) = \begin{cases} \tilde{s}_l(n, \dots, n), & n \in \{0, \dots, M + N_p - 2\} \\ 0, & \text{caso contrário} \end{cases} \quad (4.11)$$

onde $\tilde{s}_l(n_1, \dots, n_p)$ é a anti-transformada de comprimento $M + N_p - 1$

$$\tilde{s}_l(n_1, \dots, n_p) = \text{TFD}^{-1} \left[\tilde{H}_p(m_1, \dots, m_p) \tilde{U}_l(m_1) \dots \tilde{U}_l(m_p) \right] \quad (4.12)$$

sendo $\tilde{U}_l(m)$ a TFD de $\tilde{u}_l(n)$ de mesmo comprimento.

Em linhas gerais, se a implementação em tempo real de um FV é feita como descrito acima, então os blocos de amostra da saída, $s_l(n)$, vão sendo computados no domínio da frequência à medida em que novos blocos de entradas de comprimento M são disponibilizados. Dessa forma, por um lado, o uso de algoritmos de transformada rápida pode possibilitar a redução do número total de operações realizadas para o cálculo da saída. Por outro lado, há um atraso introduzido devido ao esquema de processamento em blocos, o que em algumas

aplicações pode ser inconveniente ou até mesmo inaceitável. Já em situações nas quais não existe a necessidade de cômputo em tempo real, sendo conhecido previamente um conjunto completo de Q amostras de entrada, não há esse problema e, adicionalmente, o cálculo no domínio da frequência pode ser feito usando-se as relações (4.7)–(4.12) com $L = 1$ e $M = Q$ (i.e., considerando-se um único bloco que contempla todo o conjunto de amostras da entrada).

Observe que a complexidade computacional da implementação descrita é, para cada bloco de tamanho M , da mesma ordem do algoritmo que computa a transformada rápida multidimensional, ou seja, $O((M + N_p)^p \log_2(M + N_p)^p)$ [78]. Já a implementação direta, quando empregada no cômputo de um bloco de M amostras da saída, requer um número de operações $O(MN_p^p)$. Admitindo que $M = N_p = N$, temos então que a complexidade da implementação via TFD é $O(2^p N^p \log_2(2N)^p)$ enquanto a direta possui complexidade $O(N^{p+1})$. Para se ter uma ideia concreta de como tais expressões se comportam comparativamente, calcularam-se valores da expressão

$$C(N, p) = \frac{2^p N^p \log_2(2N)^p}{N^{p+1}} = \frac{p 2^p \log_2(2N)}{N} \quad (4.13)$$

para $p \in \{1, \dots, 5\}$ e $N = 2^m$ com $m \in \{1, \dots, 20\}$, os quais são mostrados no gráfico da Figura 4.1.

Figura 4.1: Valores da função $C(N, p)$, como definida em (4.13), para $p = 1$ (—), $p = 2$ (---), $p = 3$ (■), $p = 4$ (◆) e $p = 5$ (●).

Com base nas curvas mostradas, pode-se ver que, considerando-se uma ordem

fixa p , há vantagem em se empregar o esquema de cômputo no domínio da frequência (o que corresponde, no gráfico, a valores de N e p tais que $C(N, p) < 10^0$) somente se N for suficientemente grande, i.e., maior do que um determinado γ_p . Mais especificamente, pode-se ver que, por exemplo, $\gamma_1 \approx 10$ e $\gamma_2 < 100$, enquanto $\gamma_5 > 1000$. Logo, para que a implementação através da TFD traga um ganho efetivo de eficiência é preciso que a extensão da memória do núcleo de ordem p e/ou o tamanho do bloco sejam suficientemente grandes, comportamento que se acentua com o aumento de p . Por outro lado, se M for muito grande, conforme discutido acima introduz-se um atraso que pode ser inaceitável em muitos casos práticos. Percebe-se, portanto, que há um inevitável compromisso entre a redução de complexidade obtida com o uso das TFDs e o atraso implicado pelo processamento em blocos.

Para finalizar, é importante ressaltar que, obviamente, os valores dados pela função $C(N, p)$ não caracterizam a relação entre as complexidades de forma exata. Ao invés disso, fornecem apenas uma ideia da ordem de grandeza desta relação para diferentes valores de N e p . Por extensão, o mesmo é válido para as constantes γ_p mencionadas no parágrafo anterior.

4.2 Representação dos núcleos em coordenadas diagonais

Uma ideia que tem sido explorada em alguns trabalhos para a implementação eficiente de filtros de Volterra consiste em reescrever a relação entrada/saída convencional dada por (3.135) de modo que o somatório mais interno da convolução multidimensional percorra diagonais dos núcleos [1, 27, 79]. Com isso, a saída do filtro é decomposta em um somatório de convoluções convencionais (unidimensionais) entre tais diagonais e sequências formadas por produtos de amostras da entrada, o que possibilita seu cômputo no domínio da frequência com o uso de TFDs unidimensionais.

Além da possibilidade mencionada no parágrafo anterior, a transformação para coordenadas diagonais permite descartar de forma simples diagonais de um

núcleo afastadas de sua diagonal principal, as quais, para sistemas físicos, costumam ser pouco significativas. Conforme será discutido em mais detalhes na Subseção 4.2.2, isto se deve ao fato de que os coeficientes de tais diagonais ponderam produtos de amostras da entrada relativas a instantes de tempo distantes entre si. Com tal descarte, pode-se reduzir ainda mais o custo computacional envolvido no uso do filtro. No restante desta seção, tais ideias serão discutidas em mais detalhe.

4.2.1 Implementação de filtros de Volterra na representação diagonal

Como foi colocado na seção anterior, a implementação de filtros de Volterra convencionais através da abordagem direta descrita por (4.3) costuma ser computacionalmente custosa. Por outro lado, mencionou-se que a implementação através da TFD multidimensional possui certas inconveniências. Mais especificamente, para que ela seja vantajosa é necessário que a extensão da memória dos núcleos e/ou o comprimento do bloco de amostras considerado na implementação sejam suficientemente grandes. Além disso, este tamanho mínimo necessário cresce com a ordem do núcleo.

Uma outra forma de se escrever a relação (4.2) que possibilita uma implementação no domínio da frequência é como descrita a seguir. Inicialmente, consideremos a transformação dos índices $k_1, \dots, k_p$ dada por

$$k = k_1 \quad (4.14)$$

$$w_i = k_i - k, \quad i = 2, \dots, p \quad (4.15)$$

através da qual (4.2) resulta em [27]

$$y(n) = \sum_{p=1}^P \sum_{w_2=0}^{N_p-1} \dots \sum_{w_p=w_{p-1}}^{N_p-1} \sum_{k=0}^{N_p-w_p-1} h_p(k, w_2+k, \dots, w_p+k) u(n-k) \prod_{i=2}^p u(n-w_i-k). \quad (4.16)$$

Definindo agora as seqüências

$$g_{w_2, \dots, w_p}^{(p)}(k) = h_p(k, w_2 + k, \dots, w_p + k) \quad (4.17)$$

$$u_{w_2, \dots, w_p}^{(p)}(n) = u(n) \prod_{i=2}^p u(n - w_i) \quad (4.18)$$

obtemos a representação [1]

$$y(n) = \sum_{p=1}^P \sum_{w_2=0}^{N_p-1} \dots \sum_{w_p=w_{p-1}}^{N_p-1} \sum_{k=0}^{N_p-w_p-1} g_{w_2, \dots, w_p}^{(p)}(k) u_{w_2, \dots, w_p}^{(p)}(n - k). \quad (4.19)$$

Observe que, na relação dada em (4.19), a saída do FV é decomposta em um somatório de convoluções unidimensionais entre os filtros lineares $g_{w_2, \dots, w_p}^{(p)}$ e as seqüências $u_{w_2, \dots, w_p}^{(p)}$, sendo que cada filtro $g_{w_2, \dots, w_p}^{(p)}$ corresponde a uma diagonal do núcleo de ordem p . Consequentemente, a relação entrada/saída do filtro pode ser escrita no domínio da frequência como [1]

$$Y(e^{j\omega}) = \sum_{p=1}^P \sum_{w_2=0}^{N_p-1} \dots \sum_{w_p=w_{p-1}}^{N_p-1} G_{w_2, \dots, w_p}^{(p)}(e^{j\omega}) U_{w_2, \dots, w_p}^{(p)}(e^{j\omega}) \quad (4.20)$$

em que $Y(e^{j\omega})$, $G_{w_2, \dots, w_p}^{(p)}(e^{j\omega})$ e $U_{w_2, \dots, w_p}^{(p)}(e^{j\omega})$ são as TFTDs da saída, do filtro linear $g_{w_2, \dots, w_p}^{(p)}$ e da seqüência $u_{w_2, \dots, w_p}^{(p)}$, respectivamente. Como observado em [1], (4.20) é uma representação particularmente interessante uma vez que relaciona diretamente o conteúdo espectral da saída com as transformadas das diagonais dos núcleos de Volterra.

Para um melhor entendimento da representação diagonal descrita acima, consideremos, como exemplo, o núcleo de segunda ordem triangular com memória $N_2 = 3$ ilustrado na Figura 4.2. Variando-se o índice w_2 de 0 a 2, ele é decomposto em três filtros lineares com comprimentos decrescentes que correspondem às diagonais formadas pelos coeficientes $h_2(k, w_2 + k)$, com k denotando a posição de um coeficiente dentro da diagonal a que pertence. Na literatura, os índices k e w_2 são chamados de coordenadas diagonais, enquanto aqueles empregados na formulação convencional de um FV, $k_1, \dots, k_p$, são denominados coordenadas cartesianas [27].

Figura 4.2: Representação de um núcleo de segunda ordem na forma triangular através de coordenadas diagonais.

Agora, a partir da formulação (4.19), observe que cada componente de ordem p da saída pode ser escrito como uma superposição de sequências que resultam de cada convolução com uma determinada diagonal, i.e.,

$$y_p(n) = \sum_{w_2=0}^{N_p-1} \dots \sum_{w_p=w_{p-1}}^{N_p-1} y_{w_2, \dots, w_p}^{(d,p)}(n) \quad (4.21)$$

$$y_{w_2, \dots, w_p}^{(d,p)}(n) = \sum_{k=0}^{N_p-w_p-1} g_{w_2, \dots, w_p}^{(p)}(k) u_{w_2, \dots, w_p}^{(p)}(n-k). \quad (4.22)$$

Sendo assim, a mesma estratégia descrita na Subseção 4.1.2 com $p = 1$ pode ser explorada para se calcular cada componente $y_{w_2, \dots, w_p}^{(d,p)}(n)$ a partir das TFDs unidimensionais de sequências construídas estendendo-se com zeros os filtros $g_{w_2, \dots, w_p}^{(p)}(k)$ e blocos de amostras das sequências $u_{w_2, \dots, w_p}^{(p)}(n-k)$. No presente caso, contudo, observe que no cálculo de (4.21) deve-se levar em conta a que diagonal do núcleo cada componente $y_{w_2, \dots, w_p}^{(d,p)}(n)$ se refere, para que eles sejam superpostos corretamente. Mais precisamente, como o filtro envolvido na convolução (4.22) possui comprimento $N_p - w_p$, então, as sequências $\tilde{s}_l(n)$ calculadas no domínio da frequência para compor $y_{w_2, \dots, w_p}^{(d,p)}(n)$, analogamente àquelas dadas por (4.12), possuirão comprimento $M + N_p - w_p - 1$. Logo, é preciso que seus

respectivos blocos, $s_l(n)$, sejam definidos como

$$s_l(lM + n) = \begin{cases} \tilde{s}_l(n - w_p, \dots, n - w_p), & n \in \{w_p, \dots, M + N_p - 2\} \\ 0, & \text{caso contrário} \end{cases} \quad (4.23)$$

Dessa maneira, apesar dos diferentes comprimentos apresentados pelos filtros, pode-se escrever cada componente $y_{w_2, \dots, w_p}^{(d,p)}(n)$ como um somatório na mesma forma de (4.10), sem que o resultado fique inconsistente com (4.19).

Para aplicações que necessitam do uso do FV em tempo real, é evidente que a implementação no domínio da TFD com o uso da representação diagonal traz a desvantagem de introduzir um atraso devido ao processamento de blocos, assim como foi observado em relação ao esquema descrito na Subseção 4.1.2. Porém, deve-se notar que, ao contrário daquela estratégia, nesta se emprega a TFD unidimensional de filtros lineares, o que possibilita estabelecer um tamanho M para o bloco que não implique em um atraso elevado. Isso se deve ao fato de que, como visto, o valor mínimo de M necessário para que haja ganho em termos computacionais com o uso da implementação no domínio da TFD é menor no caso unidimensional (ver Figura 4.1).

Por fim, deve-se notar que, como os filtros envolvidos em (4.19) possuem diferentes comprimentos (os quais são dados por $N_p - w_p$), pode-se optar por empregar ou não o esquema descrito acima para cada uma das convoluções desta expressão, de acordo com o tamanho da resposta do respectivo filtro. Assim, computam-se no domínio da TFD apenas as saídas das diagonais para as quais isso é vantajoso computacionalmente.

4.2.2 Filtros de Volterra diagonais aproximados

Como se mencionou no Capítulo 3, se um sistema físico possui memória desvanescente, então a influência do valor da entrada no instante t_0 sobre a saída $y(t)$ decai com o aumento de $t - t_0$. Logo, para tais sistemas, é de se esperar também que a contribuição da interação entre valores da entrada relativos a instantes de tempo próximos seja, em geral, mais significativa do que aquela dada pela

interação de valores da entrada referentes a instantes de tempo mais distantes.

Se considerarmos a representação de Volterra de um sistema físico com memória desvanescente, a propriedade de localidade temporal das interações entre valores da entrada implica na concentração da energia de cada núcleo em uma região próxima à respectiva diagonal principal. Sob tal condição, uma forma de se reduzir a complexidade computacional envolvida no cálculo da saída de um FV sem que haja perda significativa de precisão seria desconsiderar coeficientes distantes da diagonal principal. Para tanto, a representação diagonal é particularmente conveniente, uma vez que ela pode ser reescrita na forma

$$y(n) = \sum_{p=1}^P \sum_{w_2=0}^{W_p-1} \dots \sum_{w_p=w_{p-1}}^{W_p-1} \sum_{k=0}^{N_p-w_{p-1}-1} g_{w_2, \dots, w_p}^{(p)}(k) u_{w_2, \dots, w_p}^{(p)}(n-k) \quad (4.24)$$

onde $W_p \leq N_p$, de modo que a escolha de W_p determina a distância máxima que uma diagonal do núcleo de ordem p pode apresentar em relação à diagonal principal para que seja levada em conta no cômputo da saída [27]. Note que o conceito de distância aqui empregado é consistente com a definição mencionada na Subseção 3.4.4, de modo que a distância de qualquer coeficiente do filtro diagonal $g_{w_2, \dots, w_p}^{(p)}(k)$ para a diagonal principal do núcleo em que reside é dada por $\max\{k, w_2 + k, \dots, w_p + k\} - \min\{k, w_2 + k, \dots, w_p + k\} = w_p + k - k = w_p$, visto que $0 \leq w_2 \leq \dots \leq w_p$. De acordo com (4.24), tal distância não pode exceder $W_p - 1$, como desejado. No restante do texto, o modelo na forma acima será chamado de filtro de Volterra diagonal aproximado.

É interessante observar que, se tomarmos $W_p = 1$ para $p = 1, \dots, P$, então tal modelo passa a ser composto apenas pela diagonal principal de cada núcleo. Logo, a estrutura resultante possui a mesma capacidade de representação de um modelo do tipo *power filter*, o qual é um caso particular do modelo de Hammerstein generalizado (ver Subseção 3.4.1). Assim, a possibilidade de se variar W_p permite buscar um compromisso entre precisão e complexidade estrutural do FV. Além disso, diante do discutido anteriormente, trata-se de uma maneira de fazê-lo que possui fundamentação física.

Outro comentário importante é que, como visto na Subseção 3.4.4, esta carac-

terística de possuir coeficientes nulos em algumas diagonais afastadas da diagonal principal também ocorre nos núcleos de Volterra equivalentes a uma estrutura L-N-L. Todavia, os núcleos de um filtro de Volterra diagonal aproximado são mais gerais do que aqueles, visto que não estão sujeitos a algumas restrições trazidas pela forma (3.222). Por exemplo, como discutido na Subseção 3.4.4, núcleos associados a um modelo L-N-L possuem interdependências entre si, o que, em geral, não se aplica ao modelo (4.24).

Para que fique mais clara a motivação da aproximação introduzida em (4.24), é útil recorrermos a um exemplo concreto. Consideremos, então, o núcleo de segunda ordem derivado a partir de um modelo físico para alto-falantes, o qual será apresentado na Seção 6.1. Na Figura 4.3, exibem-se os módulos dos coeficientes do núcleo (na forma simétrica), os quais são proporcionais às intensidades luminosas de suas respectivas células. Observe que, à medida em que se afasta da diagonal principal, a magnitude dos valores decai rapidamente até que seja muito pouco significativa. Logo, com o uso da representação (4.24), podemos definir $W_2 \ll N_2$, de modo que o número de convoluções necessárias para o cômputo do componente de segunda ordem da saída seja consideravelmente reduzido. É importante observar que, se utilizássemos a representação convencional (4.2) e simplesmente diminuíssemos N_2 , poderíamos descartar alguns coeficientes cujos módulos são relativamente significativos, como aqueles situados próximos à diagonal principal para índices $k_1 \approx k_2 \approx 75$, o que degradaria a aproximação. No Capítulo 6 será apresentado um estudo de caso sobre a aproximação de um FV que inclui o núcleo mostrado na Figura 4.3 com o uso da representação diagonal (4.24).

4.2.2.1 Complexidades estrutural e computacional

A partir da equação (4.24), podemos caracterizar quantitativamente a redução de complexidade estrutural obtida com o uso da representação diagonal aproximada, como feito a seguir. Primeiro, observe que o número de coeficientes em cada filtro diagonal $g_{w_2, \dots, w_p}^{(p)}$ é dado por $N_p - w_p$. Segundo, para um núcleo de ordem $p = 2$, o número de diagonais com qualquer comprimento é claramente

Figura 4.3: Valores absolutos dos coeficientes de um núcleo de segunda ordem calculado a partir do modelo físico apresentado na Seção 6.1. As células mais claras correspondem aos valores de maior magnitude.

igual a um. Terceiro, para $p \geq 3$, o total de diagonais com comprimento $N_p - w_p$ é igual à cardinalidade do conjunto

$$\{(w_2, \dots, w_{p-1}) \in \mathbb{Z}_+ \mid w_2 \leq w_3 \leq \dots \leq w_{p-1} \leq w_p\} \quad (4.25)$$

uma vez que w_p é fixo para este comprimento e os demais índices w_i , em que $i \in \{2, \dots, p-1\}$, podem variar, contanto que satisfaçam a condição $w_i \leq w_{i+1}$. Sendo assim, tal cardinalidade é idêntica ao número de coeficientes de um núcleo triangular de ordem $p-2$, que corresponde ao total de índices que podem variar, com extensão de memória $w_p + 1$, que é o número máximo de valores que tais índices podem assumir. Logo, para calculá-la, emprega-se uma fórmula semelhante àquela que fornece cada termo de (4.1), i.e.,

$$\binom{w_p + p - 2}{p - 2}. \quad (4.26)$$

Então, conclui-se que o número total de coeficientes de um filtro diagonal aproximado $\mathcal{D}$ dado por (4.24) é

$$\Psi_{\text{diag}}(\mathcal{D}) = N_1 + \sum_{p=2}^P \sum_{w_p=0}^{W_p-1} (N_p - w_p) \binom{w_p + p - 2}{p - 2} \quad (4.27)$$

$$= N_1 + \sum_{p=2}^P \sum_{w_p=0}^{W_p-1} (N_p - w_p) \frac{(w_p + p - 2)!}{(p - 2)!w_p!}. \quad (4.28)$$

Observe que o número de coeficientes para o núcleo de segunda ordem é dado pela mesma fórmula usada para os núcleos de mais alta ordem, o que é válido se admitirmos que $0! = 1$.

Para se ter uma ideia de como a complexidade estrutural quantificada por (4.28) se compara com a de um FV convencional, na Tabela 4.1 são mostrados alguns valores para FVs de ordens $P \in \{1, \dots, 5\}$, considerando-se que todos os núcleos tem a mesma extensão de memória $N_1 = \dots = N_P = N$ e que a largura da representação diagonal empregada também é igual para todas as ordens, i.e., $W_1 = \dots = W_P = W$. Assim, variando-se N , W e P , foram calculados os valores dados por (4.28) e (4.1). Como é de se esperar, nota-se que, com o aumento da ordem e da extensão de memória, o ganho em compacidade obtido com o descarte de diagonais não-significativas cresce rapidamente.

	$N = 30$			$N = 50$			$N = 70$		
P	$W = 10$	$W = 20$	FV	$W = 10$	$W = 20$	FV	$W = 10$	$W = 20$	FV
1	30	30	30	50	50	50	70	70	70
2	285	440	495	505	860	1.325	725	1.280	2.555
3	1.605	4.080	5.455	2.925	8.700	23.425	4.245	13.320	62.195
4	6.720	28.335	46.375	12.440	63.755	316.250	18.160	99.175	1.150.625
5	23.022	159.389	324.631	43.042	371.909	3.478.760	63.062	584.429	17.259.389

Tabela 4.1: Comparação entre os números de coeficientes de filtros de Volterra diagonais e de filtros de Volterra convencionais.

Em relação à complexidade computacional resultante, uma implementação direta das convoluções com os filtros diagonais que fornecem o componente de ordem p da saída requer um número de operações da ordem de $O(D_p)$ por amostra,

em que D_p é o número de coeficientes que compõem o núcleo de ordem p , i.e.,

$$D_p = \sum_{w_p=0}^{W_p-1} (N_p - w_p) \frac{(w_p + p - 2)!}{(p - 2)!w_p!}. \quad (4.29)$$

Já se considerarmos uma implementação das convoluções no domínio da TFD, então tem-se uma complexidade da ordem $O(T_p)$ por bloco de amostras de comprimento M , em que

$$T_p = \sum_{w_p=0}^{W_p-1} \frac{(w_p + p - 2)!}{(p - 2)!w_p!} (M + N_p - w_p) \log(M + N_p - w_p). \quad (4.30)$$

Por fim, como mencionado anteriormente, uma possibilidade interessante seria escolher, para cada convolução de (4.24), o esquema de implementação mais adequado dentre o direto e o baseado na TFD. Considerando-se o aspecto de eficiência computacional, esta escolha deve ser feita, como visto, de acordo com o tamanho do filtro linear envolvido, i.e., $N_p - w_p$. Tal implementação mista pode ser vantajosa se adequadamente projetada, ao custo de um atraso de M amostras introduzido devido ao processamento em blocos.

4.3 Filtro de Volterra com interpolação de núcleos esparsos

Uma outra estratégia efetiva para se obter um ganho em relação à compactidade do modelo, proposta recentemente em [28], consiste em reduzir o número de graus de liberdade envolvidos na representação de cada núcleo, admitindo-se que existe um certo nível de redundância entre coeficientes próximos. Mais especificamente, uma parte dos coeficientes é descartada, de modo que seus termos correspondentes no somatório de (4.2) são aproximados através do uso de um filtro interpolador posicionado antes do FV.

A abordagem acima descrita, chamada em [28] de esquema esparsos com interpolação, é uma extensão para filtros de Volterra de uma técnica empregada na aproximação de filtros digitais com resposta impulsiva de duração finita, cuja

ideia básica é descrita a seguir.

4.3.1 Aproximação de filtros lineares pelo esquema esparsa com interpolação

Seja $h(n)$ um filtro FIR de comprimento N . Para se implementar $h(n)$ pelo esquema esparsa com interpolação, inicialmente deve-se substituí-lo por uma versão esparsa $h^{(s)}(n)$ com fator de interpolação $L \in \mathbb{N}^*$, definida como

$$h^{(s)}(n) = \begin{cases} h(n), & \text{se } \exists k \in \mathbb{N} \text{ tal que } n = kL, \\ 0, & \text{caso contrário} \end{cases}. \quad (4.31)$$

Em seguida, para se reduzir a perda de precisão causada pelo descarte de coeficientes do filtro $h(n)$, introduz-se um filtro interpolador com resposta impulsiva $g(n)$ de comprimento M , de modo que a relação entrada/saída original

$$y(n) = \sum_{k=0}^{N-1} h(k)u(n-k) \quad (4.32)$$

é aproximada por

$$\hat{y}(n) = \sum_{k=0}^{N-1} h^{(s)}(k)\tilde{u}(n-k) \quad (4.33)$$

em que $\tilde{u}(n)$ é a saída do filtro interpolador para a entrada $u(n)$, ou seja,

$$\tilde{u}(n) = \sum_{k=0}^{M-1} g(k)u(n-k). \quad (4.34)$$

Na Figura 4.4, tal esquema é mostrado através de um diagrama de blocos. Conforme indicado, ele equivale à aplicação do filtro de comprimento $M + N - 1$

$$h^{(i)}(n) = \sum_{k=0}^{M-1} g(k)h^{(s)}(n-k) \quad (4.35)$$

à entrada $u(n)$. Assim, os coeficientes de $h(n)$ descartados são “recriados” de forma aproximada devido ao efeito do filtro interpolador, como se pode ver em

(4.35). Observe que, para que haja de fato uma interpolação, é necessário que $L \leq (M+1)/2$, o que garante que cada coeficiente recriado em $h^{(i)}(n)$ contenha a contribuição de dois ou mais coeficientes próximos originais de $h(n)$ (excetuando-se, evidentemente, aqueles situados nas bordas de $h^{(i)}(n)$ [28]).

Figura 4.4: Esquema esparsa com interpolação para a aproximação de um filtro FIR [28].

Exemplo 4.1

Para que se tenha uma ideia concreta do esquema descrito acima, considere o exemplo mostrado na Figura 4.5, em que os filtros $h(n)$, $h^{(s)}(n)$ e $h^{(i)}(n)$ são representados por sequências de amostras cujos valores são dados pela alturas da respectivas hastes. Neste exemplo, emprega-se um fator de interpolação $L = 2$ para se descartar os coeficientes do filtro original de comprimento $N = 5$, de modo que apenas as amostras de $h^{(s)}(n)$ para $n \in \{0, 2, 4\}$ são não-nulas. O filtro de interpolação utilizado $g(n)$ é tal que sua resposta impulsiva pode ser representada pelo vetor

$$\mathbf{g} = [0.5 \quad 1 \quad 0.5]^T. \quad (4.36)$$

Em outras palavras, aplica-se uma interpolação linear de comprimento $M = 3$ na entrada. Como resultado, note que os coeficientes introduzidos no filtro equivalente $h^{(i)}(n)$, representados pelas amostras com círculos não-preenchidos, são dados pela média aritmética entre seus vizinhos. Observe também que, enquanto o quarto coeficiente original de $h(n)$ é reconstruído de forma exata, uma vez que seu valor é igual à media entre os coeficientes vizinhos, o segundo é reconstruído apenas aproximadamente. Além disso, observa-se ainda o surgimento de dois coeficientes adicionais, um no início e outro no final da resposta impul-

siva, o que se deve à convolução $g * h^{(s)}$. Contudo, conforme se comenta em [28], este efeito nas bordas do filtro resultante é pouco significativo quando $M \ll N$, como usualmente ocorre na prática.

Figura 4.5: Efeito do esquema esparsa com interpolação linear de comprimento $M = 3$ sobre um filtro FIR de comprimento $N = 5$, considerando-se um fator $L = 2$.

□

Com o uso da estratégia exposta acima, observa-se que o número de termos não-nulos que compõem o somatório (4.33) é reduzido, em geral, por um fator de aproximadamente L . Logo, a implementação do filtro pode ser feita tirando-se proveito deste fato, ou seja, computando-se na prática a convolução

$$\hat{y}(n) = \sum_{k=0}^{\lceil N/L \rceil - 1} h^{(s)}(kL) \tilde{u}(n - kL). \quad (4.37)$$

Assim, pode-se ver que o custo computacional por amostra envolvido na filtragem também é reduzido pelo mesmo fator mencionado.

Em contrapartida, é evidente que há uma perda de precisão, a qual será tanto menor quanto maior for a capacidade do filtro interpolador de aproximar os coeficientes de $h(n)$ descartados a partir das amostras não-nulas de $h^{(s)}(n)$. Quando se utiliza uma interpolação linear, como aquela do exemplo dado, para se obter uma boa aproximação a resposta impulsiva $h(n)$ deve variar suavemente ou, equivalentemente, ela deve ser tal que coeficientes vizinhos possuam alta correlação.

Uma desvantagem do esquema descrito, relacionada ao seu uso em tempo real, é que, como nesse caso $g(n)$ e $h^{(s)}(n)$ devem ser causais, então normalmente um atraso é introduzido para que a interpolação seja simétrica, como ocorre no exemplo mostrado. Porém, tal consequência negativa é normalmente aceitável

desde que o atraso não seja excessivo, ou seja, desde que o comprimento do filtro interpolador seja relativamente pequeno.

4.3.2 Aproximação de filtros de Volterra pelo esquema esparso com interpolação

Consideremos agora a extensão do esquema descrito na subseção anterior para filtros de Volterra, conforme proposto em [28]. Para simplificar a análise, representaremos a saída de um FV de ordem P na forma vetorial

$$y(n) = \sum_{p=1}^P \mathbf{h}_p^T \mathbf{u}_p(n) = \sum_{p=1}^P \mathbf{u}_p^T(n) \mathbf{h}_p \quad (4.38)$$

em que $\mathbf{u}_1(n)$ e $\mathbf{h}_1$ são definidos como em (4.4) e (4.5) e os vetores $\mathbf{u}_p(n)$ são dados pela fórmula recursiva

$$\mathbf{u}_p(n) = \mathbf{u}_1(n) \otimes \mathbf{u}_{p-1}(n) \quad (4.39)$$

para $p \in \{2, \dots, P\}$, onde $\otimes$ denota o produto de Kronecker (ver Subseção A.1.4 do Apêndice A). Ainda, vamos supor, também por de simplicidade, que as extensões das memórias dos núcleos do FV considerado são todas iguais, i.e., $N_1 = N_2 = \dots = N_P = N$. Note que, por consequência da definição (4.39), os vetores $\mathbf{u}_p(n)$ formam versões redundantes daqueles descritos na Seção 4.1, de modo que $\mathbf{u}_p(n)$ possui dimensão N^p . Por sua vez, os vetores $\mathbf{h}_p$ considerados nesta seção possuem todos os N^p coeficientes que caracterizam o núcleo h_p quando representado na forma simétrica (ver Subseção 3.1.1.1).

Analogamente ao caso linear, a estratégia esparsa com interpolação consiste em construir versões esparsas dos núcleos originais e introduzir um filtro de interpolação que possibilite a reconstrução *ad hoc* dos coeficientes descartados. A construção dos núcleos esparsos $h_p^{(s)}$ a partir dos originais h_p , $p \in \{1, \dots, P\}$, é

feita através da generalização de (4.31), ou seja,

$$h_p^{(s)}(k_1, \dots, k_p) = \begin{cases} h_p(k_1, \dots, k_p), & \text{se } \forall i \in \{1, \dots, p\}, \exists r_i \in \mathbb{N} \text{ tal que } k_i = r_i L \\ 0, & \text{caso contrário} \end{cases} \quad (4.40)$$

onde L é o fator de interpolação escolhido. Assim, denotando por $\mathbf{h}_p^{(s)}$ a representação vetorial do núcleo esparsa de ordem p e por $\tilde{\mathbf{u}}_p(n)$ o vetor de entrada interpolado correspondente, a aproximação de (4.38) através do esquema esparsa com interpolação pode ser expressa na forma

$$\hat{g}(n) = \sum_{p=1}^P \tilde{\mathbf{u}}_p^T(n) \mathbf{h}_p^{(s)}. \quad (4.41)$$

Observe que o mesmo resultado de (4.41) pode ser obtido se eliminarmos de $\mathbf{h}_p^{(s)}$ os componentes correspondentes a coeficientes que são necessariamente nulos devido a seus índices não satisfazerem a condição estabelecida em (4.40) e eliminarmos de $\tilde{\mathbf{u}}_p(n)$ os elementos que seriam multiplicados por tais componentes. Assim, semelhantemente à redução do número de operações obtida na formulação de (4.37), a presença de elementos nulos em $\mathbf{h}_p^{(s)}$ permite computar o produto interno considerando-se, na prática, vetores de dimensão reduzida, o que alivia a carga computacional.

Com o intuito de obter as expressões dos núcleos reconstruídos devido à ação do filtro interpolador $g(n)$ de comprimento M , vamos definir agora a matriz de interpolação de dimensão $(M + N - 1) \times N$ [28]

$$\mathbf{G} = \begin{bmatrix} g(0) & 0 & \dots & 0 \\ g(1) & g(0) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ g(M-1) & g(M-2) & \dots & g(0) \\ 0 & g(M-1) & \dots & g(1) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & g(M-1) \end{bmatrix}. \quad (4.42)$$

Em seguida, definindo-se o vetor de entrada estendido de dimensão $M + N - 1$

$$\mathbf{v}_1(n) = [u(n) \ u(n-1) \ \dots \ u(n-M-N+2)]^T \quad (4.43)$$

podemos escrever os vetores de entrada interpolados na forma vetorial como

$$\tilde{\mathbf{u}}_1(n) = \mathbf{G}^T \mathbf{v}_1(n) \quad (4.44)$$

$$\tilde{\mathbf{u}}_p(n) = \tilde{\mathbf{u}}_1(n) \otimes \tilde{\mathbf{u}}_{p-1}(n), \quad p \in \{2, \dots, P\}. \quad (4.45)$$

Logo, depreende-se de (4.44)–(4.45) que

$$\tilde{\mathbf{u}}_2(n) = \mathbf{G}^T \mathbf{v}_1(n) \otimes \mathbf{G}^T \mathbf{v}_1(n) \quad (4.46)$$

$$= [\mathbf{G}^T \otimes \mathbf{G}^T] [\mathbf{v}_1(n) \otimes \mathbf{v}_1(n)] \quad (4.47)$$

$$= \mathbf{G}_2^T \mathbf{v}_2(n) \quad (4.48)$$

em que $\mathbf{G}_2 = \mathbf{G} \otimes \mathbf{G}$, $\mathbf{v}_2(n) = \mathbf{v}_1(n) \otimes \mathbf{v}_1(n)$ e utilizou-se a propriedade de produto misto apresentada pelo produto de Kronecker (ver Subseção A.1.4 do Apêndice A). Analogamente, pode-se mostrar que, para $p = 3$,

$$\tilde{\mathbf{u}}_3(n) = \mathbf{G}_3^T \mathbf{v}_3(n) \quad (4.49)$$

onde $\mathbf{G}_3 = \mathbf{G} \otimes \mathbf{G}_2$ e $\mathbf{v}_3(n) = \mathbf{v}_1(n) \otimes \mathbf{v}_2(n)$. Generalizando-se tais relações, obtém-se

$$\tilde{\mathbf{u}}_p(n) = \mathbf{G}_p^T \mathbf{v}_p(n) \quad (4.50)$$

$$\mathbf{G}_p = \mathbf{G} \otimes \mathbf{G}_{p-1} \quad (4.51)$$

$$\mathbf{v}_p(n) = \mathbf{v}_1(n) \otimes \mathbf{v}_{p-1}(n) \quad (4.52)$$

para $p \geq 2$. Substituindo (4.50) em (4.41), obtemos, finalmente,

$$\hat{y}(n) = \sum_{p=1}^P \mathbf{v}_p^T(n) \mathbf{G}_p \mathbf{h}_p^{(s)} \quad (4.53)$$

$$= \sum_{p=1}^P \mathbf{v}_p^T(n) \mathbf{h}_p^{(i)} \quad (4.54)$$

em que $\mathbf{G}_1 = \mathbf{G}$ e $\mathbf{h}_p^{(i)} = \mathbf{G}_p \mathbf{h}_p^{(s)}$ é a representação vetorial do núcleo de ordem p reconstruído devido à presença do filtro de interpolação.

Para que tenhamos uma ideia da forma resultante de $\mathbf{h}_p^{(i)}$ e também da mecânica envolvida em sua obtenção, ao invés de derivarmos uma expressão geral a partir das equações acima, o que requer um desenvolvimento bastante laborioso, vejamos um exemplo simples que é fornecido em [28].

Exemplo 4.2

Vamos considerar a implementação de um FV composto apenas por um núcleo de segunda ordem através do esquema descrito. Para tanto, empregaremos o mesmo filtro de interpolação $g(n)$ dado pela equação (4.36), cuja matriz de interpolação associada é

$$\mathbf{G} = \frac{1}{2} \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}. \quad (4.55)$$

Admitindo que a memória do núcleo de segunda ordem tem comprimento $N = 3$, então ele pode ser representado pela matriz¹

$$\mathbf{H}_2 = \begin{bmatrix} h_2(0,0) & h_2(0,1) & h_2(0,2) \\ h_2(1,0) & h_2(1,1) & h_2(1,2) \\ h_2(2,0) & h_2(2,1) & h_2(2,2) \end{bmatrix} \quad (4.56)$$

¹É importante ressaltar que o uso das representações matriciais no exemplo é motivado apenas por elas serem mais convenientes para escrita e visualização, se comparadas às respectivas representações vetoriais.

cuja forma vetorial correspondente é dada por $\mathbf{h}_2 = \text{vec}(\mathbf{H}_2)$, em que o operador vec denota a formação de um vetor coluna composto pelas colunas de $\mathbf{H}_2$ “empilhadas” (ver Subseção A.1.5 do Apêndice A). Empregando-se um fator de interpolação $L = 2$, o núcleo esparso correspondente é

$$\mathbf{H}_2^{(s)} = \begin{bmatrix} h_2(0,0) & 0 & h_2(0,2) \\ 0 & 0 & 0 \\ h_2(2,0) & 0 & h_2(2,2) \end{bmatrix} \quad (4.57)$$

cuja representação vetorial é $\mathbf{h}_2^{(s)} = \text{vec}(\mathbf{H}_2^{(s)})$. O próximo passo é determinarmos a matriz de interpolação de segunda ordem, o que é feito com o uso de (4.55) e (4.51), resultando em

$$\mathbf{G}_2 = \frac{1}{2} \begin{bmatrix} \mathbf{G} & \mathbf{0} & \mathbf{0} \\ 2\mathbf{G} & \mathbf{G} & \mathbf{0} \\ \mathbf{G} & 2\mathbf{G} & \mathbf{G} \\ \mathbf{0} & \mathbf{G} & 2\mathbf{G} \\ \mathbf{0} & \mathbf{0} & \mathbf{G} \end{bmatrix} \quad (4.58)$$

onde $\mathbf{0} \in \mathbb{R}^{5 \times 3}$. Assim, a representação vetorial do núcleo de segunda ordem resultante, dada por $\mathbf{h}_2^{(i)} = \mathbf{G}_2 \mathbf{h}_2^{(s)}$, corresponde à matriz [28]

$$\mathbf{H}_2^{(i)} = \frac{1}{4} \begin{bmatrix} h_2(0,0) & 2h_2(0,0) & h_2(0,0) + h_2(0,2) & 2h_2(0,2) & h_2(0,2) \\ 2h_2(0,0) & 4h_2(0,0) & 2[h_2(0,0) + h_2(0,2)] & 4h_2(0,2) & 2h_2(0,2) \\ h_2(0,0) + h_2(2,0) & 2[h_2(0,0) + h_2(2,0)] & h_2(0,0) + h_2(0,2) + h_2(2,0) + h_2(2,2) & 2[h_2(0,2) + h_2(2,2)] & h_2(0,2) + h_2(2,2) \\ 2h_2(2,0) & 4h_2(2,0) & 2[h_2(2,0) + h_2(2,2)] & 4h_2(2,2) & 2h_2(2,2) \\ h_2(2,0) & 2h_2(2,0) & h_2(2,0) + h_2(2,2) & 2h_2(2,2) & h_2(2,2) \end{bmatrix}. \quad (4.59)$$

Observe que, assim como no caso linear, há um surgimento de coeficientes adicionais, o que corresponde a um “efeito de bordas” causado pela interpolação. Em $\mathbf{H}_2^{(i)}$, tais coeficientes aparecem na primeira e última linhas e na primeira e última colunas. Novamente, deve-se notar que,

admitindo-se que $M \ll N$, como normalmente é o caso na prática, este fenômeno é pouco significativo. Quanto aos coeficientes reconstruídos, que correspondem aos elementos $[\mathbf{H}_2^{(i)}]_{2,3}$, $[\mathbf{H}_2^{(i)}]_{3,2}$, $[\mathbf{H}_2^{(i)}]_{3,3}$, $[\mathbf{H}_2^{(i)}]_{3,4}$ e $[\mathbf{H}_2^{(i)}]_{4,3}$, pode-se observar que eles resultam da interpolação linear dos coeficientes originais adjacentes na matriz esparsa.

□

Embora no exemplo dado tenhamos nos restringido ao núcleo de segunda ordem por simplicidade, as observações feitas ao final dele se aplicam, com as devidas generalizações, aos demais núcleos de alta ordem. Logo, analogamente ao caso linear, para que a aproximação (4.41) seja suficientemente precisa, é necessário que os coeficientes adjacentes de cada núcleo do FV em questão apresentem alto grau de correlação. Em outras palavras, os núcleos do FV devem possuir um comportamento caracterizado por variações suaves.

Assim como ocorre no caso linear, o uso de um filtro interpolador causal e simétrico, como aquele do Exemplo 4.2, introduz um atraso no processamento realizado pela estrutura resultante. Todavia, em ambos os casos, pode-se empregar uma técnica que é discutida em [80] para a remoção do efeito de bordas e do atraso associado. Uma outra possibilidade interessante explorada em [80] é o uso de um esquema adaptativo para a otimização do filtro interpolador, o que possibilita aprimorar o desempenho da estrutura no que se refere à precisão da aproximação por ela fornecida.

Por fim, deve-se notar que, para fins de implementação, evidentemente utilizam-se vetores $\mathbf{h}_p^{(s)}$ que constituem representações triangulares dos núcleos e vetores de entrada interpolados não-redundantes, de forma análoga à implementação descrita na Seção 4.1. Todavia, permanecem válidas as conclusões obtidas nesta subseção, devido à equivalência entre as representações triangular e simétrica de quaisquer núcleos, sejam esparsos ou não.

4.3.2.1 Complexidades estrutural e computacional

No que se refere ao aumento de compacidade trazido pelo esquema apresentado, como o critério de descarte dos coeficientes leva em conta todos os p índices

de um núcleo de grau p , então, o número total de coeficientes envolvidos na representação esparsa com interpolação é idêntico àquele de um FV cujos núcleos possuem extensões de memória dadas por [28]

$$N_p^{(s)} = \left\lfloor \frac{N_p - 1}{L} \right\rfloor + 1 \quad (4.60)$$

em que L é o fator de interpolação utilizado e N_p é a extensão de memória do núcleo de ordem p original. Tal equivalência é explicada pelo fato de que, devido ao descarte dos coeficientes, o núcleo esparsa de ordem p possui o mesmo número de graus de liberdade que um núcleo de mesma ordem, porém cuja memória possui extensão $N_p^{(s)}$.

Analogamente, a complexidade computacional associada ao uso deste núcleo também é da mesma ordem que aquela relativa a um núcleo de ordem p com extensão de memória dada pela fórmula (4.60). Quanto à complexidade total do filtro de Volterra resultante, é evidente que há ainda a necessidade de algumas operações adicionais para o cômputo da saída do filtro interpolador. No entanto, tal filtro costuma possuir uma resposta impulsiva de curta duração, e portanto é pouco significativo para a complexidade total do esquema.

Em decorrência do que foi discutido acima, conclui-se que o ganho em eficiência obtido com esta abordagem, assim como o próprio número de coeficientes que compõem um núcleo de Volterra, cresce muito rapidamente com sua ordem. Logo, trata-se de uma alternativa particularmente atrativa quando se necessita de núcleos de alta ordem para a representação precisa de um determinado sistema.

Com o uso da equação (4.1), podemos comparar quantitativamente a complexidade estrutural de um FV convencional em relação a um FV esparsa com interpolação, como é feito na Tabela 4.2 para alguns valores da ordem dos modelos, P , do fator de interpolação, L , e da extensão de memória, N , admitindo-se que $N_1 = N_2 = \dots = N_P = N$ em todos os casos. Deve-se notar que, de acordo com o que foi mencionado ao final da subseção anterior, considera-se que os núcleos são representados sem redundâncias, como ocorre na prática. A partir dos dados mostrados e da argumentação apresentada, constata-se que, com o crescimento da ordem ou da memória, a complexidade estrutural é significativamente

reduzida pelo uso do esquema esparsos com interpolação, o que implica em uma economia computacional considerável.

	$N = 30$			$N = 50$			$N = 70$		
P	$L = 4$	$L = 2$	FV	$L = 4$	$L = 2$	FV	$L = 4$	$L = 2$	FV
1	8	15	30	13	25	50	18	35	70
2	44	135	495	104	350	1.325	189	665	2.555
3	164	815	5.455	559	3.275	23.425	1.329	8.435	62.195
4	494	3.875	46.375	2.379	23.750	316.250	7.314	82.250	1.150.625
5	1.286	15.503	324.631	8.567	142.505	3.478.760	33.648	658.007	17.259.389

Tabela 4.2: Comparação entre os números de coeficientes de filtros de Volterra esparsos com interpolação e de filtros de Volterra convencionais.

4.4 Estruturas baseadas no modelo de Wiener

Na Subseção 3.4.5, vimos que a representação matricial do núcleo de segunda ordem de um modelo de Volterra equivalente a um modelo de Wiener com M filtros lineares é dada por

$$\mathbf{H}_2 = \sum_{i=1}^M \sum_{j=1}^M \alpha_{i,j} \mathbf{g}_i \mathbf{g}_j^T \quad (4.61)$$

onde $\mathbf{g}_i$ é a representação vetorial do filtro $g_i(n)$, $i \in \{1, \dots, M\}$. Assim, considerando-se que qualquer núcleo admite uma representação simétrica, então, qualquer núcleo de segunda ordem com posto M pode ser escrito² na forma particular de (4.61) dada por

$$\mathbf{H}_2 = \sum_{i=1}^M \lambda_i \mathbf{g}_i \mathbf{g}_i^T \quad (4.62)$$

em que $\mathbf{g}_i$ são autovetores ortogonais de $\mathbf{H}_2$ e os escalares λ_i são seus respectivos autovalores. Consequentemente, com uma escolha adequada dos filtros lineares e da não-linearidade sem memória, qualquer filtro de Volterra homogêneo de grau dois com memória finita de N amostras e cujo núcleo de segunda ordem possui posto M pode ser implementado exatamente através de uma estrutura de Wiener

²Na realidade, qualquer matriz simétrica pode ser escrita na forma (4.62), e não apenas núcleos de Volterra.

contendo um banco de M filtros lineares, cada um possuindo comprimento N , os quais compõem uma expansão do núcleo de segunda ordem na forma (4.62).

Esta interpretação estrutural de (4.62) em termos do modelo de Wiener associado é mostrada na Figura 4.6. Note que a não-linearidade sem memória não envolve produtos cruzados entre as saídas dos filtros, uma vez que não há produtos internos entre vetores $\mathbf{g}_i$ e $\mathbf{g}_j$ com $i \neq j$ em (4.62). Tal desacoplamento se deve, basicamente, à existência de uma matriz unitária (cujas colunas são os vetores $\mathbf{g}_i$) que diagonaliza $\mathbf{H}_2$.

Figura 4.6: Modelo de Wiener equivalente ao modelo de Volterra homogêneo de grau dois cujo núcleo é dado por (4.62).

O mesmo argumento apresentado nos parágrafos anteriores é válido, com as devidas generalizações, para sistemas homogêneos de maior grau, cujos núcleos podem ser interpretados como tensores de alta ordem [29]. Computacionalmente, a decomposição dos núcleos através do uso de uma estrutura de Wiener apresenta vantagens interessantes em relação a um FV convencional, a saber:

- 1) Com frequência, os núcleos de sistemas físicos podem ser aproximados precisamente por tensores de posto reduzido, o que permite empregar uma estrutura contendo poucos filtros lineares, reduzindo-se, portanto, consideravelmente o número de parâmetros em relação ao FV. Consequentemente, tem-se um ganho em eficiência, o qual será tanto maior quanto menor for o posto que possibilite atingir a precisão desejada.
- 2) O cálculo da saída dos filtros lineares pode ser realizado eficientemente no domínio da frequência, sem que seja necessário considerar uma formulação especial do modelo como a representação diagonal para FVs descrita na Seção 4.2.

- 3) A implementação de um modelo de Wiener é mais simples, em geral, do que a de um FV convencional, seja esta na forma direta ou na forma diagonal. Isso decorre do fato de que o modelo de Wiener envolve apenas filtragens lineares da sequência de entrada e o cálculo de uma função não-linear sem memória, enquanto a implementação de um FV convencional requer ainda a formação de vetores (não-redundantes) contendo produtos de amostras da entrada de várias ordens.

Devido às propriedades do modelo de Wiener mencionadas, diversos trabalhos recentes tem buscado explorá-lo com o intuito de atenuar a complexidade computacional tipicamente associada a filtros de Volterra convencionais. Em linhas gerais, a estratégia empregada consiste em formular uma estrutura de Volterra alternativa baseada no modelo de Wiener que forneça um compromisso entre complexidade computacional e precisão. A diferença entre as estruturas propostas reside, basicamente, na forma como sua não-linearidade sem memória e seus filtros lineares são obtidos. Nesta seção, serão discutidas de forma breve duas abordagens baseadas nesta estratégia propostas recentemente.

4.4.1 O modelo PARAFAC-Volterra

Nesta subseção descreve-se, em linhas gerais, a ideia proposta em [29]. Para tanto, empregaremos a notação e as definições de operadores relacionadas a tensores que são introduzidas no Apêndice A.

Ao longo deste capítulo e do anterior, fez-se uso da representação matricial de núcleos de segunda ordem de tempo discreto algumas vezes, o que facilita a análise de certos aspectos e a derivação de alguns resultados. Formalmente, a ideia subjacente a este recurso consiste em estabelecer um mapeamento bijetivo entre um núcleo

$$h_2 : I_2 \times I_2 \rightarrow \mathbb{R} \quad (4.63)$$

onde $I_2 = \{0, 1, \dots, N_2 - 1\}$, sendo N_2 a extensão da memória deste núcleo, e uma

matriz $\mathbf{H}_2 \in \mathbb{R}^{N_2 \times N_2}$ cujos elementos são dados por

$$[\mathbf{H}_2]_{i_1, i_2} = h_2(i_1 - 1, i_2 - 1). \quad (4.64)$$

Generalizando-se esta ideia, podemos representar o núcleo de ordem p

$$h_p : I_p \times I_p \times \dots \times I_p \rightarrow \mathbb{R} \quad (4.65)$$

pelo tensor $\mathcal{H}_p \in \mathbb{R}^{N_p^{(p)}}$ cujos elementos são dados por

$$[\mathcal{H}_p]_{i_1, \dots, i_p} = h_p(i_1 - 1, \dots, i_p - 1) \quad (4.66)$$

sendo $I_p = \{0, 1, \dots, N_p - 1\}$, onde N_p é a extensão de memória do núcleo. Com isso, uma série de conceitos e resultados existentes na literatura relacionados à álgebra de tensores podem ser aplicados aos núcleos de filtros de Volterra.

Em particular, como discutido no Capítulo 3, tais núcleos sempre podem ser representados em uma forma simétrica, a qual é única. Correspondentemente, para uma representação tensorial destas funções, podemos nos restringir a tensores simétricos, i.e., tais que

$$\forall i_1, \dots, i_p \in \{1, \dots, N_p\}, \quad [\mathcal{H}_p]_{i_1, \dots, i_p} = [\mathcal{H}_p]_{\pi_1(i), \dots, \pi_p(i)} \quad (4.67)$$

em que $\pi = (\pi_1(i), \dots, \pi_p(i))$ é uma permutação qualquer do multi-índice $i = (i_1, \dots, i_p)$. Por consequência, temos que os núcleos de um filtro de Volterra sempre podem ser decompostos na forma [29, 81]

$$\mathcal{H}_p = \sum_{r=1}^{R_p} \mathbf{a}_{p,r} \circ \mathbf{a}_{p,r} \circ \dots \circ \mathbf{a}_{p,r} \quad (4.68)$$

onde R_p é o posto da decomposição, o símbolo $\circ$ denota o operador de produto externo (ver Subseção A.1.3 do Apêndice A), cada termo do somatório envolve $p-1$ produtos externos (ou seja, o vetor $\mathbf{a}_{p,r}$ é repetido p vezes em sua expressão)

e $\mathbf{a}_{p,r} \in \mathbb{C}^{N_p}$. Assim, os elementos de $\mathcal{H}_p$ são dados por

$$[\mathcal{H}_p]_{i_1, \dots, i_p} = \sum_{r=1}^{R_p} \prod_{j=1}^p [\mathbf{a}_{p,r}]_{i_j}. \quad (4.69)$$

Observe que, para $p = 2$, tal decomposição pode ser escrita como

$$\mathbf{H}_2 = \sum_{r=1}^{R_2} \mathbf{a}_{2,r} \mathbf{a}_{2,r}^T \quad (4.70)$$

sendo possível obtê-la diretamente a partir da decomposição por autovalores de qualquer matriz real simétrica de posto R_2 na forma

$$\mathbf{H}_2 = \sum_{r=1}^{R_2} \lambda_r \mathbf{g}_r \mathbf{g}_r^T \quad (4.71)$$

definindo-se, para tanto,

$$\mathbf{a}_{2,r} = \sqrt{\lambda_r} \mathbf{g}_r, \quad r \in \{1, \dots, R_2\}. \quad (4.72)$$

É importante que se note que, embora estejamos considerando o caso em que os núcleos de Volterra são definidos sobre o campo dos reais, os vetores $\mathbf{a}_{p,r}$ que compõem a expressão (4.68) devem ser definidos sobre o campo dos complexos para que se garanta a existência da decomposição. No caso de segunda ordem, por exemplo, esta garantia decorre do fato de que, tendo sido feita a generalização de tais vetores para o campo dos complexos, então não há problema em relação ao valor de $\sqrt{\lambda_r}$ envolvido em (4.72), o qual é sempre definido para qualquer autovalor λ_r . Caso os vetores $\mathbf{a}_{2,r}$ fossem definidos sobre o campo dos reais, então a incorporação do fator $\sqrt{\lambda_r}$ nos vetores $\mathbf{a}_{2,r}$ para se obter a forma (4.70) seria possível apenas para matrizes positivas semidefinidas ou positivas definidas, pois somente assim os autovalores λ_i seriam garantidamente não-negativos. Como observado em [81], um argumento semelhante é válido para tensores reais de alta ordem, os quais sempre podem ser decompostos na forma (4.68) desde que os vetores envolvidos sejam definidos sobre o campo dos números complexos, visto que qualquer número real admite uma p -ésima raiz complexa que pode ser

incorporada nos vetores $\mathbf{a}_{p,r}$. Sendo assim, a generalização feita acima tem a finalidade de assegurar a existência da decomposição (4.68).

A decomposição na forma (4.68) é chamada na literatura de PARAFAC (do inglês *parallel factor decomposition*) [29, 82], sendo os vetores $\mathbf{a}_{p,r}$ chamados de fatores e a matriz $N_p \times R_p$

$$\mathbf{A}_p = [\mathbf{a}_{p,1} \ \mathbf{a}_{p,2} \ \dots \ \mathbf{a}_{p,R_p}] \quad (4.73)$$

denominada matriz de fatores.

Com base na decomposição PARAFAC, a saída de ordem p de um filtro de Volterra (com núcleo simétrico) pode ser escrita como [29]

$$y_p(n) = \sum_{k_1=0}^{N_p-1} \dots \sum_{k_p=0}^{N_p-1} [\mathcal{H}_p]_{k_1+1, \dots, k_p+1} \prod_{i=1}^p u(n - k_i) \quad (4.74)$$

$$= \sum_{k_1=0}^{N_p-1} \dots \sum_{k_p=0}^{N_p-1} \left\{ \sum_{r=1}^{R_p} \prod_{j=1}^p [\mathbf{a}_{p,r}]_{k_j+1} \right\} \prod_{i=1}^p u(n - k_i) \quad (4.75)$$

$$= \sum_{k_1=0}^{N_p-1} \dots \sum_{k_p=0}^{N_p-1} \sum_{r=1}^{R_p} \prod_{i=1}^p [\mathbf{a}_{p,r}]_{k_i+1} u(n - k_i) \quad (4.76)$$

$$= \sum_{r=1}^{R_p} \sum_{k_1=0}^{N_p-1} \dots \sum_{k_p=0}^{N_p-1} \prod_{i=1}^p [\mathbf{a}_{p,r}]_{k_i+1} u(n - k_i). \quad (4.77)$$

Agora, se fatorarmos o resultado dos p somatórios mais internos de (4.77) em expressões que dependam cada uma de apenas um índice k_i , obtemos a forma equivalente

$$y_p(n) = \sum_{r=1}^{R_p} \prod_{i=1}^p \sum_{k_i=0}^{N_p-1} [\mathbf{a}_{p,r}]_{k_i+1} u(n - k_i). \quad (4.78)$$

Assim, definindo-se o vetor de entradas

$$\mathbf{u}^{(p)}(n) = [u(n) \ u(n-1) \ \dots \ u(n - N_p + 1)]^T \quad (4.79)$$

a expressão (4.78) pode ser reescrita na forma

$$y_p(n) = \sum_{r=1}^{R_p} [\mathbf{a}_{p,r}^T \mathbf{u}^{(p)}(n)]^p. \quad (4.80)$$

O modelo resultante, denominado PARAFAC-Volterra [29], é representado pelo diagrama de blocos da Figura 4.7. Nele, para cada fator paralelo $\mathbf{a}_{p,r}$ da relação (4.80) há um filtro FIR correspondente, $A_{p,r}(z)$, cuja resposta impulsiva na forma vetorial é dada pelo próprio fator associado. Note que se trata de uma estrutura cuja implementação é simples e pode ser paralelizada completamente, uma vez que inclusive a não-linearidade sem memória pode ser decomposta em um somatório de termos tal que cada um dependa apenas da saída de um dos filtros.

Figura 4.7: Estrutura de Volterra alternativa PARAFAC-Volterra [29].

4.4.1.1 Obtenção dos fatores a partir de um FV de referência

Dado um FV de referência de ordem P , a chave para se obter uma aproximação do tipo (4.80) é a decomposição das representações tensoriais (simétricas) de seus núcleos na forma (4.68). Para um núcleo de ordem $p = 2$, como discutido, isto

pode ser obtido diretamente a partir da decomposição por autovalores da matriz correspondente.

Já no caso de ordem $p > 2$, apesar de ter sido mostrado na literatura que tensores simétricos sempre podem ser escritos nesta forma, a determinação de vetores que componham uma decomposição do tipo (4.68) não é um problema simples. De fato, os conceitos algébricos envolvidos nesse caso são consideravelmente mais complexos, existindo uma série de problemas ainda em aberto, os quais são bastante estudados atualmente [81, 83, 84, 85].

Uma das dificuldades envolvidas neste cômputo é a determinação de um número de fatores adequados, ou, equivalentemente, do posto simétrico do tensor em questão. Para um tensor simétrico qualquer $\mathcal{T} \in \mathbb{C}^{d^{(p)}}$, define-se o posto simétrico como o menor número $R_s \in \mathbb{N}$ tal que $\mathcal{T}$ possa ser escrito como [81]

$$\mathcal{T} = \sum_{r=1}^{R_s} \mathbf{v}_r \circ \mathbf{v}_r \circ \dots \circ \mathbf{v}_r \quad (4.81)$$

onde $\mathbf{v}_r \in \mathbb{C}^d$ e cada termo do somatório envolve $p - 1$ produtos externos. O cômputo de R_s e até mesmo a determinação de seu máximo valor possível³ em função da ordem p e da dimensão d não são problemas resolvidos na literatura [81].

Na prática, muitas vezes se escolhe um valor reduzido para o número de fatores considerados, buscando-se, na realidade, uma aproximação na forma (4.81) [84]. Uma razão para isso é que, sob algumas condições, a escolha de um valor subdimensionado para o posto da decomposição permite obter uma decomposição única [84].

Assim, partindo do pressuposto de que um valor para R_s é escolhido *a priori*, alguns algoritmos foram propostos na literatura para o cômputo da decomposição (4.81). Em [83], por exemplo, propõe-se um algoritmo que explora a dualidade entre tensores simétricos e polinômios multivariável homogêneos.

Generalizando-se a decomposição para tensores quaisquer (i.e., possivelmente

³Para tensores simétricos, sabe-se que $R_s \leq \binom{d+p-1}{p}$ [81], fato que, no entanto, possui pouca utilidade prática, visto que tal limite é com frequência superestimado.

assimétricos), ela assume a forma

$$\mathcal{T} = \sum_{r=1}^R \mathbf{v}_{1,r} \circ \mathbf{v}_{2,r} \circ \dots \circ \mathbf{v}_{p,r}. \quad (4.82)$$

Para determiná-la, uma família de algoritmos existente na literatura cuja estratégia básica é bastante simples é chamada de algoritmos do tipo ALS (*alternating least squares*) [85]. Basicamente, tal estratégia consiste em computar os vetores da decomposição iterativamente, sendo que, em cada iteração, fixam-se os vetores $\mathbf{v}_{i,r}$ tais que $i \in \{1, \dots, j-1, j+1, \dots, p\}$ e calculam-se os vetores $\mathbf{v}_{j,r}$ que minimizam o critério

$$J = \left\| \mathcal{T} - \sum_{r=1}^R \mathbf{v}_{1,r} \circ \mathbf{v}_{2,r} \circ \dots \circ \mathbf{v}_{p,r} \right\|_F \quad (4.83)$$

em que $\|\cdot\|_F$ denota a norma de Frobenius (i.e., a raiz do somatório dos quadrados dos componentes do tensor). Dessa forma, pode-se mostrar que, a cada iteração, os vetores que minimizam J são dados pelo estimador de mínimos quadrados⁴ [85].

Embora apresentem grande simplicidade, tais algoritmos possuem como desvantagem o fato de que não há garantia de convergência e tampouco de que a solução computada é ótima [84]. Sendo assim, mesmo que eles forneçam um resultado insatisfatório em uma dada situação, isso não quer dizer que uma boa aproximação do tipo (4.81) não existe para o tensor em questão. Apesar dos problemas mencionados, do ponto de vista prático algoritmos ALS constituem uma alternativa pragmática que frequentemente conduz a aproximações razoáveis.

Para finalizar, deve-se ressaltar que os algoritmos mencionados acima são aplicáveis somente quando os núcleos do FV que se deseja decompor são previamente conhecidos. Em aplicações práticas, este nem sempre é o caso, sendo muitas vezes necessário estimar os fatores paralelos a partir de um conjunto de dados. Adicionalmente, pode ser necessário que se faça uma atualização contínua (i.e., adaptação) da estrutura PARAFAC-Volterra. Para tais finalidades, algoritmos de

⁴Observe que os vetores da decomposição são únicos somente sob a suposição de que $p-1$ deles possuem norma unitária, devido à ambiguidade de escala introduzida pelos produtos entre os componentes dos vetores.

estimação dos parâmetros desta estrutura são discutidos, por exemplo, em [29].

4.4.1.2 Complexidades estrutural e computacional

Consideremos agora o uso do esquema descrito em comparação com o de um FV de referência de ordem P , cujos núcleos possuem extensões de memórias N_p , $p \in \{1, \dots, P\}$. De acordo com a discussão realizada, é evidente que a redução de complexidade estrutural obtida com a adoção da estrutura PARAFAC-Volterra de ordem P depende dos postos $R_2, \dots, R_P$ suficientes para se obter uma aproximação satisfatória dos núcleos do FV convencional. O número total de coeficientes necessários para se caracterizar um modelo PARAFAC-Volterra $\mathcal{P}$ completamente é

$$\Psi_{\text{PARAFAC}}(\mathcal{P}) = N_1 + \sum_{p=2}^P R_p N_p. \quad (4.84)$$

Para se ter uma ideia quantitativa do que isso representa em termos de ganho em compacidade, na Tabela 4.3 são mostrados alguns valores de (4.84) e também de (4.1) para modelos de ordem P . Visando simplificar a comparação, admitiu-se que os postos para as diferentes ordens são $R_2 = R_3 = \dots = R_5 = R$ e as extensões das memórias dos núcleos são $N_1 = N_2 = N_3 = \dots = N_5 = N$. Assim, variando-se N , P e R , os valores resultantes das fórmulas mencionadas são como dados na tabela. Para cada N , as duas primeiras colunas correspondem aos números de coeficientes obtidos com respectivos valores de R a que se referem, enquanto a terceira corresponde ao número de coeficientes de um FV convencional.

P	$N = 30$			$N = 50$			$N = 70$		
	$R = 10$	$R = 20$	FV	$R = 10$	$R = 20$	FV	$R = 10$	$R = 20$	FV
1	30	30	30	50	50	50	70	70	70
2	360	660	495	600	1.100	1.325	840	1.540	2.555
3	990	1.890	5.455	1.650	3.150	23.425	2.310	4.410	62.195
4	1.920	3.720	46.375	3.200	6.200	316.250	4.480	8.680	1.150.625
5	3.150	6.150	324.631	5.250	10.250	3.478.760	7.350	14.350	17.259.389

Tabela 4.3: Comparação entre os números de coeficientes de modelos PARAFAC-Volterra e de filtros de Volterra convencionais.

A partir dos dados mostrados na Tabela 4.3, observa-se que o número de parâmetros que compõem o FV cresce muito mais rapidamente do que o número de parâmetros de um modelo PARAFAC-Volterra para os valores de R utilizados.

No entanto, vê-se que há vantagem significativa somente quando R é consideravelmente menor que N e/ou o modelo é de alta ordem. Em caso contrário, o número de parâmetros do modelo PARAFAC-Volterra pode ser comparável ou até mesmo superior ao de um FV convencional, como ocorre na tabela para $P = 2$, $N = 30$ e $R = 20$. Isto pode ser explicado pelo fato de que, nesse caso, a representação triangular do núcleo de segunda ordem proporciona uma redução de coeficientes maior do que a decomposição PARAFAC do núcleo.

Deve-se ressaltar que, obviamente, os valores da Tabela 4.3 são fornecidos apenas como exemplos, uma vez que um valor adequado para R depende do sistema sendo modelado. No Capítulo 6, será feita uma análise mais detalhada sobre o ganho em eficiência obtido pelo uso de modelos PARAFAC-Volterra em relação a um FV de referência de um sistema físico específico, considerando-se valores de posto que permitam atingir diferentes níveis de precisão.

No que se refere à complexidade computacional de sua implementação, o modelo PARAFAC-Volterra requer, para se computar o componente de ordem p da saída, R_p filtragens lineares, $(p - 1)R_p$ multiplicações relativas às potências e $R_p - 1$ adições para agregar as saídas dos filtros. Assim, se os filtros lineares forem implementados na forma direta, o cálculo de $y_p(n)$ requer um número de operações da ordem de $O(R_p(N_p + p))$ por amostra. Já se a implementação deles é no domínio da frequência e se computam blocos de tamanho M , então são necessárias $O(R_p(M + N_p) \log(M + N_p) + pMR_p)$ operações por bloco de $y_p(n)$.

4.4.2 O modelo de modos dinâmicos principais

Analogamente ao que foi discutido na Subseção 3.4.5 para modelos de Volterra de tempo contínuo, uma justificativa para a capacidade ampla de representação dos modelos de Wiener é que, admitindo-se que os filtros lineares empregados, $g_i(n)$, $i = 1, \dots, M$, fornecem um número suficiente de graus de liberdade, então os núcleos de um FV equivalente, dados por

$$h_p(k_1, \dots, k_p) = \sum_{j_1=1}^M \dots \sum_{j_p=1}^M \beta_p(j_1, \dots, j_p) g_{j_1}(k_1) \dots g_{j_p}(k_p) \quad (4.85)$$

em que os coeficientes β_p são derivados da expansão em série de Taylor da não-linearidade sem memória, podem assumir (ou, ao menos, aproximar precisamente) quaisquer formas arbitrárias. É nítido então que o aspecto chave para a obtenção de um modelo de Wiener satisfatório é a definição do banco de filtros lineares que o compõe, os quais também são chamados de modos [9].

No contexto do modelamento de sistemas de tempo contínuo, Wiener sugeriu o uso de funções de Laguerre para o banco de filtros como forma de se produzir expansões funcionais que aproximassem os núcleos de uma ampla classe de sistemas. Tal escolha foi feita devido às funções de Laguerre serem ortogonais entre si e também por apresentarem uma forma exponencial consistente com a resposta de sistemas físicos [9].

Em se tratando de sistemas de tempo discreto, [9] descreve um método análogo para a obtenção de um banco de filtros dados por funções de Laguerre discretas⁵ (FLD), as quais também são, por construção, ortogonais. Em linhas gerais, utilizam-se as informações fornecidas por um conjunto de medições de entradas e saídas para se estimar um valor apropriado para o parâmetro α que caracteriza a taxa de decaimento exponencial das FLDs. Assim, uma vez obtidos os filtros lineares, podem-se estimar os coeficientes da expansão polinomial da não-linearidade sem memória com o uso do estimador de mínimos quadrados. Este método empírico é chamado em [9] de técnica de expansão de Laguerre (LET, do inglês *Laguerre expansion technique*).

Embora o método LET ofereça uma forma prática de se aproximar a relação entrada/saída de um sistema de tempo discreto através de uma estrutura de Wiener, o caráter geral das funções de Laguerre discretas geralmente impede que a expansão funcional dos núcleos seja eficiente. Em outras palavras, por não utilizar filtros especializados para o sistema sendo modelado, tal método não fornece, em geral, um modelo de Wiener tão parcimonioso quanto possível [9].

A ideia da abordagem de modos dinâmicos principais (MDP) é buscar um banco com o menor número possível de filtros lineares tais que os núcleos resultantes permitam aproximar com precisão a relação entrada/saída de um sistema

⁵Deve-se notar que as funções de Laguerre discretas *não* são iguais às funções de Laguerre discretizadas, as quais não são, em geral, ortogonais [9].

de interesse [9]. Em outras palavras, deseja-se encontrar um conjunto mínimo de modos que capturem as características dinâmicas essenciais que compõem os núcleos da representação de Volterra de tempo discreto associada a tal sistema.

Usualmente, a não-linearidade do modelo MDP é definida como um polinômio de grau finito P , o que traz vantagens em termos de simplicidade do modelo. Adicionalmente, como neste caso a saída do modelo é escrita em uma forma linear nos parâmetros que caracterizam a não-linearidade, pode-se estimá-los facilmente usando-se, por exemplo, o método de mínimos quadrados. Porém, deve-se notar que, em contrapartida, tal escolha limita o grau máximo de não-linearidade do modelo.

No que segue, consideraremos, assim como é feito em [9], um modelo MDP com não-linearidade sem memória polinomial de grau finito P . Neste caso, a relação entrada/saída do modelo é dada por

$$y(n) = \sum_{p=1}^P \sum_{j_1=1}^M \dots \sum_{j_p=j_{p-1}}^M \beta_p(j_1, \dots, j_p) v_{j_1}(n) \dots v_{j_p}(n) \quad (4.86)$$

$$v_i(n) = \sum_{k=0}^{N-1} g_i(k) u(n-k), \quad i = 1, \dots, M \quad (4.87)$$

onde $g_i(k)$ é o i -ésimo modo dinâmico principal e os coeficientes $\beta_p(j_1, \dots, j_p)$ definem a não-linearidade polinomial multivariável

$$F(x_1, \dots, x_M) = \sum_{p=1}^P \sum_{j_1=1}^M \dots \sum_{j_p=j_{p-1}}^M \beta_p(j_1, \dots, j_p) x_{j_1} \dots x_{j_p}. \quad (4.88)$$

É importante que sejam feitas duas observações a respeito da forma (4.86)–(4.87). Primeiro, note que, em decorrência do fato de que o mesmo conjunto de filtros lineares “gera” todos os núcleos do filtro de Volterra equivalente ao MDP formulado, então tais núcleos possuem a mesma extensão de memória N . Segundo, observe que os índices $j_1, \dots, j_p$ de (4.86) variam sobre um domínio triangular, i.e., $j_1 \leq j_2 \leq \dots \leq j_p$. Fazendo um paralelo com o que foi discutido na Subseção 3.1.1.1 para núcleos de Volterra, trata-se de uma forma mais apropriada do que a do somatório em (4.85) do ponto de vista de unicidade da definição da

não-linearidade sem memória. Adicionalmente, para fins de implementação, o uso de um domínio triangular também é vantajoso, uma vez que permite reduzir o número de operações envolvidas no somatório. Evidentemente, dada uma função β não-triangular que define os coeficientes da não-linearidade, como aquela em (4.85), os valores de uma função triangular que produz o mesmo resultado podem ser calculados através de uma expressão análoga a (3.17).

O modelo (4.86)–(4.87) é ilustrado pelo diagrama da Figura 4.8. Observe que se trata de um modelo estruturalmente idêntico àquele da Figura 3.7. De fato, o modelo MDP é um caso particular do modelo de Wiener em que os filtros possuem uma interpretação especial como os modos básicos que compõem a dinâmica do sistema modelado.

Figura 4.8: Modelo de modos dinâmicos principais.

Conceitualmente, quando o banco de filtros lineares da estrutura é formado pelos modos dinâmicos principais do sistema, o modelo apresenta duas propriedades importantes. Em primeiro lugar, obtém-se uma representação eficiente em termos do número de parâmetros que compõem o modelo. Além disso, tem-se também um ganho em interpretabilidade do modelo, uma vez que, em princípio, os modos dinâmicos principais fornecem informações importantes a respeito das características dinâmicas do sistema de interesse [9]. Na realidade, como discutido em [9], esta última vantagem mencionada é também uma das motivações do modelo MDP.

Comparando-se a discussão realizada em relação ao MDP com aquela em que se descreveu o modelo PARAFAC-Volterra, nota-se que, além de ambas as EVAs

serem casos particulares da estrutura de Wiener, elas possuem ainda outra característica em comum: à semelhança dos modos dinâmicos principais, o modelo PARAFAC-Volterra é formado por um banco de filtros lineares escolhidos com o intuito de se obter uma boa aproximação dos núcleos com um número reduzido de filtros. No entanto, enquanto a justificativa dos fatores paralelos é algébrica, os modos dinâmicos principais são justificados em [9], de certa forma, mais intuitivamente.

Uma outra diferença ainda mais evidente, é que, por definição, a não-linearidade sem memória dos modelos PARAFAC-Volterra é tal que não há acoplamento entre saídas dos diferentes filtros que compõem o banco, o que não é necessariamente verdadeiro para modelos MDP. Em outras palavras, o polinômio multivariável de um modelo PARAFAC-Volterra não possui produtos cruzados entre variáveis, correspondendo a uma versão de (4.86)–(4.87) em que

$$\beta_p(j_1, \dots, j_p) \neq 0 \implies j_1 = j_2 = \dots = j_p \quad (4.89)$$

sendo que os filtros $g_i(n)$ são dados, nesse caso, pelos fatores paralelos das representações tensoriais dos núcleos.

4.4.2.1 Obtenção dos modos dinâmicos principais

De acordo com a discussão realizada, fica claro que o requisito central para a construção de um modelo MDP é a determinação de um conjunto de filtros lineares que apresentem as propriedades mencionadas acima. Em [9], discutem-se algumas formas de fazê-lo em situações nas quais o modelo deve ser obtido apenas com base em entradas e saídas medidas. Basicamente, propõe-se que os modos dinâmicos principais sejam derivados de um outro modelo, o qual é identificado através de um método geral como o LET ou a abordagem connexionista.

No contexto deste trabalho, o interesse recai sobre a situação em que um conjunto de núcleos de referência é conhecido, como será discutido nos Capítulos 5 e 6. Quando este é o caso, os modos dinâmicos principais desejados podem ser obtidos a partir das próprias informações fornecidas pelos núcleos. Assim,

supõe-se que são conhecidos os núcleos de ordens 1 a P , os quais possuem, respectivamente, extensões de memória dadas por N_p , $p \in \{1, \dots, P\}$. No que segue, nos basearemos nos conceitos de álgebra de tensores apresentados no Apêndice A para mostrar uma forma de se obter modos dinâmicos para o modelo.

Em relação ao núcleo de segunda ordem, uma forma de obter os modos dinâmicos principais a partir de sua decomposição por valores singulares consiste em tomar os vetores singulares relacionados aos valores singulares mais significativos, adotando-se, para tanto, um critério quantitativo. Considere, por exemplo, a decomposição em valores singulares de um núcleo representado matricialmente por $\mathbf{H}_2 \in \mathbb{R}^{N_2 \times N_2}$, a qual tem a forma

$$\mathbf{H}_2 = \mathbf{Q}\mathbf{S}\mathbf{W}^T \quad (4.90)$$

em que $\mathbf{Q}$ e $\mathbf{W}$ são matrizes $N_2 \times N_2$ unitárias e

$$\mathbf{S} = \text{diag}(\sigma_1, \dots, \sigma_R, 0, \dots, 0) \in \mathbb{R}^{N_2 \times N_2} \quad (4.91)$$

onde $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_R$ e R é o posto de $\mathbf{H}_2$. Assim, para se obter os modos dinâmicos principais, uma possibilidade seria escolher os primeiros M vetores $\mathbf{q}_i$ singulares esquerdos (i.e., as M primeiras colunas de $\mathbf{Q}$) cujos valores singulares associados satisfazem

$$\sum_{i=1}^M \sigma_i^2 \geq \gamma \sum_{i=1}^R \sigma_i^2 \quad (4.92)$$

com $0 \ll \gamma \leq 1$. Deve-se observar que, admitindo-se que $\mathbf{H}_2$ é uma matriz simétrica, então os vetores $\mathbf{q}_i$ e $\mathbf{w}_i$ (i.e., as i -ésimas colunas de $\mathbf{Q}$ e $\mathbf{W}$, respectivamente), possuem a mesma direção, isto é

$$\mathbf{q}_i = \kappa_i \mathbf{w}_i, \quad \kappa_i \in \{-1, 1\}. \quad (4.93)$$

Logo, para simplificar a estrutura obtida com esta estratégia, podemos definir as representações vetoriais dos modos dinâmicos principais por $\tilde{\mathbf{q}}_i = \sqrt{\sigma_i} \mathbf{q}_i$.⁶ Com

⁶Note que não há problema nesta definição, uma vez que os valores singulares são, por definição, não-negativos.

isso, como o núcleo de segunda ordem pode ser escrito na forma

$$h_2(k_1, k_2) = \sum_{i=1}^M \kappa_i \tilde{q}_i(k_1) \tilde{q}_i(k_2) \quad (4.94)$$

onde $\tilde{q}_i$ é o filtro linear correspondente ao i -ésimo modo dinâmico principal, então, de (4.85) e (4.86) tem-se que o componente de segunda ordem da saída do modelo MDP é dado por

$$y_2(n) = \sum_{i=1}^M \kappa_i v_i^2(n) \quad (4.95)$$

em que $v_i(n)$ é a saída de $\tilde{q}_i$.

A ideia descrita no parágrafo anterior pode ser estendida para os núcleos de ordem $p > 2$ através do uso da decomposição em valores singulares de alta ordem (HOSVD, do inglês *high-order singular value decomposition*) [82]. Neste caso, a representação tensorial do núcleo de ordem p , $\mathcal{H}_p \in \mathbb{R}^{N_p^{(p)}}$, é escrita na forma

$$\mathcal{H}_p = \mathcal{S} \times_1 \mathbf{W}^{(1)} \times_2 \mathbf{W}^{(2)} \times_3 \dots \times_p \mathbf{W}^{(p)} \quad (4.96)$$

em que cada matriz $\mathbf{W}^{(s)} \in \mathbb{R}^{N_p \times N_p}$ contém os vetores singulares do modo⁷ s de $\mathcal{H}_p$, $\mathcal{S} \in \mathbb{R}^{N_p^{(p)}}$ é o tensor núcleo (do inglês *core tensor*) e $\times_s$ denota o produto de Tucker no modo s (ver Subseção A.1.2 do Apêndice A). Em linhas gerais, (4.96) pode ser vista como uma generalização de (4.90), ou seja, pode-se pensar nas matrizes $\mathbf{W}^{(s)}$ como generalizações de $\mathbf{Q}$ e $\mathbf{W}$ e no tensor núcleo $\mathcal{S}$ como uma generalização da matriz de valores singulares $\mathbf{S}$ [82]. No caso particular em que o tensor é simétrico, tem-se que $\mathbf{W}^{(1)} = \mathbf{W}^{(2)} = \dots = \mathbf{W}^{(p)} = \mathbf{W}$, de maneira que a decomposição (4.96) assume a forma mais simples

$$\mathcal{H}_p = \mathcal{S} \times_1 \mathbf{W} \times_2 \mathbf{W} \times_3 \dots \times_p \mathbf{W}. \quad (4.97)$$

⁷O termo “modo” aqui utilizado não deve ser confundido com a denominação dos filtros lineares que compõem um modelo MDP. Neste caso, ele se refere aos chamados modos de um tensor, que correspondem a matrizes construídas a partir de seus elementos [29, 82].

É importante salientar que uma outra forma de se escrever (4.97) é

$$\mathcal{H}_p = \sum_{k_1=1}^{N_p} \dots \sum_{k_p=1}^{N_p} [\mathcal{S}]_{k_1, \dots, k_p} \mathbf{w}_{k_1} \circ \dots \circ \mathbf{w}_{k_p} \quad (4.98)$$

em que $\mathbf{w}_k$ denota a k -ésima coluna de $\mathbf{W}$, $k \in \{1, \dots, N_p\}$. Em outras palavras, trata-se de um somatório de produtos externos ponderados por componentes do tensor núcleo $\mathcal{S}$. Logo, uma decomposição PARAFAC com N_p fatores paralelos pode ser vista como um caso particular de (4.97) em que o tensor núcleo $\mathcal{S}$ é a identidade

$$[\mathcal{I}]_{k_1, \dots, k_p} = \begin{cases} 1, & k_1 = \dots = k_p \\ 0, & \text{caso contrário} \end{cases} \quad (4.99)$$

o que resulta em uma forma bem mais simples que (4.97). Além disso, em termos práticos, a decomposição PARAFAC envolve um número bem menor de coeficientes que o caso mais geral (4.97), devido à forma particular de $\mathcal{S}$. Contudo, conforme discutido na Subseção 4.4.1.1, o cômputo da decomposição PARAFAC para um tensor qualquer não é um problema bem resolvido na literatura, ao contrário da HOSVD, que pode ser calculada através de expressões como aquelas dadas, por exemplo, em [82].

Para tensores de segunda ordem (simétricos), devido à diagonalidade de $\mathbf{S}$, tem-se que a decomposição em valores singulares de $\mathbf{H}_2$ pode ser escrita também na forma

$$[\mathbf{H}_2]_{k_1, k_2} = \sum_{n=1}^{N_2} [\mathbf{S}]_{n, n} [\mathbf{Q}]_{k_1, n} [\mathbf{W}]_{k_2, n} \quad (4.100)$$

em que a n -ésima coluna de $\mathbf{Q}$ é acoplada somente à n -ésima coluna de $\mathbf{W}$. No caso de alta ordem, isto não é, em geral, válido, sendo que a HOSVD (para tensores simétricos) pode ser escrita como

$$[\mathcal{H}_p]_{k_1, \dots, k_p} = \sum_{n_1=1}^{N_p} \dots \sum_{n_p=1}^{N_p} [\mathcal{S}]_{n_1, \dots, n_p} [\mathbf{W}]_{k_1, n_1} \dots [\mathbf{W}]_{k_p, n_p} \quad (4.101)$$

havendo, portanto, acoplamento entre diferentes colunas de $\mathbf{W}$. Ainda assim, é possível estender o critério usado para seleção dos vetores singulares, como

descrito a seguir. Para tensores de ordem $p > 2$, uma generalização dos valores singulares é dada pela norma de Frobenius de certos subtensores de $\mathcal{S}$ (ver Apêndice A). Mais precisamente, o i -ésimo valor singular de modo⁸ s é dado por [82]

$$\sigma_i^{(s)} = \|\mathcal{S}_{k_s=i}\|_F \quad (4.102)$$

em que $\mathcal{S}_{k_s=i} \in \mathbb{R}^{N_p^{(p-1)}}$ denota o subtensor de $\mathcal{S}$ obtido fixando-se o s -ésimo índice em $k_s = i$. Deve-se observar que, devido à simetria da representação tensorial do núcleo de ordem p , $\mathcal{S}$ também é simétrico, e portanto

$$\sigma_i^{(1)} = \sigma_i^{(2)} = \dots = \sigma_i^{(p)} = \sigma_i \quad (4.103)$$

ou seja, a escolha do modo s é irrelevante. Assim, uma extensão natural do critério anterior é selecionar como modos dinâmicos principais os M primeiros vetores coluna de $\mathbf{W}$ que satisfaçam (4.92), considerando-se a generalização dos valores singulares acima discutida. É importante notar que, por definição da HOSVD, o tensor $\mathcal{S}$ é tal que

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{N_p} \quad (4.104)$$

assim como ocorre no caso de ordem $p = 2$ [82].

Usando-se a ideia discutida no parágrafo anterior, o componente da saída de ordem p do modelo MDP pode ser escrito, de forma análoga a (4.95), como

$$y_p(n) = \sum_{j_1=1}^M \dots \sum_{j_p=j_{p-1}}^M [\tilde{\mathcal{S}}]_{j_1, \dots, j_p} v_{j_1}(n) \dots v_{j_p}(n) \quad (4.105)$$

em que $v_i(n)$ é a saída do filtro linear associado ao i -ésimo modo dinâmico principal e $\tilde{\mathcal{S}}$ é o tensor obtido a partir da triangularização de $\mathcal{S}$, i.e.,

$$[\tilde{\mathcal{S}}]_{k_1, \dots, k_p} = \begin{cases} \sum_{\pi} [\mathcal{S}]_{\pi_1(k), \dots, \pi_p(k)}, & \text{se } k_1 \leq k_2 \leq \dots \leq k_p \\ 0, & \text{caso contrário} \end{cases} \quad (4.106)$$

⁸Idem à nota anterior.

onde o somatório é sobre todas as permutações possíveis $\pi = (\pi_1(k), \dots, \pi_p(k))$ de $k = (k_1, \dots, k_p)$ que produzem sequências distintas. De (4.105) observa-se que os coeficientes da não-linearidade polinomial do modelo MDP resultante são dados, nesse caso, pelos elementos do tensor triangular $\tilde{\mathcal{S}}$.

Para finalizar, uma observação importante deve ser feita. Como foi discutido na apresentação do modelo MDP, de modo geral, busca-se um conjunto reduzido de filtros lineares que, ao terem suas saídas combinadas por uma não-linearidade sem memória adequada, resultem em núcleos capazes de aproximar aqueles associados ao sistema de interesse. Em geral, tais filtros são “compartilhados” entre as expansões dos diversos núcleos, ao contrário do procedimento acima descrito para a obtenção dos modos, em que cada núcleo é decomposto separadamente. No entanto, como visto, tais decomposições possuem fundamentação algébrica, e portanto se espera que sejam eficientes na aproximação dos núcleos. Para um conjunto de núcleos quaisquer, pode ser difícil se obter um único conjunto reduzido de modos que forneça uma expansão eficiente para todos os núcleos.

4.4.2.2 Complexidades estrutural e computacional

A quantificação da redução de complexidade estrutural obtida com o uso do modelo MDP depende não só do número de modos dinâmicos empregados, mas também da forma particular do modelo no que se refere ao acoplamento entre modos, uma vez que isto determinará o número de coeficientes necessários para descrever a não-linearidade sem memória polinomial. Na análise aqui realizada, admitiremos que não há acoplamento entre os modos obtidos a partir dos núcleos de diferentes ordens, em consonância com a estratégia de obtenção dos modos discutida na subseção anterior.

Considere um modelo MDP de ordem P formado por $1 + M_2 + \dots + M_P$ modos dinâmicos, em que M_i , $i \in \{2, \dots, P\}$ é o número de modos obtidos com a abordagem acima descrita. Suponha, ainda, que os modos dinâmicos de ordem p possuem comprimento N_p . Então, o número total de coeficientes dos filtros

lineares é dado por

$$N_1 + \sum_{p=2}^P M_p N_p. \quad (4.107)$$

Para a caracterização da não-linearidade, conforme foi discutido, no caso de segunda ordem são necessários apenas M_2 coeficientes, devido ao desacoplamento entre os modos. Já para $p > 2$, o número de coeficientes necessário para cobrir todas as combinações entre os modos é igual àquele apresentado por núcleo de Volterra de ordem p e extensão de memória M_p . Logo, o total de coeficientes de um modelo MDP $\mathcal{M}$ construído da forma descrita é

$$\Psi_{\text{MDP}}(\mathcal{M}) = N_1 + M_2(N_2 + 1) + \sum_{p=3}^P \left[M_p N_p + \frac{(M_p + p - 1)!}{p!(M_p - 1)!} \right]. \quad (4.108)$$

Assim como realizado nos casos anteriores, computaram-se os totais dados pela expressão acima, considerando-se algumas combinações diferentes das características estruturais do modelo, para que seja fornecida uma ideia quantitativa da redução de complexidade estrutural através de alguns exemplos. Por simplicidade, admitiu-se que $M_2 = \dots = M_P = M$ e $N_1 = N_2 = \dots = N_P = N$. Os resultados obtidos são mostrados na Tabela 4.4, sendo comparados com aqueles de filtros de Volterra convencionais.

P	$N = 30$			$N = 50$			$N = 70$		
	$M = 10$	$M = 20$	FV	$M = 10$	$M = 20$	FV	$M = 10$	$M = 20$	FV
1	30	30	30	50	50	50	70	70	70
2	340	650	495	560	1.070	1.325	780	1.490	2.555
3	860	2.790	5.455	1.280	3.610	23.425	1.700	4.430	62.195
4	1.875	12.245	46.375	2.495	13.465	316.250	3.115	14.685	1.150.625
5	4.177	55.349	324.631	4.997	56.969	3.478.760	5.817	58.589	17.259.389

Tabela 4.4: Comparação entre os números de coeficientes de modelos de modos dinâmicos principais e de filtros de Volterra convencionais.

A partir dos dados mostrados, pode-se ver que, com o aumento da memória e da ordem do modelo, o uso do modelo MDP possibilita uma redução considerável da complexidade estrutural, admitindo-se, evidentemente, que o número de modos dinâmicos principais é baixo em comparação com a extensão da memória do modelo. No caso em que $N = 30$, por exemplo, vê-se que um modelo MDP de ordem $P = 2$ com $M = 20$ modos dinâmicos possui um número de parâme-

tros maior que um FV convencional, assim como se observou para um modelo PARAFAC-Volterra com $R = 20$, $P = 2$ e $N = 30$. Portanto, o modelo MDP não traz, neste caso, vantagem em relação à eficiência estrutural.

Em termos de complexidade computacional, o cálculo dos componentes de primeira e segunda ordem da saída envolve o mesmo número de operações que aquele dado para o modelo PARAFAC-Volterra. Já com relação aos núcleos de ordem $p > 2$, a complexidade do modelo MDP para um mesmo número de filtros lineares empregados é maior, devido ao acoplamento que há entre suas saídas (i.e., à presença de produtos cruzados entre tais saídas na não-linearidade). Neste caso, com o uso de uma implementação direta para as filtragens lineares, tem-se um total de $O(M_p N_p + T_p)$ operações por amostra, em que T_p é o custo da avaliação do componente de ordem p não-linearidade, F_p , sendo dado por

$$T_p = D_p(p-1) + D_p + D_p - 1 = D_p(p+1) - 1 \quad (4.109)$$

onde

$$D_p = \frac{(M_p + p - 1)!}{p!(M_p - 1)!}. \quad (4.110)$$

O primeiro termo da soma em (4.109) se refere ao cálculo dos produtos cruzados, o segundo ao produto de cada termo da expansão de F_p por seu respectivo coeficiente e o terceiro à soma de todos os termos. Se a implementação dos filtros lineares é feita no domínio da TFD usando-se blocos de tamanho L , então, nesse caso, o cômputo de $y_p(n)$ requer $O(M_p(L + N_p) \log(L + N_p) + LT_p)$ por bloco de L amostras.

É importante salientar que a análise acima realizada se refere ao pior caso, uma vez que, na prática, não necessariamente F_p contém efetivamente um termo para cada produto cruzado de ordem p possível entre as saídas dos filtros. Além disso, caso os termos de F_p possuam fatores em comum – como possuem, por exemplo, os termos $\alpha_1 v_1 v_2 v_3$ e $\alpha_2 v_1 v_2 v_4$, sendo v_i a saída do i -ésimo filtro linear e α_1, α_2 coeficientes da expansão de F_3 –, pode-se reaproveitar alguns produtos computados de modo a se reduzir o primeiro termo de (4.109).

Capítulo 5

Derivação eficiente de filtros de Volterra de referência

A ideia central da metodologia proposta neste trabalho consiste no uso de informações fornecidas por um modelo físico do sistema de interesse. Mais especificamente, obtém-se um conjunto de núcleos de Volterra de referência a partir de tal modelo, os quais podem ser explorados para se avaliar a adequação de diferentes estruturas ao modelamento do sistema em questão, como filtros de Volterra convencionais, modelos orientados a blocos e estruturas de Volterra alternativas.

Neste capítulo, será mostrado como núcleos de Volterra de referência podem ser calculados a partir de um modelo físico invariante no tempo com o uso do método de bilinearização de Carleman, o qual se mostrou eficaz e computacionalmente conveniente para tal finalidade. Uma implementação eficiente deste método, proposta pelos autores recentemente [33], é descrita e sua complexidade computacional é comparada com a da implementação direta. Além disso, apontam-se os problemas envolvidos no uso prático de outros três métodos de derivação de representações de Volterra: os métodos da função de fluxo e da expansão em série de Taylor dos núcleos, baseados no conteúdo de [31, Cap. 3], e o método da exponencial crescente [32].

Para a obtenção dos núcleos de referência, considera-se a presença de filtros de reconstrução e anti-rebatimento e de conversores analógico/digital e digital/analógico, os quais são elementos tipicamente envolvidos no contexto de aplicações de processamento digital de sinais. Além disso, mostra-se que tais núcleos de referência são úteis mesmo em situações que envolvem fatores ambientais,

como reverberação acústica.

5.1 Cálculo de núcleos de referência em aplicações de PDS

5.1.1 Definição do problema

Os modelos matemáticos empregados em aplicações de PDS são, naturalmente, realizações de formalismos de tempo discreto. Assim, conceitualmente, o modelamento de um sistema físico de tempo contínuo neste contexto, o qual denotaremos pelo operador $\mathbf{V}$, corresponde à busca de um modelo de tempo discreto capaz de representar a estrutura mostrada na Figura 5.1. Nela, a excitação do sistema de interesse, $u_c(t)$, é produzida a partir da sequência de entrada, $u(n)$, pela ação de um conversor digital/analogico seguido de um filtro de reconstrução, $G_1(j\Omega)$. Analogamente, a resposta produzida por $\mathbf{V}$ é submetida a um filtro anti-rebatimento, $G_2(j\Omega)$, e subsequentemente amostrada, resultando na sequência de saída $y(n)$.

Figura 5.1: Estrutura considerada para obtenção dos núcleos de referência do sistema físico $\mathbf{V}$ no contexto de aplicações de PDS.

A motivação de se considerarem os filtros G_1 e G_2 é como descrita a seguir. Primeiro, note que em algumas aplicações, como a de cancelamento de eco acústico (ver Subseção 1.1.2), de fato há a presença de um arranjo de dispositivos físicos análogo à estrutura representada na Figura 5.1. Sendo assim, para que o modelo de referência seja realista, é preciso considerar as características dos filtros envolvidos. Segundo, mesmo em aplicações que envolvem apenas sinais de tempo discreto, como a de simulação digital de dispositivos de áudio (descrita na Subseção 1.1.3), para fins práticos é importante que o modelo resultante incorpore ao menos o efeito do filtro anti-rebatimento G_2 , uma vez que as distorções não-lineares produzidas pela ação de $\mathbf{V}$ podem conter componentes de frequências

maiores que $\Omega_s/2$, sendo Ω_s a frequência de amostragem adotada.

De acordo com o que foi discutido acima, a derivação de um filtro de Volterra de referência requer a determinação dos núcleos de tempo discreto que constituem a representação de Volterra do operador associado à estrutura completa da Figura 5.1, o qual denotaremos por V_d . Com o uso das expressões dadas na Subseção 3.1.7, pode-se verificar que, se escrevermos a representação de Volterra do sistema (de tempo contínuo) constituído pela conexão em série de G_1 , V e G_2 como

$$y(t) = \sum_{p=1}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \tilde{h}_p(\tau_1, \dots, \tau_p) u(t - \tau_1) \dots u(t - \tau_p) d\tau_1 \dots d\tau_p \quad (5.1)$$

então as FRFGs dos núcleos que a compõem são dadas por

$$\tilde{H}_p(j\Omega_1, \dots, j\Omega_p) = \left[\prod_{q=1}^p G_1(j\Omega_q) \right] H_p(j\Omega_1, \dots, j\Omega_p) G_2(j\Omega_1 + \dots + j\Omega_p) \quad (5.2)$$

em que $H_p(j\Omega_1, \dots, j\Omega_p)$ é a FRFG de ordem p de V , o qual se supõe possuir uma representação de Volterra com núcleos absolutamente somáveis.

Devido à presença do filtro de reconstrução G_1 na saída do conversor D/A, pode-se escrever, sem perda de generalidade,

$$u(t) = T_s \sum_{k=-\infty}^{\infty} u(k) \delta(t - kT_s) \quad (5.3)$$

onde T_s é o intervalo de amostragem correspondente a Ω_s , i.e., $T_s = 2\pi/\Omega_s$. Logo, temos que o componente de ordem p da saída de tempo discreto é dado por

$$y_p(n) = y_p(t = nT_s) \quad (5.4)$$

$$= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \tilde{h}_p(\tau_1, \dots, \tau_p) u(nT_s - \tau_1) \dots u(nT_s - \tau_p) d\tau_1 \dots d\tau_p \quad (5.5)$$

$$= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \tilde{h}_p(\tau_1, \dots, \tau_p) \prod_{i=1}^p \left[T_s \sum_{k_i=-\infty}^{\infty} u(k_i) \delta(nT_s - \tau_i - k_i T_s) \right] d\tau_1 \dots d\tau_p \quad (5.6)$$

em que $y_p(t)$ é o componente de ordem p de $y(t)$. Definindo-se

$$v_p(k_1, \dots, k_p) = T_s^p \tilde{h}_p(k_1 T_s, \dots, k_p T_s) \quad (5.7)$$

resulta

$$y_p(n) = \sum_{k_1=-\infty}^{\infty} \dots \sum_{k_p=-\infty}^{\infty} v_p(n - k_1, \dots, n - k_p) u(k_1) \dots u(k_p) \quad (5.8)$$

$$= \sum_{k_1=-\infty}^{\infty} \dots \sum_{k_p=-\infty}^{\infty} v_p(k_1, \dots, k_p) u(n - k_1) \dots u(n - k_p). \quad (5.9)$$

Em outras palavras, o uso do método de discretização que corresponde à invariância dos núcleos, o qual foi apresentado na Subseção 3.1.10.2, conduz a um modelo de tempo discreto da estrutura V_d .

Por outro lado, se admitirmos que o filtro G_1 é tal que

$$|\Omega| > \frac{\pi}{T_s} \implies G_1(j\Omega) = 0 \quad (5.10)$$

então, de (5.2) conclui-se que

$$|\Omega_i| > \frac{\pi}{T_s} \text{ para algum } i \in \{1, \dots, p\} \implies \tilde{H}_p(j\Omega_1, \dots, j\Omega_p) = 0. \quad (5.11)$$

Consequentemente, como discutido no final da Subseção 3.1.10.2, tem-se que nestas condições o método de invariância dos núcleos é equivalente à forma de discretização que corresponde a impor que as FRFGs de tempo discreto do modelo de Volterra resultante sejam tais que

$$V_p(e^{j\omega_1}, \dots, e^{j\omega_p}) = \tilde{H}_p\left(j\frac{\omega_1}{\pi}\Omega_N, \dots, j\frac{\omega_p}{\pi}\Omega_N\right) \quad (5.12)$$

para $-\pi \leq \omega_i \leq \pi$, $i = 1, \dots, p$, onde $\Omega_N = \Omega_s/2$ é a frequência de Nyquist.

Assim, supondo que Ω_s é suficientemente grande para representar as características de interesse do modelo físico, então o cômputo das anti-transformadas de (5.12) possibilita obter um modelo de tempo discreto análogo a V na banda de interesse, levando-se em conta também as características dos filtros G_1 e G_2

usados na prática. Note que o uso desta abordagem, ao invés do cômputo dos núcleos diretamente no domínio do tempo, é conveniente pois permite incorporar facilmente as propriedades dos filtros G_1 e G_2 , uma vez que elas são usualmente especificadas no domínio da frequência.

Com base no conteúdo acima, vê-se que a determinação dos núcleos do filtro de Volterra de referência requer que se disponha das FRFGs de V e também das respostas em frequência de G_1 e G_2 . Para fins de derivação de tais núcleos, vamos admitir que tais filtros são conhecidos, o que em geral é válido em um contexto prático. Já em relação a V , adota-se o pressuposto de que se dispõe de um modelo físico constituído por equações de espaço de estados. Sendo assim, o problema se reduz à derivação de expressões para as FRFGs de V a partir de tal modelo, assunto que será abordado nas demais seções deste capítulo.

No restante desta seção, discutiremos aspectos adicionais importantes envolvidos no problema acima descrito e, subsequentemente, forneceremos uma descrição detalhada da forma que se supõe para o modelo físico do sistema de interesse.

5.1.2 Fatores ambientais

Uma questão que pode surgir em relação à estrutura discutida na subseção anterior, a qual é mostrada na Figura 5.1, é a de como se pode levar em conta a presença de fatores ambientais envolvidos em certas aplicações. Por exemplo, na aplicação de cancelamento de eco acústico, a saída do alto-falante (que corresponde ao sistema de interesse, V) sofre a ação da resposta acústica do ambiente em que o dispositivo é utilizado. Portanto, na realidade, o sinal submetido ao filtro anti-rebatimento é o resultado de uma convolução entre a saída do alto-falante e a resposta impulsiva que caracteriza a acústica do ambiente.

Há basicamente duas maneiras de se lidar com a presença de fatores ambientais que podem ser representados por um sistema linear, como é o caso do exemplo citado no parágrafo anterior. A primeira delas seria considerar um novo filtro, $G_A(j\Omega)$, posicionado após V , o qual representa tais fatores ambientais de interesse.

Com isso, a equação (5.2) teria de ser reescrita na forma

$$\begin{aligned} \tilde{H}_p(j\Omega_1, \dots, j\Omega_p) = & \left[\prod_{q=1}^p G_1(j\Omega_q) \right] H_p(j\Omega_1, \dots, j\Omega_p) \\ & \times G_A(j\Omega_1 + \dots + j\Omega_p) G_2(j\Omega_1 + \dots + j\Omega_p). \end{aligned} \quad (5.13)$$

Isto corresponde a admitir que a estrutura que o filtro de Volterra de referência deve modelar é como aquela mostrada na Figura 5.2.

Quando se adota esta alternativa, se a anti-transformada de $G_A(j\Omega)$ corresponde a uma resposta impulsiva cujo decaimento é lento, então usualmente a extensão da memória dos núcleos resultantes será longa, o que implica na necessidade de se computar e armazenar tensores com um número elevado de elementos. Tal fato corresponde a uma desvantagem significativa especialmente se núcleos de referência de alta ordem são desejados. Por outro lado, esta alternativa pode ser interessante justamente por permitir uma estimativa do número de coeficientes envolvidos no filtro de Volterra quando ele é utilizado para modelar o sistema não-linear e os fatores ambientais conjuntamente (admitindo-se, obviamente, que $G_A(j\Omega)$ é conhecido).

Figura 5.2: Inclusão de um filtro linear que representa fatores ambientais, $G_A(j\Omega)$, na estrutura da Figura 5.1.

Na prática, contudo, não é necessário que os núcleos incorporem as características dos fatores ambientais. Uma outra maneira de se lidar com a presença deles consiste em considerar que o filtro de Volterra obtido como descrito na subseção anterior será conectado em série com um filtro linear de tempo discreto, $\tilde{G}_A(z)$, o qual modela tais fatores. A estrutura que corresponde a este esquema é mostrada na Figura 5.3. Nela, a saída do operador que corresponde à estrutura da Figura 5.1 é denotada por $y_V(n)$.

Seja Ω_N a frequência de Nyquist, i.e., $\Omega_N = \Omega_s/2$, em que $\Omega_s = 2\pi/T_s$ é a frequência de amostragem escolhida. No que segue, vamos demonstrar que, se a

Figura 5.3: Conexão da estrutura da Figura 5.1 em série com um filtro linear de tempo discreto que representa fatores ambientais, $\tilde{G}_A(z)$.

resposta em frequência do filtro $\tilde{G}_A(z)$ é tal que

$$\tilde{G}_A(e^{j\omega}) = G_A\left(j\frac{\omega}{\pi}\Omega_N\right) \quad (5.14)$$

para $|\omega| \leq \pi$ e o filtro anti-rebatimento G_2 satisfaz

$$|\Omega| > \Omega_N = \frac{\pi}{T_s} \implies G_2(j\Omega) = 0 \quad (5.15)$$

então as estruturas das Figuras 5.2 e 5.3 são, de fato, equivalentes. Em outras palavras, as sequências de saída $y(n)$ produzidas por ambas estruturas são idênticas desde que $\tilde{G}_A(z)$ corresponda a uma versão em tempo discreto de $G_A(j\Omega)$ na banda de interesse e G_2 elimine quaisquer componentes de frequências que não se pode representar com uma taxa de amostragem de $1/T_s$.

Para tanto, será necessário utilizar o resultado dado pelo teorema abaixo.

Teorema 5.1 A conexão em série de um filtro de Volterra cuja relação entrada/saída é dada por

$$y_V(n) = \sum_{p=1}^P \sum_{k_1=0}^{N_p-1} \dots \sum_{k_p=k_{p-1}}^{N_p-1} h_p(k_1, \dots, k_p) u(n-k_1) \dots u(n-k_p) \quad (5.16)$$

com um filtro linear de tempo discreto com resposta em frequência $\tilde{G}_A(e^{j\omega})$ é equivalente a um filtro de Volterra com P núcleos cujas FRFGs são dadas por

$$W_p(e^{j\omega_1}, \dots, e^{j\omega_p}) = H_p(e^{j\omega_1}, \dots, e^{j\omega_p}) \tilde{G}_A(e^{j(\omega_1 + \dots + \omega_p)}) \quad (5.17)$$

em que H_p é a FRFG associada ao núcleo h_p , com $p \in \{1, \dots, P\}$.

Demonstração. Inicialmente, temos que a saída da conexão em série entre o FV

e o filtro linear é dada por

$$y(n) = \sum_{k=-\infty}^{\infty} \tilde{g}_A(k) y_V(n-k) \quad (5.18)$$

onde $\tilde{g}_A(n)$ é a resposta impulsiva do filtro linear. Usando (5.16), obtemos

$$y(n) = \sum_{k=-\infty}^{\infty} \tilde{g}_A(k) \sum_{p=1}^P \sum_{k_1=0}^{N_p-1} \dots \sum_{k_p=k_{p-1}}^{N_p-1} h_p(k_1, \dots, k_p) u(n-k-k_1) \dots u(n-k-k_p) \quad (5.19)$$

$$= \sum_{p=1}^P \sum_{k=-\infty}^{\infty} \sum_{k_1=0}^{N_p-1} \dots \sum_{k_p=k_{p-1}}^{N_p-1} \tilde{g}_A(k) h_p(k_1, \dots, k_p) u(n-k-k_1) \dots u(n-k-k_p). \quad (5.20)$$

Fazendo $m_i \triangleq k + k_i$, com $i \in \{1, \dots, p\}$, temos

$$y(n) = \sum_{p=1}^P \sum_{m_1=-\infty}^{\infty} \dots \sum_{m_p=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \tilde{g}_A(k) h_p(m_1-k, \dots, m_p-k) \times u(n-m_1) \dots u(n-m_p) \quad (5.21)$$

$$= \sum_{p=1}^P \sum_{m_1=-\infty}^{\infty} \dots \sum_{m_p=-\infty}^{\infty} w_p(m_1, \dots, m_p) u(n-m_1) \dots u(n-m_p) \quad (5.22)$$

em que os novos núcleos w_p são dados por

$$w_p(m_1, \dots, m_p) = \sum_{k=-\infty}^{\infty} \tilde{g}_A(k) h_p(m_1-k, \dots, m_p-k) \quad (5.23)$$

com $p \in \{1, \dots, P\}$. Logo, as FRFGs a eles associadas resultam de

$$W_p(e^{j\omega_1}, \dots, e^{j\omega_p}) = \sum_{m_1=-\infty}^{\infty} \dots \sum_{m_p=-\infty}^{\infty} w_p(m_1, \dots, m_p) e^{-j\omega_1 m_1} \dots e^{-j\omega_p m_p} \quad (5.24)$$

$$= \sum_{m_1=-\infty}^{\infty} \dots \sum_{m_p=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \tilde{g}_A(k) h_p(m_1-k, \dots, m_p-k) \times e^{-j\omega_1 m_1} \dots e^{-j\omega_p m_p}. \quad (5.25)$$

Então, definindo-se $k_i \triangleq m_i - k$, com $i \in \{1, \dots, p\}$, obtém-se

$$W_p(e^{j\omega_1}, \dots, e^{j\omega_p}) = \sum_{k_1=-\infty}^{\infty} \dots \sum_{k_p=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \tilde{g}_A(k) h_p(k_1, \dots, k_p) \times e^{-j\omega_1(k+k_1)} \dots e^{-j\omega_p(k+k_p)} \quad (5.26)$$

$$= \sum_{k_1=-\infty}^{\infty} \dots \sum_{k_p=-\infty}^{\infty} h_p(k_1, \dots, k_p) e^{-j\omega_1 k_1} \dots e^{-j\omega_p k_p} \times \sum_{k=-\infty}^{\infty} \tilde{g}_A(k) e^{-j(\omega_1 + \dots + \omega_p)k} \quad (5.27)$$

$$= H_p(e^{j\omega_1}, \dots, e^{j\omega_p}) \tilde{G}_A(e^{j(\omega_1 + \dots + \omega_p)}). \quad (5.28)$$

■

O Teorema 5.1 pode ser visto como uma versão de tempo discreto da relação descrita na Subseção 3.1.7 relativa à interconexão em série de um sistema de Volterra com um filtro linear.

Uma vez obtido este resultado, vamos então proceder à demonstração da equivalência entre as estruturas das Figuras 5.2 e 5.3. Observe que, se o princípio de modelamento em tempo discreto apresentado na subseção anterior é aplicado à estrutura da Figura 5.2, então teremos um filtro de Volterra de referência cujas FRFGs são dadas por

$$\begin{aligned} \tilde{V}_p(e^{j\omega_1}, \dots, e^{j\omega_p}) &= \left[\prod_{q=1}^p G_1 \left(j \frac{\omega_q}{\pi} \Omega_N \right) \right] H_p \left(j \frac{\omega_1}{\pi} \Omega_N, \dots, j \frac{\omega_p}{\pi} \Omega_N \right) \\ &\times G_A \left(j \frac{\omega_1}{\pi} \Omega_N + \dots + j \frac{\omega_p}{\pi} \Omega_N \right) G_2 \left(j \frac{\omega_1}{\pi} \Omega_N + \dots + j \frac{\omega_p}{\pi} \Omega_N \right) \end{aligned} \quad (5.29)$$

para $|\omega_i| \leq \pi$, $i = 1, \dots, p$.

Agora, se considerarmos as FRFGs resultantes da conexão em série do FV de referência descrito na subseção anterior com o filtro de tempo discreto que modela os fatores ambientais, como representado na Figura 5.3, então, de acordo com o Teorema 5.1, elas são dadas por

$$W_p(e^{j\omega_1}, \dots, e^{j\omega_p}) = V_p(e^{j\omega_1}, \dots, e^{j\omega_p}) \tilde{G}_A(e^{j(\omega_1 + \dots + \omega_p)}). \quad (5.30)$$

Usando (5.2) e (5.12), temos

$$W_p(e^{j\omega_1}, \dots, e^{j\omega_p}) = \left[\prod_{q=1}^p G_1 \left(j \frac{\omega_q}{\pi} \Omega_N \right) \right] H_p \left(j \frac{\omega_1}{\pi} \Omega_N, \dots, j \frac{\omega_p}{\pi} \Omega_N \right) \\ \times G_2 \left(j \frac{\omega_1}{\pi} \Omega_N + \dots + j \frac{\omega_p}{\pi} \Omega_N \right) \tilde{G}_A(e^{j(\omega_1 + \dots + \omega_p)}) \quad (5.31)$$

para $|\omega_i| \leq \pi$, $i = 1, \dots, p$. Note que, como o somatório de frequências $\omega_1 + \dots + \omega_p$ que compõe o argumento de $\tilde{G}_A$ pode assumir valores maiores, em módulo, que π , então a condição (5.14) não é suficiente para se afirmar que (5.31) é equivalente a (5.29). Todavia, se o filtro anti-rebatimento satisfizer (5.15), não há preocupação em relação a tais valores, uma vez que, neste caso,

$$|\omega_1 + \dots + \omega_p| > \pi \implies W_p(e^{j\omega_1}, \dots, e^{j\omega_p}) = 0 \quad (5.32)$$

e, por sua vez,

$$|\Omega_1 + \dots + \Omega_p| > \frac{\pi}{T_s} \implies \tilde{V}_p(e^{j\omega_1}, \dots, e^{j\omega_p}) = 0. \quad (5.33)$$

Logo, de (5.14), (5.15), (5.29) e (5.31) tem-se que

$$W_p(e^{j\omega_1}, \dots, e^{j\omega_p}) \equiv \tilde{V}_p(e^{j\omega_1}, \dots, e^{j\omega_p}) \quad (5.34)$$

para $|\omega_i| \leq \pi$. Em outras palavras, sob tais condições, as estruturas das Figuras 5.2 e 5.3 são, de fato, equivalentes.

Dessa maneira, restringindo-nos à banda de interesse, a resposta acústica pode ser modelada por um filtro de tempo discreto posicionado após a conversão A/D, produzindo-se o mesmo resultado que teríamos considerando núcleos que já incluem o efeito desta resposta. Então, conclui-se que a formulação do problema apresentada na subseção anterior é válida mesmo em situações que envolvem fatores ambientais, uma vez que eles podem ser modelados separadamente, caso necessário. Por fim, deve-se ressaltar que, se desejado, eles podem também ser absorvidos pelos núcleos na derivação do filtro de Volterra de referência, como foi discutido acima.

Através de um raciocínio análogo ao que foi desenvolvido nesta subseção, é possível mostrar também que a presença de fatores ambientais representados por um filtro linear $G_A(s)$ posicionado antes do sistema de interesse, V , pode ser levada em conta introduzindo-se um filtro linear de tempo discreto antes do conversor D/A. Para tanto, novamente é necessário que tal filtro, $\tilde{G}_A(z)$, satisfaça (5.14), o que neste caso é suficiente uma vez que não ocorre um somatório de argumentos $\omega_1 + \dots + \omega_p$ na FRFG de tempo discreto resultante. O motivo disto é que, analogamente ao caso de tempo contínuo discutido na Subseção 3.1.7, a introdução de um filtro linear de tempo discreto antes da estrutura da Figura 5.1 faz com que as FRFGs resultantes incorporem produtórios na forma $\tilde{G}_A(e^{j\omega_1}) \dots \tilde{G}_A(e^{j\omega_p})$ ao invés de um termo na forma $\tilde{G}_A(e^{j\omega_1 + \dots + j\omega_p})$, o qual ocorre no caso acima considerado, conforme descrito por (5.30).

5.1.3 Forma considerada para o modelo físico

Neste trabalho, consideraremos que se dispõe de um modelo físico invariante no tempo constituído de equações de espaço de estados na forma linear-analítica [31, 32, 36]

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)) + \mathbf{g}(\mathbf{x}(t))u(t), \quad t \geq 0 \text{ e } \mathbf{x}(0) = \mathbf{x}_0 \quad (5.35)$$

$$y(t) = \lambda(\mathbf{x}(t)) \quad (5.36)$$

onde $\mathbf{x} \in X \subset \mathbb{R}^n$ é o vetor de estados e a entrada e a saída são escalares. Tais equações, as quais já foram mencionadas no Capítulo 3, são assim chamadas devido a (5.35) ser afim¹ na entrada e também por se supor que tanto os campos vetoriais $\mathbf{f} : X \rightarrow X$ e $\mathbf{g} : X \rightarrow X$ quanto a função $\lambda : X \rightarrow \mathbb{R}$ são analíticos no conjunto aberto X .

As razões pelas quais são considerados modelos físicos na forma (5.35)–(5.36) são como descritas a seguir. Primeiro, ela simplifica em alguns aspectos a derivação dos núcleos, em comparação com uma forma irrestrita para as equações.

¹Note que a terminologia “linear-analítica”, a qual é normalmente usada, é imprecisa, uma vez que, na realidade, trata-se de uma forma afim na entrada.

Segundo, ela é capaz de representar uma classe ampla de sistemas reais que ocorrem em problemas de engenharia [36]. Terceiro e último, como discutido na Subseção 3.1.8, sistemas lineares-analíticos sempre admitem uma representação de Volterra convergente para um conjunto de entradas limitadas, desde que exista solução para o caso não forçado de (5.35) [32].

Uma observação importante sobre o sistema (5.35)–(5.36) é que, para que ele seja de fato invariante no tempo, é preciso que o estado inicial considerado, $\mathbf{x}_0$, seja um ponto fixo de $\mathbf{f}$, i.e., $\mathbf{f}(\mathbf{x}_0) = \mathbf{0}$. Note que, caso isso não seja válido, então é fácil encontrar um par de entradas $u(t)$ e $v(t) = u(t - t_0)$ em face das quais as respectivas trajetórias percorridas pelo sistema, $\mathbf{x}_u(t)$ e $\mathbf{x}_v(t)$, sejam tais que, em geral, $\mathbf{x}_v(t) \neq \mathbf{x}_u(t - t_0)$. Por consequência, a condição (2.6) não seria satisfeita.

Com o intuito de facilitar ainda mais a apresentação dos métodos, duas simplificações adicionais podem ser feitas em relação ao modelo físico dado por (5.35)–(5.36), as quais não implicam em perda de generalidade no caso em que tal modelo é invariante no tempo. A primeira delas consiste em admitir que λ é linear em $\mathbf{x}$, i.e.,

$$\lambda(\mathbf{x}) = \mathbf{c}^T \mathbf{x}. \quad (5.37)$$

Note que, caso a função λ original não satisfaça tal condição, então, derivando a saída em relação a t , temos

$$\dot{y}(t) = (\nabla_{\mathbf{x}} \lambda)^T \dot{\mathbf{x}}(t) \quad (5.38)$$

$$= (\nabla_{\mathbf{x}} \lambda)^T [\mathbf{f}(\mathbf{x}(t)) + \mathbf{g}(\mathbf{x}(t))u(t)] \quad (5.39)$$

$$= f_y(\mathbf{x}(t)) + g_y(\mathbf{x}(t))u(t) \quad (5.40)$$

em que $f_y(\mathbf{x}(t)) = (\nabla_{\mathbf{x}} \lambda)^T \mathbf{f}(\mathbf{x}(t))$ e $g_y(\mathbf{x}(t)) = (\nabla_{\mathbf{x}} \lambda)^T \mathbf{g}(\mathbf{x}(t))$. Como (5.40) possui uma forma igual à de (5.35), então o vetor de estados original pode ser estendido com um componente adicional x_{n+1} cuja derivada em relação a t é dada pelo lado direito de (5.40). Assim, fazendo-se $y(t) = \mathbf{e}_{n+1}^T \tilde{\mathbf{x}}(t)$, em que $\tilde{\mathbf{x}}(t)$ é o vetor de estados estendido e $\mathbf{e}_{n+1}$ o $(n+1)$ -ésimo vetor da base canônica do $\mathbb{R}^{n+1}$, obtém-se um sistema equivalente cuja equação de saída é linear no vetor de estados estendido, como desejado [32].

A segunda simplificação é obtida supondo-se que $\mathbf{x}_0 = \mathbf{0}$. Também não há perda de generalidade introduzida porque, se este não for o caso, definindo-se desta vez $\tilde{\mathbf{x}}(t) = \mathbf{x}(t) - \mathbf{x}_0$ podemos obter o sistema equivalente [32]

$$\dot{\tilde{\mathbf{x}}}(t) = \dot{\mathbf{x}}(t) \quad (5.41)$$

$$= \mathbf{f}(\mathbf{x}(t)) + \mathbf{g}(\mathbf{x}(t))u(t) \quad (5.42)$$

$$= \mathbf{f}(\tilde{\mathbf{x}}(t) + \mathbf{x}_0) + \mathbf{g}(\tilde{\mathbf{x}}(t) + \mathbf{x}_0)u(t) \quad (5.43)$$

$$= \tilde{\mathbf{f}}(\tilde{\mathbf{x}}(t)) + \tilde{\mathbf{g}}(\tilde{\mathbf{x}}(t))u(t), \quad t \geq 0 \text{ e } \tilde{\mathbf{x}}(0) = \mathbf{0} \quad (5.44)$$

$$y(t) = \mathbf{c}^T \tilde{\mathbf{x}}(t) + \mathbf{c}^T \mathbf{x}_0 \quad (5.45)$$

para o qual tal suposição é válida, sendo

$$\tilde{\mathbf{f}}(\tilde{\mathbf{x}}(t)) = \mathbf{f}(\tilde{\mathbf{x}}(t) + \mathbf{x}_0) \quad (5.46)$$

$$\tilde{\mathbf{g}}(\tilde{\mathbf{x}}(t)) = \mathbf{g}(\tilde{\mathbf{x}}(t) + \mathbf{x}_0). \quad (5.47)$$

Desse modo, em conformidade com tais simplificações, o sistema linear-analítico invariante no tempo que consideraremos durante o restante desta dissertação possui a forma

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)) + \mathbf{g}(\mathbf{x}(t))u(t), \quad t \geq 0 \text{ e } \mathbf{x}(0) = \mathbf{0} \quad (5.48)$$

$$y(t) = \mathbf{c}^T \mathbf{x}(t). \quad (5.49)$$

É importante enfatizar que a analiticidade das funções de (5.48) é fundamental, o que se deve ao fato de que a representação de Volterra surge da expansão em série de Taylor de tais funções em torno do estado inicial, tanto no domínio da frequência quanto no do tempo, como será visto.

Apesar da equação diferencial considerada ser afim na entrada, deve-se notar que os métodos apresentados adiante podem também ser utilizados para se derivar os núcleos de sistemas que podem ser escritos como

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)) + \mathbf{g}(\mathbf{x}(t))F[u(t)], \quad t \geq 0 \text{ e } \mathbf{x}(0) = \mathbf{0} \quad (5.50)$$

em que $F(x)$ é uma não-linearidade polinomial sem memória. Para tanto, introduz-se uma nova variável $w(t) = F[u(t)]$ de maneira que, obtendo-se a representação de Volterra para a relação entre $y(t)$ e $w(t)$ (a qual é definida por um sistema linear-analítico), a substituição de $w(t)$ pelo polinômio em $u(t)$ na expressão da série resultante possibilita determinar indiretamente os núcleos que relacionam $y(t)$ a $u(t)$. Uma consequência da existência de um bloco polinomial sem memória na entrada é o aparecimento de impulsos nas expressões que fornecem os núcleos do sistema [32].

5.2 Abordagem da exponencial crescente

A abordagem da exponencial crescente (do inglês *growing exponential approach*), a qual é descrita em [32], consiste em um método analítico que possibilita determinar as FRFGs associadas às equações de estado de um sistema invariante no tempo. Para tanto, explora-se a propriedade apresentada por sistemas de Volterra homogêneos que é mencionada na Subseção 3.1.6, como explicado a seguir.

Deve-se notar que a descrição aqui fornecida é mais detalhada que aquela dada em [32], em que se expõem os passos apenas em linhas gerais. Por exemplo, [32] não esclarece concretamente como o método é aplicado a um sistema cujo vetor de estados possui mais de um componente, o que requer o uso de uma série de potências definida com o uso do produto de Kronecker (ver Subseção A.2.2 do Apêndice A).

Suponha que se deseje determinar as FRFGs de ordens até P associadas ao modelo linear-analítico (5.48)–(5.49). Inicialmente, admite-se que sua entrada é dada por

$$u(t) = \sum_{m=1}^P e^{\lambda_m t} \quad (5.51)$$

com $\text{Re}\{\lambda_1\}, \dots, \text{Re}\{\lambda_P\} > 0$. Comparando (5.51) com a expressão da entrada considerada na Subseção 3.1.6, (3.82), vê-se que a primeira é um caso particular

da segunda em que os coeficientes da combinação linear das exponenciais são todos unitários. Como visto naquela subseção, quando a entrada possui tal forma, o componente de ordem p da saída, o qual é produzido pelo subsistema homogêneo de grau p , pode ser escrito como

$$y_p(t) = \sum_{\substack{i_1, \dots, i_P=0 \\ i_1 + \dots + i_P=p}}^p p! H_p \left(\lambda_1^{(i_1)}, \dots, \lambda_P^{(i_P)} \right) e^{(i_1 \lambda_1 + \dots + i_P \lambda_P)t} \quad (5.52)$$

em que $H_p(s_1, \dots, s_p)$ é a transformada de Laplace simétrica de $h_p(\tau_1, \dots, \tau_p)$, contanto que tal subsistema seja estável.² Como explicado na Subseção 3.1.6, a notação $\lambda_j^{(i_j)}$ denota i_j ocorrências consecutivas do argumento λ_j . Logo, podemos escrever a saída como

$$y(t) = \sum_{p=1}^{\infty} y_p(t) \quad (5.53)$$

$$= \sum_{\substack{i_1, \dots, i_P=0 \\ i_1 + \dots + i_P \geq 1}}^{\infty} H_{i_1, \dots, i_P}^* (\lambda_1, \dots, \lambda_P) e^{(i_1 \lambda_1 + \dots + i_P \lambda_P)t} \quad (5.54)$$

onde

$$H_{i_1, \dots, i_P}^* (\lambda_1, \dots, \lambda_P) = (i_1 + \dots + i_P)! H_{i_1 + \dots + i_P} \left(\lambda_1^{(i_1)}, \dots, \lambda_P^{(i_P)} \right). \quad (5.55)$$

Observe que em (5.54) todos os componentes da saída são dados por um único somatório, o qual contém todos produtórios de todas as ordens possíveis das exponenciais da entrada, de maneira que cada produtório é ponderado pelo coeficiente que o multiplica em (5.52), de acordo com sua ordem.

Agora, considerando a equação de saída do sistema, (5.49), juntamente com (5.54), temos que

$$\mathbf{c}^T \mathbf{x}(t) = \sum_{\substack{i_1, \dots, i_P=0 \\ i_1 + \dots + i_P \geq 1}}^{\infty} H_{i_1, \dots, i_P}^* (\lambda_1, \dots, \lambda_P) e^{(i_1 \lambda_1 + \dots + i_P \lambda_P)t}. \quad (5.56)$$

²Como foi discutido na Subseção 3.1.6, a estabilidade do subsistema de ordem p e o fato de que $\text{Re}\{\lambda_1\}, \dots, \text{Re}\{\lambda_P\} > 0$ garantem que não haverá problemas de convergência no cálculo da transformada. A necessidade de que $\text{Re}\{\lambda_m\} > 0$ em (5.51) origina o nome de “método da exponencial crescente”.

Então, se supusermos que a solução de (5.48) possui, para tal entrada, a forma

$$\mathbf{x}(t) = \sum_{\substack{i_1, \dots, i_P=0 \\ i_1 + \dots + i_P \geq 1}}^{\infty} \mathbf{a}_{i_1, \dots, i_P}(\lambda_1, \dots, \lambda_P) e^{(i_1 \lambda_1 + \dots + i_P \lambda_P)t} \quad (5.57)$$

onde $\mathbf{a}_{i_1, \dots, i_P} \in \mathbb{R}^n$, então, temos que

$$H_{i_1, \dots, i_P}^*(\lambda_1, \dots, \lambda_P) = \mathbf{c}^T \mathbf{a}_{i_1, \dots, i_P}(\lambda_1, \dots, \lambda_P). \quad (5.58)$$

Assim, se obtivermos os vetores $\mathbf{a}_{i_1, \dots, i_P}$, podemos calcular as FRFGs desejadas com o uso de (5.55) e (5.58).

Vejamos, então, como se calculam tais vetores associados a um sistema do tipo (5.48)–(5.49). Para tanto, é necessário determinar as expansões em série de Taylor de $\mathbf{f}$ e $\mathbf{g}$ em torno do estado inicial, $\mathbf{0}$, as quais permitem obter uma expressão na forma³

$$\dot{\mathbf{x}}(t) = \sum_{i=1}^{\infty} \mathbf{F}_i \mathbf{x}_i(t) + \sum_{i=0}^{\infty} \mathbf{G}_i \mathbf{x}_i(t) u(t) \quad (5.59)$$

em que os vetores $\mathbf{x}_i(t) \in \mathbb{R}^{n^i}$ são dados por

$$\mathbf{x}_0(t) = 1 \quad (5.60)$$

$$\mathbf{x}_1(t) = \mathbf{x}(t) \quad (5.61)$$

$$\mathbf{x}_i(t) = \mathbf{x}(t) \otimes \mathbf{x}_{i-1}(t) \quad (5.62)$$

onde $\otimes$ denota o produto de Kronecker. Em outras palavras, o vetor $\mathbf{x}_i(t)$ contém como componentes todos os possíveis produtos de ordem i entre os componentes

³A derivação da expansão em série de Taylor de campos vetoriais é descrita na Subseção A.2.2 do Apêndice A.

de $\mathbf{x}(t)$. Por exemplo, o vetor $\mathbf{x}_2(t) \in \mathbb{R}^{n^2}$ resultante desta definição é

$$\mathbf{x}_2(t) = \mathbf{x}_1(t) \otimes \mathbf{x}_1(t) = \begin{bmatrix} x_1^2(t) \\ x_1(t)x_2(t) \\ \vdots \\ x_1(t)x_n(t) \\ x_2(t)x_1(t) \\ x_2^2(t) \\ \vdots \\ x_n^2(t) \end{bmatrix}. \quad (5.63)$$

Já as matrizes $\mathbf{F}_i, \mathbf{G}_i \in \mathbb{R}^{n \times n^i}$ de (5.59) contêm os coeficientes das expansões em série de Taylor de $\mathbf{f}$ e $\mathbf{g}$, respectivamente, ordenados analogamente aos vetores $\mathbf{x}_i(t)$. Observe que, em (5.59), considerou-se que o termo de ordem zero da expansão de $\mathbf{f}$ é nulo, consistentemente com o fato de que $\mathbf{f}(\mathbf{x}_0) = \mathbf{x}_0$, i.e., $\mathbf{f}(\mathbf{0}) = \mathbf{0}$.

Uma vez obtidas as expansões de Taylor de $\mathbf{f}$ e $\mathbf{g}$, substituindo-se (5.51) e (5.57) em (5.59), obtém-se

$$\sum_{\substack{i_1, \dots, i_P=0 \\ i_1 + \dots + i_P \geq 1}}^{\infty} \left(\sum_{p=1}^P i_p \lambda_p \right) \mathbf{a}_{i_1, \dots, i_P}(\lambda_1, \dots, \lambda_P) e^{(i_1 \lambda_1 + \dots + i_P \lambda_P)t} = \sum_{i=1}^{\infty} \mathbf{F}_i \mathbf{x}_i(t) + \sum_{i=0}^{\infty} \mathbf{G}_i \mathbf{x}_i(t) \left(\sum_{p=1}^P e^{\lambda_p t} \right). \quad (5.64)$$

Após isso, igualando-se os coeficientes que multiplicam a exponencial $e^{\lambda_1 t}$, forma-se a equação

$$\lambda_1 \mathbf{a}_{1,0,\dots,0}(\lambda_1, \dots, \lambda_P) = \mathbf{F}_1 \mathbf{a}_{1,0,\dots,0}(\lambda_1, \dots, \lambda_P) + \mathbf{G}_0 \quad (5.65)$$

cujas solução para $\mathbf{a}_{1,0,\dots,0}(\lambda_1, \dots, \lambda_P)$ é

$$\mathbf{a}_{1,0,\dots,0}(\lambda_1) = (\lambda_1 \mathbf{I} - \mathbf{F}_1)^{-1} \mathbf{G}_0 \quad (5.66)$$

em que foram omitidos os argumentos $\lambda_2, \dots, \lambda_P$ de $\mathbf{a}_{1,0,\dots,0}$ porque, como visto, tal vetor depende apenas de λ_1 . Como (5.64) é simétrica em relação aos vetores $\mathbf{a}_{1,0,0,\dots,0}$, $\mathbf{a}_{0,1,0,\dots,0}$, $\dots$, $\mathbf{a}_{0,0,\dots,0,1}$, então, de forma análoga a (5.66),

$$\mathbf{a}_{0,1,0,\dots,0}(\lambda_2) = (\lambda_2 \mathbf{I} - \mathbf{F}_1)^{-1} \mathbf{G}_0 \quad (5.67)$$

$$\mathbf{a}_{0,0,1,0,\dots,0}(\lambda_3) = (\lambda_3 \mathbf{I} - \mathbf{F}_1)^{-1} \mathbf{G}_0 \quad (5.68)$$

e assim por diante.

Agora, usando-se o mesmo procedimento para $e^{(\lambda_1 + \lambda_2)t}$, obtém-se

$$\begin{aligned} (\lambda_1 + \lambda_2) \mathbf{a}_{1,1,0,\dots,0}(\lambda_1, \dots, \lambda_P) = \\ \mathbf{F}_1 \mathbf{a}_{1,1,0,\dots,0}(\lambda_1, \dots, \lambda_P) + \mathbf{F}_2 [\mathbf{a}_{1,0,\dots,0}(\lambda_1) \otimes \mathbf{a}_{0,1,0,\dots,0}(\lambda_2) \\ + \mathbf{a}_{0,1,0,\dots,0}(\lambda_2) \otimes \mathbf{a}_{1,0,\dots,0}(\lambda_1)] + \mathbf{G}_1 [\mathbf{a}_{1,0,\dots,0}(\lambda_1) + \mathbf{a}_{0,1,0,\dots,0}(\lambda_2)]. \end{aligned} \quad (5.69)$$

Resolvendo-se (5.69) para $\mathbf{a}_{1,1,0,\dots,0}(\lambda_1, \dots, \lambda_P)$, tem-se

$$\begin{aligned} \mathbf{a}_{1,1,0,\dots,0}(\lambda_1, \lambda_2) = [(\lambda_1 + \lambda_2) \mathbf{I} - \mathbf{F}_1]^{-1} \{ \mathbf{F}_2 [\mathbf{a}_{1,0,\dots,0}(\lambda_1) \otimes \mathbf{a}_{0,1,0,\dots,0}(\lambda_2) \\ + \mathbf{a}_{0,1,0,\dots,0}(\lambda_2) \otimes \mathbf{a}_{1,0,\dots,0}(\lambda_1)] + \mathbf{G}_1 [\mathbf{a}_{1,0,\dots,0}(\lambda_1) + \mathbf{a}_{0,1,0,\dots,0}(\lambda_2)] \}. \end{aligned} \quad (5.70)$$

Desta forma, substituindo-se (5.66) e (5.67) em (5.70), obtém-se uma solução para $\mathbf{a}_{1,1,0,\dots,0}(\lambda_1, \lambda_2)$ apenas em termos de λ_1, λ_2 e das matrizes $\mathbf{F}_i, \mathbf{G}_i$ que compõem (5.59).

Ao procedermos de maneira análoga para $e^{(\lambda_1 + \dots + \lambda_p)t}$, com $p \in \{3, \dots, P\}$, podemos encontrar uma solução para cada um dos vetores $\mathbf{a}_{1,1,1,0,\dots,0}(\cdot)$, $\mathbf{a}_{1,1,1,1,0,\dots,0}(\cdot)$, $\dots$, $\mathbf{a}_{1,1,\dots,1}(\cdot)$. Finalmente, empregando agora (5.55) e (5.58), temos que as funções de transferência generalizadas simétricas associadas às equações (5.48)–(5.49) são

dadas por [32]

$$H_1(s_1) = \mathbf{c}^T \mathbf{a}_{1,0,\dots,0}(s_1) \quad (5.71)$$

$$H_2(s_1, s_2) = \frac{1}{2!} \mathbf{c}^T \mathbf{a}_{1,1,0,\dots,0}(s_1, s_2) \quad (5.72)$$

$$\vdots$$

$$H_P(s_1, \dots, s_P) = \frac{1}{P!} \mathbf{c}^T \mathbf{a}_{1,1,\dots,1}(s_1, \dots, s_P). \quad (5.73)$$

Logo, as FRFGs desejadas são obtidas fazendo-se a substituição $s_i = j\Omega_i$ nas equações (5.71)–(5.73), com $i \in \{1, \dots, P\}$.

5.2.1 Implementação computacional

Do ponto de vista computacional, o método da exponencial crescente fornece uma forma recursiva de se calcular as FRFGs numericamente. Para tanto, inicialmente, deve-se escolher um conjunto de valores de frequência de interesse, $\Theta = \{\theta_1, \dots, \theta_M\}$. Uma vez definido tal conjunto, determinam-se então os valores dos vetores $\mathbf{a}_{1,0,\dots,0}(j\Omega_1)$ para $\Omega_1 \in \Theta$. Em seguida, todos os valores de $\mathbf{a}_{1,1,0,\dots,0}(j\Omega_1, j\Omega_2)$ tais que $\Omega_1, \Omega_2 \in \Theta$ podem ser calculados usando-se os resultados obtidos para $\mathbf{a}_{1,0,\dots,0}(j\Omega_1)$. Este passo é repetido para os demais vetores $\mathbf{a}$ envolvidos nas equações (5.71)–(5.73) de modo que, ao final, computam-se os valores da FRFG de ordem p através de

$$H_p(j\Omega_1, \dots, j\Omega_p) = \mathbf{c}^T \mathbf{a}_{1,1,\dots,1,0,\dots,0}(j\Omega_1, \dots, j\Omega_p) \quad (5.74)$$

para todas as $\binom{M+p-1}{p}$ combinações distintas de frequências $\Omega_1, \dots, \Omega_p \in \Theta$, com $p \in \{1, \dots, P\}$, que satisfazem a condição de triangularidade $\Omega_1 \leq \Omega_2 \leq \dots \leq \Omega_p$. Note que, devido à simetria das FRFGs, não é necessário computar todos os M^p valores que a FRFG de ordem p assume considerando-se que os índices de frequência pertencem ao conjunto Θ . Ao invés disso, computam-se apenas os valores de $H_p(j\Omega_1, \dots, j\Omega_p)$ para índices $\Omega_1, \dots, \Omega_p$ sobre um domínio triangular, os quais caracterizam completamente a FRFG simétrica.

Uma característica indesejável associada ao método da exponencial crescente

é que o número de termos envolvidos na equação relativa ao vetor $\mathbf{a}$ que fornece a FRFG de ordem p cresce muito rapidamente com p . Por exemplo, as equações associadas aos vetores $\mathbf{a}_{1,0,\dots,0}$ e $\mathbf{a}_{1,1,0,\dots,0}$ possuem apenas dois e cinco termos do lado direito, respectivamente. Já a equação associada a $\mathbf{a}_{1,1,1,0,\dots,0}$ é dada por

$$\begin{aligned}
(\lambda_1 + \lambda_2 + \lambda_3)\mathbf{a}_{1,1,1} = & \mathbf{F}_1\mathbf{a}_{1,1,1} + \mathbf{F}_2[\mathbf{a}_{1,1,0} \otimes \mathbf{a}_{0,0,1} + \mathbf{a}_{0,0,1} \otimes \mathbf{a}_{1,1,0} + \mathbf{a}_{1,0,1} \otimes \mathbf{a}_{0,1,0} \\
& + \mathbf{a}_{0,1,0} \otimes \mathbf{a}_{1,0,1} + \mathbf{a}_{0,1,1} \otimes \mathbf{a}_{1,0,0} + \mathbf{a}_{1,0,0} \otimes \mathbf{a}_{0,1,1}] + \mathbf{F}_3[\mathbf{a}_{1,0,0} \otimes \mathbf{a}_{0,1,0} \otimes \mathbf{a}_{0,0,1} \\
& + \mathbf{a}_{1,0,0} \otimes \mathbf{a}_{0,0,1} \otimes \mathbf{a}_{0,1,0} + \mathbf{a}_{0,1,0} \otimes \mathbf{a}_{1,0,0} \otimes \mathbf{a}_{0,0,1} + \mathbf{a}_{0,1,0} \otimes \mathbf{a}_{0,0,1} \otimes \mathbf{a}_{1,0,0} \\
& + \mathbf{a}_{0,0,1} \otimes \mathbf{a}_{1,0,0} \otimes \mathbf{a}_{0,1,0} + \mathbf{a}_{0,0,1} \otimes \mathbf{a}_{0,1,0} \otimes \mathbf{a}_{1,0,0}] + \mathbf{G}_1[\mathbf{a}_{1,1,0} + \mathbf{a}_{0,1,1} + \mathbf{a}_{1,0,1}] \\
& + \mathbf{G}_2[\mathbf{a}_{1,0,0} \otimes \mathbf{a}_{0,1,0} + \mathbf{a}_{0,1,0} \otimes \mathbf{a}_{1,0,0} + \mathbf{a}_{0,1,0} \otimes \mathbf{a}_{0,0,1} + \mathbf{a}_{0,0,1} \otimes \mathbf{a}_{0,1,0} \\
& + \mathbf{a}_{1,0,0} \otimes \mathbf{a}_{0,0,1} + \mathbf{a}_{0,0,1} \otimes \mathbf{a}_{1,0,0}] \quad (5.75)
\end{aligned}$$

em que, por questões de clareza, os argumentos dos vetores $\mathbf{a}$ são omitidos e exibem-se apenas seus três primeiros subíndices, sendo todos os demais iguais a zero. Como visto acima, trata-se de uma equação bastante extensa, envolvendo 22 termos. Semelhantemente, pode-se mostrar que equação associada a $\mathbf{a}_{1,1,1,1,0,\dots,0}$, por sua vez, contém 85 termos do lado direito. Assim, é nítido que a derivação analítica de FRFGs com ordem $p > 2$ implica em uma laboriosa formulação de equações extensas, o que se agrava com o aumento de p .

Como consequência do problema apontado no parágrafo anterior, a implementação computacional da estratégia acima para a obtenção de FRFGs de ordens quaisquer não é uma tarefa simples. A principal razão disso é que há uma dificuldade considerável envolvida na elaboração de um algoritmo para determinar a equação que fornece a solução para o vetor $\mathbf{a}$ relativo a uma FRFG qualquer de ordem p .

Por outro lado, o método da exponencial crescente é útil em algumas situações como, por exemplo, na obtenção analítica de fórmulas fechadas para as FRFGs associadas a um sistema bilinear (o qual é um caso particular da classe de sistemas lineares-analíticos). Uma vez que tais fórmulas serão necessárias mais adiante para a obtenção dos núcleos de referência de um sistema físico no contexto de aplicações de PDS, mostraremos a seguir como elas podem ser derivadas. Adi-

cionalmente, isto fornecerá um exemplo de aplicação do método da exponencial crescente.

5.2.2 Obtenção das FRFGs de sistemas bilineares

Considere o sistema bilinear invariante no tempo

$$\dot{\mathbf{x}}(t) = \mathbf{F}\mathbf{x}(t) + \mathbf{G}\mathbf{x}(t)u(t) + \mathbf{b}u(t), \quad \mathbf{x}(0) = \mathbf{0} \quad (5.76)$$

$$y(t) = \mathbf{c}^T \mathbf{x}(t) \quad (5.77)$$

e uma entrada na forma

$$u(t) = e^{\lambda_1 t} + e^{\lambda_2 t}. \quad (5.78)$$

Agora, procedendo-se como descrito, suponha que a trajetória do sistema é dada por

$$\mathbf{x}(t) = \sum_{\substack{i_1, i_2=0 \\ i_1+i_2 \geq 1}}^{\infty} \mathbf{a}_{i_1, i_2} e^{(i_1 \lambda_1 + i_2 \lambda_2)t}. \quad (5.79)$$

Se substituirmos (5.78) e (5.79) na equação (5.76), obtemos

$$\sum_{\substack{i_1, i_2=0 \\ i_1+i_2 \geq 1}}^{\infty} (i_1 \lambda_1 + i_2 \lambda_2) \mathbf{a}_{i_1, i_2} e^{(i_1 \lambda_1 + i_2 \lambda_2)t} = \mathbf{F} \sum_{\substack{i_1, i_2=0 \\ i_1+i_2 \geq 1}}^{\infty} \mathbf{a}_{i_1, i_2} e^{(i_1 \lambda_1 + i_2 \lambda_2)t} + \left[\mathbf{G} \sum_{\substack{i_1, i_2=0 \\ i_1+i_2 \geq 1}}^{\infty} \mathbf{a}_{i_1, i_2} e^{(i_1 \lambda_1 + i_2 \lambda_2)t} + \mathbf{b} \right] (e^{\lambda_1 t} + e^{\lambda_2 t}). \quad (5.80)$$

Ao se igualarem os coeficientes que multiplicam termos $e^{\lambda_1 t}$ em (5.80), obtém-se a equação linear

$$\lambda_1 \mathbf{a}_{1,0} = \mathbf{F} \mathbf{a}_{1,0} + \mathbf{b} \quad (5.81)$$

cujas solução é

$$\mathbf{a}_{1,0} = (\lambda_1 \mathbf{I} - \mathbf{F})^{-1} \mathbf{b}. \quad (5.82)$$

Assim, a transformada de Laplace do núcleo de primeira ordem é dada por

$$H_1(s) = \mathbf{c}^T (s\mathbf{I} - \mathbf{F})^{-1} \mathbf{b} \quad (5.83)$$

e, portanto, a FRFG de primeira ordem é

$$H_1(j\Omega) = \mathbf{c}^T (j\Omega \mathbf{I} - \mathbf{F})^{-1} \mathbf{b}. \quad (5.84)$$

Em seguida, a FRFG de segunda ordem pode ser obtida igualando-se os coeficientes dos termos $e^{(\lambda_1 + \lambda_2)t}$ que aparecem em ambos os lados de (5.80). Com isto, temos

$$(\lambda_1 + \lambda_2) \mathbf{a}_{1,1} = \mathbf{F} \mathbf{a}_{1,1} + \mathbf{G} \mathbf{a}_{1,0} + \mathbf{G} \mathbf{a}_{0,1}. \quad (5.85)$$

Substituindo-se $\mathbf{a}_{1,0} = (\lambda_1 \mathbf{I} - \mathbf{F})^{-1} \mathbf{b}$ e $\mathbf{a}_{0,1} = (\lambda_2 \mathbf{I} - \mathbf{F})^{-1} \mathbf{b}$ em (5.85), chega-se a

$$\mathbf{a}_{1,1} = [(\lambda_1 + \lambda_2) \mathbf{I} - \mathbf{F}]^{-1} \mathbf{G} [(\lambda_1 \mathbf{I} - \mathbf{F})^{-1} \mathbf{b} + (\lambda_2 \mathbf{I} - \mathbf{F})^{-1} \mathbf{b}] \quad (5.86)$$

que resulta na FRFG simétrica

$$H_2(j\Omega_1, j\Omega_2) = \frac{1}{2} \mathbf{c}^T [j(\Omega_1 + \Omega_2) \mathbf{I} - \mathbf{F}]^{-1} \mathbf{G} [(j\Omega_1 \mathbf{I} - \mathbf{F})^{-1} + (j\Omega_2 \mathbf{I} - \mathbf{F})^{-1}] \mathbf{b}. \quad (5.87)$$

De modo geral, a aplicação do método da exponencial crescente a um sistema bilinear conduz a FRFGs que podem ser escritas na forma simétrica

$$H_p(j\Omega_1, \dots, j\Omega_p) = \frac{1}{p!} \mathbf{c}^T \mathbf{H}^{(p)}(j\Omega_1, \dots, j\Omega_p) \mathbf{b} \quad (5.88)$$

onde

$$\begin{aligned} \mathbf{H}^{(p)}(\theta_1, \dots, \theta_p) &= [(\theta_1 + \dots + \theta_p) \mathbf{I} - \mathbf{F}]^{-1} \mathbf{G} [\mathbf{H}^{(p-1)}(\theta_1, \theta_2, \dots, \theta_{p-1}) \\ &\quad + \mathbf{H}^{(p-1)}(\theta_1, \theta_2, \dots, \theta_{p-2}, \theta_p) + \dots + \mathbf{H}^{(p-1)}(\theta_2, \theta_3, \dots, \theta_p)] \end{aligned} \quad (5.89)$$

para $p > 1$ e

$$\mathbf{H}^{(1)}(\theta) = (\theta \mathbf{I} - \mathbf{F})^{-1}. \quad (5.90)$$

Diante das expressões acima é nítido que, assim como no caso mais geral linear-analítico, o cômputo das FRFGs associadas a um sistema bilinear pode ser feito recursivamente. Com o uso desta estratégia e explorando-se a simetria de tais funções, pode-se reduzir consideravelmente o esforço computacional envolvido

no cômputo numérico de tensores compostos por pontos delas.

5.3 Método de bilinearização de Carleman

A ideia central do método de bilinearização de Carleman⁴ é representar a dinâmica não-linear do sistema através de um conjunto de equações de espaço de estados na forma bilinear apresentada na Subseção 3.1.2. Para tanto, faz-se uso da expansão em série de Taylor de $\mathbf{f}$ e $\mathbf{g}$ em torno do estado inicial. Assim, uma vez que tais equações são obtidas, seus núcleos de Volterra associados podem ser determinados pela fórmula direta dada na Subseção 3.1.2, ou, caso se deseje uma representação no domínio da frequência, pode-se recorrer às fórmulas das FRFGs de um sistema bilinear dadas na Subseção 5.2.2. Observe que isto é uma vantagem em relação ao método da exponencial crescente, o qual fornece fórmulas apenas para as FRFGs do sistema.

No que segue, descreveremos o método de Carleman através de um desenvolvimento semelhante àquele feito em [32]. Para simplificar a notação, o argumento relativo ao tempo será omitido.

Suponhamos que o modelo físico de interesse é dado por um sistema linear-analítico na forma (5.48)–(5.49) e que se deseja calcular os núcleos de ordens 1 a P . Inicialmente, assim como foi feito para o método da exponencial crescente, calcula-se uma expansão em série de potências da equação (5.48) na forma

$$\dot{\mathbf{x}} = \sum_{i=1}^P \mathbf{F}_i \mathbf{x}_i + \sum_{i=0}^{P-1} \mathbf{G}_i \mathbf{x}_i u \quad (5.91)$$

em que os vetores $\mathbf{x}_i \in \mathbb{R}^{n^i}$ são dados por (5.60)–(5.62) e $\mathbf{F}_i, \mathbf{G}_i$ são matrizes $n \times n^i$, sendo n a dimensão do vetor de estados $\mathbf{x}$. Note que, desta vez, a expansão em série de potências da equação (5.48) foi truncada na ordem P . Logo, o sistema considerado na derivação dos núcleos através do método de Carleman é uma aproximação do modelo físico original. Entretanto, como será discutido

⁴Embora o nome “linearização de Carleman” seja por vezes empregado na literatura, o termo “bilinearização” é mais apropriado, visto que a forma final obtida é bilinear.

adiante nas Subseções 5.3.2 e 5.3.3, os núcleos de ordens 1 a P associados a tal aproximação são *exatamente* iguais àqueles do modelo físico original, ou seja, trata-se de uma aproximação no sentido da série de Volterra truncada na ordem P .

O próximo passo é escrever equações diferenciais para os vetores $\mathbf{x}_i$, $i \in \{2, \dots, P\}$, fazendo-se uso da regra [32, 86]

$$\frac{d}{dt}(\mathbf{x} \otimes \mathbf{x}_{i-1}) = \frac{d}{dt}\mathbf{x} \otimes \mathbf{x}_{i-1} + \mathbf{x} \otimes \frac{d}{dt}(\mathbf{x}_{i-1}). \quad (5.92)$$

Para $i = 2$, temos

$$\frac{d}{dt}\mathbf{x}_2 = \dot{\mathbf{x}} \otimes \mathbf{x} + \mathbf{x} \otimes \dot{\mathbf{x}} \quad (5.93)$$

$$= \left(\sum_{i=1}^P \mathbf{F}_i \mathbf{x}_i + \sum_{i=0}^{P-1} \mathbf{G}_i \mathbf{x}_i u \right) \otimes \mathbf{x} + \mathbf{x} \otimes \left(\sum_{i=1}^P \mathbf{F}_i \mathbf{x}_i + \sum_{i=0}^{P-1} \mathbf{G}_i \mathbf{x}_i u \right) \quad (5.94)$$

$$= \sum_{i=1}^P (\mathbf{F}_i \otimes \mathbf{I}_n + \mathbf{I}_n \otimes \mathbf{F}_i) \mathbf{x}_{i+1} + \sum_{i=0}^{P-1} (\mathbf{G}_i \otimes \mathbf{I}_n + \mathbf{I}_n \otimes \mathbf{G}_i) \mathbf{x}_{i+1} u \quad (5.95)$$

em que $\mathbf{I}_n$ é a matriz identidade $n \times n$ e se usaram a propriedade de produto misto do operador de Kronecker (ver Subseção A.1.4 do Apêndice A) e o fato de que $\mathbf{x} \otimes \mathbf{x}_i = \mathbf{x}_i \otimes \mathbf{x}$. Observe que, descartando-se os termos de ordem maior que P e definindo-se

$$\mathbf{F}_{2,i} = \mathbf{F}_{i-1} \otimes \mathbf{I}_n + \mathbf{I}_n \otimes \mathbf{F}_{i-1}, \quad i \in \{2, \dots, P\} \quad (5.96)$$

$$\mathbf{G}_{2,i} = \mathbf{G}_{i-1} \otimes \mathbf{I}_n + \mathbf{I}_n \otimes \mathbf{G}_{i-1}, \quad i \in \{1, \dots, P-1\} \quad (5.97)$$

a equação resultante pode ser escrita em um forma semelhante a (5.91), i.e.,

$$\frac{d^{(P)}}{dt} \mathbf{x}_2 = \sum_{i=1}^{P-1} \mathbf{F}_{2,i+1} \mathbf{x}_{i+1} + \sum_{i=0}^{P-2} \mathbf{G}_{2,i+1} \mathbf{x}_{i+1} u \quad (5.98)$$

com $\mathbf{F}_{2,i+1}, \mathbf{G}_{2,i+1} \in \mathbb{R}^{n^2 \times n^{i+1}}$, em que o símbolo

$$\frac{d^{(P)}}{dt} \quad (5.99)$$

denota a expansão em série de Taylor da derivada em questão truncada na ordem P , a qual resulta após o descarte de termos de ordem maior que P . Analogamente, com

$$\mathbf{F}_{3,i} = \mathbf{F}_{i-2} \otimes \mathbf{I}_n \otimes \mathbf{I}_n + \mathbf{I}_n \otimes \mathbf{F}_{i-2} \otimes \mathbf{I}_n + \mathbf{I}_n \otimes \mathbf{I}_n \otimes \mathbf{F}_{i-2}, \quad i \in \{3, \dots, P\} \quad (5.100)$$

$$\mathbf{G}_{3,i} = \mathbf{G}_{i-2} \otimes \mathbf{I}_n \otimes \mathbf{I}_n + \mathbf{I}_n \otimes \mathbf{G}_{i-2} \otimes \mathbf{I}_n + \mathbf{I}_n \otimes \mathbf{I}_n \otimes \mathbf{G}_{i-2}, \quad i \in \{2, \dots, P-1\} \quad (5.101)$$

pode-se formular a equação diferencial para $\mathbf{x}_3$

$$\frac{d}{dt} {}^{(P)} \mathbf{x}_3 = \sum_{i=1}^{P-2} \mathbf{F}_{3,i+2} \mathbf{x}_{i+2} + \sum_{i=0}^{P-3} \mathbf{G}_{3,i+2} \mathbf{x}_{i+2} u \quad (5.102)$$

em que $\mathbf{F}_{3,i+2}, \mathbf{G}_{3,i+2} \in \mathbb{R}^{n^3 \times n^{i+2}}$, e assim por diante. Procedendo-se desta maneira, obtêm-se $P-1$ equações na forma

$$\frac{d}{dt} {}^{(P)} \mathbf{x}_p = \sum_{i=1}^{P-p+1} \mathbf{F}_{p,i+p-1} \mathbf{x}_{i+p-1} + \sum_{i=0}^{P-p} \mathbf{G}_{p,i+p-1} \mathbf{x}_{i+p-1} u \quad (5.103)$$

com $p \in \{2, \dots, P\}$, onde $\mathbf{F}_{p,i+p-1}, \mathbf{G}_{p,i+p-1} \in \mathbb{R}^{n^p \times n^{i+p-1}}$ são dadas por

$$\begin{aligned} \mathbf{F}_{p,i} = & \mathbf{F}_{i-p+1} \otimes \mathbf{I}_n \otimes \dots \otimes \mathbf{I}_n + \mathbf{I}_n \otimes \mathbf{F}_{i-p+1} \otimes \mathbf{I}_n \otimes \dots \otimes \mathbf{I}_n \\ & + \dots + \mathbf{I}_n \otimes \mathbf{I}_n \otimes \dots \otimes \mathbf{I}_n \otimes \mathbf{F}_{i-p+1} \end{aligned} \quad (5.104)$$

$$\begin{aligned} \mathbf{G}_{p,i} = & \mathbf{G}_{i-p+1} \otimes \mathbf{I}_n \otimes \dots \otimes \mathbf{I}_n + \mathbf{I}_n \otimes \mathbf{G}_{i-p+1} \otimes \mathbf{I}_n \otimes \dots \otimes \mathbf{I}_n \\ & + \dots + \mathbf{I}_n \otimes \mathbf{I}_n \otimes \dots \otimes \mathbf{I}_n \otimes \mathbf{G}_{i-p+1} \end{aligned} \quad (5.105)$$

sendo que ambas as expressões (5.104) e (5.105) consistem em um somatório de p termos e cada termo envolve $p-1$ produtos de Kronecker.

Com isso, denotando-se

$$\mathbf{x}^\otimes = [\mathbf{x}^T \quad \mathbf{x}_2^T \quad \mathbf{x}_3^T \quad \dots \quad \mathbf{x}_P^T]^T \in \mathbb{R}^{n+n^2+\dots+n^P} \quad (5.106)$$

e notando que o sistema dado por (5.91) é tal que

$$\dot{\mathbf{x}} = \frac{d^{(P)}}{dt} \mathbf{x} \quad (5.107)$$

as equações diferenciais derivadas podem ser agrupadas na forma

$$\begin{aligned} \frac{d^{(P)}}{dt} \mathbf{x}^{\otimes} = & \begin{bmatrix} \mathbf{F}_{1,1} & \mathbf{F}_{1,2} & \mathbf{F}_{1,3} & \dots & \mathbf{F}_{1,P} \\ \mathbf{0} & \mathbf{F}_{2,2} & \mathbf{F}_{2,3} & \dots & \mathbf{F}_{2,P} \\ \mathbf{0} & \mathbf{0} & \mathbf{F}_{3,3} & \dots & \mathbf{F}_{3,P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{F}_{P,P} \end{bmatrix} \mathbf{x}^{\otimes} + \begin{bmatrix} \mathbf{G}_{1,1} & \mathbf{G}_{1,2} & \dots & \mathbf{G}_{1,P-1} & \mathbf{0} \\ \mathbf{G}_{2,1} & \mathbf{G}_{2,2} & \dots & \mathbf{G}_{2,P-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{3,2} & \dots & \mathbf{G}_{3,P-1} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{G}_{P,P-1} & \mathbf{0} \end{bmatrix} \mathbf{x}^{\otimes} u \\ & + \begin{bmatrix} \mathbf{G}_{1,0} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} u \quad (5.108) \end{aligned}$$

onde $\mathbf{F}_{1,i} = \mathbf{F}_i$ e $\mathbf{G}_{1,i} = \mathbf{G}_i$. De maneira mais compacta, (5.108) pode ser escrita como

$$\frac{d^{(P)}}{dt} \mathbf{x}^{\otimes} = \mathbf{F} \mathbf{x}^{\otimes} + \mathbf{G} \mathbf{x}^{\otimes} u + \mathbf{b} u \quad (5.109)$$

em que $\mathbf{x}^{\otimes}, \mathbf{b} \in \mathbb{R}^{N_P}$ e $\mathbf{F}, \mathbf{G} \in \mathbb{R}^{N_P \times N_P}$, sendo $N_P = n + n^2 + \dots + n^P$. Observe que a coluna de blocos nulos da matriz $\mathbf{G}$ se deve ao fato de que, como estamos considerando uma expansão de Taylor truncada na ordem P , os termos dados pelos componentes do produto $\mathbf{x}_P u$ não fazem parte da série, pois possuem ordem $P+1$. Considerando-se a equação diferencial (5.109) determinada para o sistema, sua saída é então dada por

$$y(t) = \mathbf{c}_P^T \mathbf{x}^{\otimes} \quad (5.110)$$

em que

$$\mathbf{c}_P = [\mathbf{c}^T \ 0 \ 0 \ \dots \ 0]^T \in \mathbb{R}^{N_P}. \quad (5.111)$$

Assim, como as equações (5.109) e (5.110) descrevem um sistema bilinear, então pode-se lançar mão das fórmulas dadas nas Subseções 3.1.2 e 5.2.2 para o

cálculo de seus núcleos ou de suas FRFGs, respectivamente.

5.3.1 Implementação computacional

Como visto na subseção anterior, a chave para se bilinearizar um sistema na forma considerada é a construção das matrizes $\mathbf{F}$ e $\mathbf{G}$ e dos vetores $\mathbf{b}$ e $\mathbf{c}_P$ a partir da expansão em série de potências das equações de estado. Além disso, vê-se também que as expressões envolvidas na construção de $\mathbf{F}$ e $\mathbf{G}$ podem ser escritas nas formas gerais (5.103)–(5.105), enquanto os vetores $\mathbf{b}$ e $\mathbf{c}_P$ são facilmente obtidos a partir do sistema linear-analítico. Por conseguinte, ao contrário do método da exponencial crescente, o aumento da ordem máxima considerada, P , não incorre na necessidade de se considerarem expressões de grande complexidade. Esta simplicidade constitui uma vantagem do método de Carleman, fazendo com que ele seja conveniente para implementação computacional.

É importante observar que, após a bilinearização, há ainda um passo adicional necessário: o cálculo das FRFGs do sistema bilinear construído. Contudo, também não há dificuldade envolvida neste passo, o qual se pode cumprir aplicando-se um procedimento recursivo para o cômputo de pontos das FRFGs dadas pelas expressões (5.88)–(5.90), as quais são trivialmente formuláveis para qualquer ordem p .

A seguir, discutiremos brevemente certas ineficiências envolvidas na implementação direta do método de Carleman, i.e., naquela em que se utilizam as expressões envolvidas na apresentação dada. Em seguida, descreveremos uma estratégia proposta pelos autores visando uma implementação mais eficiente [33] e forneceremos o algoritmo correspondente.

5.3.1.1 Implementação direta

Um possível procedimento para se construir as estruturas de (5.109), o qual decorre diretamente do desenvolvimento exposto para o método, consiste em obter as expansões de Taylor das funções $\mathbf{f}$ e $\mathbf{g}$ truncadas na ordem P e, após isso, utilizar as expressões (5.104) e (5.105). Note que quando tal estratégia é adotada,

considera-se que o vetor de estados é como definido por (5.106).

Um problema envolvido nesta abordagem direta é que, para um sistema linear-analítico em que o vetor de estados $\mathbf{x}$ possui dimensão $n > 1$, ela é ineficiente, visto que os vetores $\mathbf{x}_p$ contêm, por definição, componentes redundantes para $p > 1$. Por exemplo, considere o vetor $\mathbf{x}_2 \in \mathbb{R}^9$ dado pela definição (5.62) no caso em que $\mathbf{x} \in \mathbb{R}^3$

$$\mathbf{x}_2 = \mathbf{x} \otimes \mathbf{x} \quad (5.112)$$

$$= [x_1^2 \ x_1x_2 \ x_1x_3 \ x_2x_1 \ x_2^2 \ x_2x_3 \ x_3x_1 \ x_3x_2 \ x_3^2]^T \quad (5.113)$$

onde x_i denota o i -ésimo componente de $\mathbf{x}$, $i \in \{1, \dots, n\}$, tendo sido omitido o argumento relativo ao tempo por simplicidade de notação. Devido à equivalência de produtos $x_i x_j$ e $x_j x_i$, vê-se claramente que há componentes repetidos em $\mathbf{x}_2$. Por consequência disto, haverá também redundâncias no vetor de estados, $\mathbf{x}^\otimes$, e nas matrizes $\mathbf{F}_p$ e $\mathbf{G}_p$ tais que $p \geq 2$, fazendo com que o cômputo de cada FRFG de ordem $p > 1$ através da estratégia acima descrita envolva um número de operações maior que o necessário. Além disso, tal comportamento se agrava com o aumento de p , dado que o número de permutações de índices $i_1, \dots, i_p$ que resultam em produtos equivalentes a $x_{i_1} \dots x_{i_p}$ aumenta com p .

Deve-se notar que, embora em [32] a existência de redundâncias nas fórmulas do método de Carleman tenha sido comentada, não se mostra como é possível contorná-la na prática. A seguir, abordaremos esta questão.

5.3.1.2 Implementação eficiente

A implementação eficiente do método de Carleman requer que se eliminem as redundâncias do vetor de estados $\mathbf{x}^\otimes$ mencionadas na subseção anterior. Para tanto, vamos introduzir a nova definição

$$\mathbf{x}^\otimes = [\mathbf{x} \ \bar{\mathbf{x}}_2 \ \bar{\mathbf{x}}_3 \ \dots \ \bar{\mathbf{x}}_P]^T \in \mathbb{R}^{\bar{N}_p} \quad (5.114)$$

onde $\bar{\mathbf{x}}_p \in \mathbb{R}^{\bar{n}_p}$ corresponde a uma versão não-redundante de $\mathbf{x}_p$ em que é mantida apenas a primeira ocorrência de cada componente, eliminando-se as ocorrências equivalentes posteriores, e $\bar{N}_p = n + \bar{n}_2 + \dots + \bar{n}_p$. Por exemplo, a versão não-redundante de (5.113) é

$$\bar{\mathbf{x}}_2 = [x_1^2 \ x_1x_2 \ x_1x_3 \ x_2^2 \ x_2x_3 \ x_3^2]^T. \quad (5.115)$$

Assim, neste caso, tem-se uma redução do número de componentes por um fator de $\frac{2}{3}$. Já no caso do vetor $\bar{\mathbf{x}}_3$, por exemplo, seria obtida uma redução do número de componentes por um fator de $\frac{10}{27}$ em relação a $\mathbf{x}_3$, ou seja, elimina-se mais da metade deles. Consistentemente com aquilo que foi discutido na subseção anterior, tal redução aumenta, como se vê, com a ordem a que se refere o vetor considerado.

Uma vez estabelecida a nova definição do vetor de estados $\mathbf{x}^\otimes$ não-redundante, é necessário agora que seja desenvolvido um procedimento que permita derivar versões não-redundantes correspondentes das matrizes $\mathbf{F}$ e $\mathbf{G}$, as quais compõem o sistema de equações bilinear, com dimensões $\bar{N}_p \times \bar{N}_p$. Para tanto, propõe-se o uso de uma estratégia que consiste em computar as i -ésimas linhas de $\mathbf{F}$ e $\mathbf{G}$ recursivamente a partir de linhas previamente determinadas, o que corresponde à derivação sequencial recursiva das equações que compõem o sistema bilinear.

A estratégia proposta baseia-se em duas propriedades do método. Em primeiro lugar, note que os elementos que constituem as i -ésimas linhas de $\mathbf{F}$ e $\mathbf{G}$ e $\mathbf{b}$ correspondem aos coeficientes das expansões de Taylor⁵ das funções f_i e g_i que descrevem a evolução temporal do i -ésimo componente do vetor de estados, $\mathbf{x}^\otimes$. Em outras palavras, tais elementos definem a equação

$$\frac{d}{dt}(x_{i_1} \dots x_{i_p}) = f_i(\mathbf{x}) + g_i(\mathbf{x})u \quad (5.116)$$

em que $x_{i_1} \dots x_{i_p}$ é o i -ésimo componente⁶ de $\mathbf{x}^\otimes$, possuindo ordem p . No contexto

⁵A expansão de Taylor de campos vetoriais é descrita na Subseção A.2.2 do Apêndice A.

⁶Mais adiante, uma fórmula que possibilita determinar i a partir de $x_{i_1} \dots x_{i_p}$ será dada, para fins de implementação. Por ora, contudo, tal relação é irrelevante para a exposição da estratégia eficiente.

do método de Carleman, considera-se a expansão em série de Taylor de (5.116) truncada na ordem P , a qual pode ser escrita como

$$\frac{d^{(P)}}{dt}(x_{i_1} \dots x_{i_p}) = \mathbf{f}_i \mathbf{x}^\otimes + \mathbf{g}_i \mathbf{x}^\otimes u + b_i u \quad (5.117)$$

em que $\mathbf{f}_i \in \mathbb{R}^{1 \times \overline{N}_p}$ é a i -ésima linha de $\mathbf{F}$, $\mathbf{g}_i \in \mathbb{R}^{1 \times \overline{N}_p}$ é a i -ésima linha de $\mathbf{G}$ e b_i é o i -ésimo componente de $\mathbf{b} \in \mathbb{R}^{\overline{N}_p}$. Consequentemente, as n primeiras linhas das matrizes $\mathbf{F}$ e $\mathbf{G}$ e do vetor $\mathbf{b}$ podem ser completamente determinadas pela expansão em série de Taylor das funções originais $\mathbf{f}$ e $\mathbf{g}$.

Segundo, cada componente $x_{i_1} \dots x_{i_p}$ de $\mathbf{x}^\otimes$ com ordem $p \geq 2$ pode ser escrito como um produto de componentes anteriores. Logo, se o procedimento de cômputo for feito sequencialmente, as equações diferenciais associadas a tais componentes anteriores já foram determinadas no momento em que se deriva a equação associada a $x_{i_1} \dots x_{i_p}$. Isso possibilita fazê-lo recursivamente através da regra do produto

$$\frac{d}{dt}(x_{i_1} \dots x_{i_p}) = x_{i_1} \dots x_{i_q} \frac{d}{dt}(x_{i_{q+1}} \dots x_{i_p}) + x_{i_{q+1}} \dots x_{i_p} \frac{d}{dt}(x_{i_1} \dots x_{i_q}) \quad (5.118)$$

em que $q = \lfloor p/2 \rfloor$ e $i_1, \dots, i_p \in \{1, \dots, n\}$. Usando a equação (5.117), pode-se escrever a regra análoga

$$\begin{aligned} \frac{d^{(P)}}{dt}(x_{i_1} \dots x_{i_p}) &= x_{i_1} \dots x_{i_q} (\mathbf{f}_j \mathbf{x}^\otimes + \mathbf{g}_j \mathbf{x}^\otimes u + b_j u) \\ &\quad + x_{i_{q+1}} \dots x_{i_p} (\mathbf{f}_k \mathbf{x}^\otimes + \mathbf{g}_k \mathbf{x}^\otimes u + b_k u) \end{aligned} \quad (5.119)$$

$$= \mathbf{f}_m \mathbf{x}^\otimes + \mathbf{g}_m \mathbf{x}^\otimes u \quad (5.120)$$

em que se admite que $x_{i_1} \dots x_{i_q}$, $x_{i_{q+1}} \dots x_{i_p}$ e $x_{i_1} \dots x_{i_p}$ são o k -ésimo, o j -ésimo e o m -ésimo componentes de $\mathbf{x}^\otimes$, respectivamente.

Observe que a ideia chave da implementação eficiente é que a eliminação das redundâncias pode ser facilmente obtida através do uso da regra (5.118) ao invés de (5.92). Isto é, considerando-se as equações diferenciais individualmente, é possível desprezar aquelas que não acrescentam informações às estruturas resultantes.

Para simplificar a notação do que segue, vamos definir o operador

$$\odot : I \times \mathbb{R}^{1 \times \overline{N}_P} \rightarrow \mathbb{R}^{1 \times \overline{N}_P} \quad (5.121)$$

em que I é o conjunto de multi-índices $(i_1, \dots, i_q)$ com comprimento $q \leq P$ e tais que seus componentes satisfazem $i_j \in \{1, \dots, n\}$, de maneira que o resultado de

$$\mathbf{f}_m = (i_1, \dots, i_q) \odot \mathbf{f}_k \quad (5.122)$$

é tal que, se $\mathbf{f}_k \mathbf{x}^\otimes$ fornece a expansão polinomial

$$\mathbf{f}_k \mathbf{x}^\otimes = \sum_{i=1}^P \sum_{j_1=1}^n \sum_{j_2=j_1}^n \dots \sum_{j_i=j_{i-1}}^n \alpha(j_1, \dots, j_i) x_{j_1} \dots x_{j_i} \quad (5.123)$$

então $\mathbf{f}_m \mathbf{x}^\otimes$ é igual a

$$\mathbf{f}_m \mathbf{x}^\otimes = x_{i_1} \dots x_{i_q} \sum_{i=1}^{P-q} \sum_{j_1=1}^n \sum_{j_2=j_1}^n \dots \sum_{j_i=j_{i-1}}^n \alpha(j_1, \dots, j_i) x_{j_1} \dots x_{j_i}. \quad (5.124)$$

Analogamente,

$$\mathbf{g}_m = (i_1, \dots, i_q) \odot \mathbf{g}_k \quad (5.125)$$

é tal que, se $\mathbf{g}_k \mathbf{x}^\otimes$ fornece a expansão polinomial

$$\mathbf{g}_k \mathbf{x}^\otimes = u \sum_{i=1}^{P-1} \sum_{j_1=1}^n \sum_{j_2=j_1}^n \dots \sum_{j_i=j_{i-1}}^n \beta(j_1, \dots, j_i) x_{j_1} \dots x_{j_i} \quad (5.126)$$

então $\mathbf{g}_m \mathbf{x}^\otimes$ é igual a

$$\mathbf{g}_m \mathbf{x}^\otimes = x_{i_1} \dots x_{i_q} u \sum_{i=1}^{P-1-q} \sum_{j_1=1}^n \sum_{j_2=j_1}^n \dots \sum_{j_i=j_{i-1}}^n \beta(j_1, \dots, j_i) x_{j_1} \dots x_{j_i}. \quad (5.127)$$

Em outras palavras, $\mathbf{f}_m$ e $\mathbf{g}_m$ fornecem os vetores que representam expansões de Taylor de $(x_{i_1} \dots x_{i_q}) f_k(\mathbf{x})$, truncada na ordem P , e de $(x_{i_1} \dots x_{i_q}) g_k(\mathbf{x})$, truncada na ordem $P - 1$, respectivamente.

Com esta definição, podemos formalizar a derivação das linhas de $\mathbf{F}$ e $\mathbf{G}$ pela

regra descrita em (5.119)–(5.120) da seguinte maneira. Denotando

$$\begin{aligned} \mathbf{f}_{m,1} &= (i_1, \dots, i_q) \odot \mathbf{f}_j, & \mathbf{g}_{m,1} &= (i_1, \dots, i_q) \odot \mathbf{g}_j + b_j \mathbf{e}_k^T \\ \mathbf{f}_{m,2} &= (i_{q+1}, \dots, i_p) \odot \mathbf{f}_k, & \mathbf{g}_{m,2} &= (i_{q+1}, \dots, i_p) \odot \mathbf{g}_k + b_k \mathbf{e}_j^T \end{aligned} \quad (5.128)$$

em que $\mathbf{e}_j$ e $\mathbf{e}_k$ são o j -ésimo e o k -ésimo vetores da base canônica do $\mathbb{R}^{\overline{N}_P}$, podemos escrever

$$\mathbf{f}_m = \mathbf{f}_{m,1} + \mathbf{f}_{m,2} \quad (5.129)$$

$$\mathbf{g}_m = \mathbf{g}_{m,1} + \mathbf{g}_{m,2}. \quad (5.130)$$

Observe que a presença do termo $b_j \mathbf{e}_k^T$ em (5.128) se deve ao fato de que, no caso em que $j \leq n$, pode haver um termo $b_j u$ não-nulo compondo a j -ésima equação. Logo, ao se derivar o vetor $\mathbf{g}_{m,1}$, o coeficiente de tal termo deve ser levado à posição que se refere ao componente $x_{i_1} \dots x_{i_q}$ de $\mathbf{x}^\otimes$, que é justamente a k -ésima. Uma observação análoga é válida em relação ao termo $b_k \mathbf{e}_j^T$, o qual, por sua vez, compõe $\mathbf{g}_{m,2}$. Ainda, note que em (5.120) não foi incluído um termo na forma $b_m u$ porque, como estamos admitindo que o m -ésimo componente possui ordem $p \geq 2$, então não haverá um termo de primeira ordem envolvendo apenas u , como se pode ver na forma do sistema bilinear dada na equação (5.108).

Em suma, com base no conteúdo acima descrito, vê-se que a derivação recursiva das linhas de $\mathbf{F}$ e $\mathbf{G}$, as quais são representadas pelos vetores $\mathbf{f}_m$ e $\mathbf{g}_m$, $m \in \{1, \dots, \overline{N}_P\}$, pode ser realizada explorando-se os fatos enumerados abaixo.

- (i) Cada termo na forma $(i_1, \dots, i_q) \odot \mathbf{f}_j$ pode ser determinado simplesmente construindo-se o vetor resultante de forma que os componentes de $\mathbf{f}_j$ sejam deslocados de posição em função da multiplicação pelas variáveis $x_{i_1} \dots x_{i_q}$, considerando-se uma ordenação dos coeficientes da expansão de Taylor análoga à dos componentes de $\mathbf{x}^\otimes$. Note que os coeficientes são sempre deslocados para posições mais à direita de suas posições originais no vetor $\mathbf{f}_j$, uma vez que eles passam a multiplicar termos de ordem maior da expansão.
- (ii) O resultado de cada produto na forma $\mathbf{g}_{m,1} = (i_1, \dots, i_q) \odot \mathbf{g}_j$ é determi-

nado da mesma forma descrita no item (i). Contudo, neste caso, deve-se ter em mente que os componentes ao final do vetor resultante que representam coeficientes para os termos de ordem P da expansão de g_m são necessariamente nulos, como se pode ver em (5.108). Isto deve ser levado em conta na determinação dos vetores $\mathbf{g}_{m,1}$ e $\mathbf{g}_{m,2}$, os quais são somados para se obter $\mathbf{g}_m$.

- (iii) O resultado de cada soma nas formas $\mathbf{g}_{m,1} = (i_1, \dots, i_q) \odot \mathbf{g}_j + b_j \mathbf{e}_k^T$, $\mathbf{f}_m = \mathbf{f}_{m,1} + \mathbf{f}_{m,2}$ e $\mathbf{g}_m = \mathbf{g}_{m,1} + \mathbf{g}_{m,2}$ é obtido através da adição usual dos vetores envolvidos, uma vez que se admite que os coeficientes das expansões de Taylor contidos em ambos estão ordenados da mesma forma.

Com o intuito de clarificar os passos envolvidos na estratégia apresentada, um exemplo concreto é dado a seguir.

Exemplo 5.1

Considere o sistema linear-analítico dado por

$$\dot{x}_1 = \sin(x_1) - \left(x_1^2 + e^{x_2} - \frac{1}{2}\right)u \quad (5.131)$$

$$\dot{x}_2 = -x_1^2 - x_1x_2^2 + x_1x_2u \quad (5.132)$$

com $\mathbf{x}_0 = [0 \ 0]^T$. A partir das equações dadas, vê-se que

$$\begin{aligned} f_1(\mathbf{x}) &= \sin(x_1), & g_1(\mathbf{x}) &= -x_1^2 - e^{x_2} + \frac{1}{2} \\ f_2(\mathbf{x}) &= -x_1^2 - x_1x_2^2, & g_2(\mathbf{x}) &= x_1x_2. \end{aligned} \quad (5.133)$$

No que segue, vamos derivar a forma bilinear de Carleman de ordem $P = 3$ a partir do sistema (5.131)–(5.132). Inicialmente, tem-se que o vetor de estados não-redundante é dado por

$$\mathbf{x}^\otimes = [x_1 \ x_2 \ x_1^2 \ x_1x_2 \ x_2^2 \ x_1^3 \ x_1^2x_2 \ x_1x_2^2 \ x_2^3]^T. \quad (5.134)$$

Agora, vamos calcular as expansões de Taylor em torno de $\mathbf{x}_0$ truncadas na ordem P das equações de estado original e considerar o sistema aproxi-

mado cujas equações diferenciais são definidas por tais expansões. Nessas novas equações, os componentes da função $\mathbf{f}$ são dados por

$$f_1(\mathbf{x}) = x_1 - \frac{x_1^3}{6} \quad (5.135)$$

$$f_2(\mathbf{x}) = -x_1^2 - x_1x_2^2. \quad (5.136)$$

Por sua vez, as expansões de g_1 e g_2 que compõem a função $\mathbf{g}$ do sistema aproximado são dadas por

$$g_1(\mathbf{x}) = -\frac{1}{2} - x_2 - x_1^2 - \frac{x_2^2}{2} \quad (5.137)$$

$$g_2(\mathbf{x}) = x_1x_2 \quad (5.138)$$

em que se retêm apenas os termos de ordem menor ou igual a dois uma vez que tais termos, ao serem multiplicados pela entrada, possuirão ordem menor ou igual a três. Assim, as duas primeiras equações do sistema bilinear correspondente são dadas por

$$\frac{d}{dt} x_1 = \mathbf{f}_1 \mathbf{x}^\otimes + \mathbf{g}_1 \mathbf{x}^\otimes u + b_1 u \quad (5.139)$$

$$\frac{d}{dt} x_2 = \mathbf{f}_2 \mathbf{x}^\otimes + \mathbf{g}_2 \mathbf{x}^\otimes u + b_2 u \quad (5.140)$$

em que

$$\begin{aligned} \mathbf{f}_1 &= [1 \quad 0 \quad 0 \quad 0 \quad 0 \quad -1/6 \quad 0 \quad 0 \quad 0] \\ \mathbf{g}_1 &= [0 \quad -1 \quad -1 \quad 0 \quad -1/2 \quad 0 \quad 0 \quad 0 \quad 0] \\ \mathbf{f}_2 &= [0 \quad 0 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad -1 \quad 0] \\ \mathbf{g}_2 &= [0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] \end{aligned} \quad (5.141)$$

$b_1 = -1/2$ e $b_2 = 0$.

Estabelecidas as n primeiras equações, pode-se agora empregar o método recursivo para derivação das demais. A equação diferencial truncada na

ordem P para o terceiro componente de $\mathbf{x}^\otimes$ é dada por

$$\frac{d^{(P)}}{dt}(x_1^2) = 2x_1(\mathbf{f}_1\mathbf{x}^\otimes + \mathbf{g}_1\mathbf{x}^\otimes u + b_1u) \quad (5.142)$$

$$= \mathbf{f}_3\mathbf{x}^\otimes + \mathbf{g}_3\mathbf{x}^\otimes u. \quad (5.143)$$

Nesse caso, temos que os vetores parciais $\mathbf{f}_{3,1}$ e $\mathbf{f}_{3,2}$ que compõem a soma $\mathbf{f}_3 = \mathbf{f}_{3,1} + \mathbf{f}_{3,2}$ são ambos iguais a

$$\mathbf{f}_{3,1} = \mathbf{f}_{3,2} = (1) \odot \mathbf{f}_1 = [0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0] \quad (5.144)$$

e podem ser determinados através do deslocamento do primeiro componente de $\mathbf{f}_1$ para a posição relativa ao termo x_1^2 . Note que, como o coeficiente de valor $-1/6$ já se referia a um termo de ordem $P = 3$, então ele não está mais presente nos vetores $\mathbf{f}_{3,1}$ e $\mathbf{f}_{3,2}$ uma vez que ele corresponderia ao termo x_1^4 , porém os vetores contêm apenas coeficientes que multiplicam termos de ordem menor ou igual a três. Assim

$$\mathbf{f}_3 = [0 \ 0 \ 2 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]. \quad (5.145)$$

Os vetores $\mathbf{g}_{3,1}$ e $\mathbf{g}_{3,2}$ são, assim como $\mathbf{f}_{3,1}$ e $\mathbf{f}_{3,2}$, iguais, sendo dados por

$$\mathbf{g}_{3,1} = \mathbf{g}_{3,2} = (1) \odot \mathbf{g}_1 + b_1\mathbf{e}_1 = [-1/2 \ 0 \ 0 \ -1 \ 0 \ 0 \ 0 \ 0 \ 0] \quad (5.146)$$

de forma que

$$\mathbf{g}_3 = \mathbf{g}_{3,1} + \mathbf{g}_{3,2} = [-1 \ 0 \ 0 \ -2 \ 0 \ 0 \ 0 \ 0 \ 0]. \quad (5.147)$$

Para a equação diferencial relativa ao quarto componente de $\mathbf{x}^\otimes$, dada por

$$\frac{d^{(P)}}{dt}(x_1x_2) = x_1(\mathbf{f}_2\mathbf{x}^\otimes + \mathbf{g}_2\mathbf{x}^\otimes u) + x_2(\mathbf{f}_1\mathbf{x}^\otimes + \mathbf{g}_1\mathbf{x}^\otimes u + b_1u) \quad (5.148)$$

$$= \mathbf{f}_4\mathbf{x}^\otimes + \mathbf{g}_4\mathbf{x}^\otimes u \quad (5.149)$$

e

$$\mathbf{G} = \begin{bmatrix} 0 & -1 & -1 & 0 & -1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1/2 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -3/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \quad (5.157)$$

Já o vetor $\mathbf{b}$ resultante é dado por

$$\mathbf{b} = [-1/2 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]^T. \quad (5.158)$$

□

Para fins de implementação, é importante que se tenha uma fórmula para determinar a posição de um componente qualquer $x_{i_1} \dots x_{i_p}$ no vetor de estados $\mathbf{x}^\otimes$. Tal fórmula é útil na implementação de (5.119)–(5.120), tanto para se recuperar os vetores $\mathbf{f}_j$, $\mathbf{g}_j$, $\mathbf{f}_k$ e $\mathbf{g}_k$ previamente calculados, com base nos componentes a que eles se referem, quanto para se implementar o operador $\odot$ de modo que os coeficientes dos vetores resultantes sejam deslocados para as posições corretas.

Uma relação entre o componente $x_{i_1} \dots x_{i_p}$ e sua posição i no vetor de estados pode ser formulada recursivamente como descrito a seguir. Primeiro, consideramos a representação do componente de interesse na forma $x_1^{r_1} \dots x_n^{r_n}$, em que $r_1 + \dots + r_n = p$. Então, podemos escrever a fórmula que fornece i como

$$\Upsilon(r_1, \dots, r_n, n) = \begin{cases} 0, & r_1 = \dots = r_n = 0 \\ j, & r_1 = \dots = r_{j-1} = r_{j+1} = \dots = r_n = 0 \text{ e } r_j = 1 \\ L(r_1, \dots, r_n, n) & \text{caso contrário} \end{cases} \quad (5.159)$$

em que

$$L(r_1, \dots, r_n, n) = \frac{(\sum_{j=1}^n r_j - 1 + n)!}{(\sum_{j=1}^n r_j - 1)!n!} + \Upsilon(r_2, \dots, r_n, n - 1). \quad (5.160)$$

Embora possa parecer elaborada, a ideia em que se baseia a fórmula (5.159) é simples. O primeiro caso, que resulta em zero, é definido apenas para fornecer um caso base para a recursão. O segundo se aplica a todo componente na forma x_j , com $j \in \{1, \dots, n\}$, cuja posição em $\mathbf{x}^\otimes$ é trivialmente dada por j , fornecendo também um caso base para a recursão.

Para a formulação do terceiro caso, usou-se o fato de que, se o componente de interesse é $x_{i_1} \dots x_{i_p}$, possuindo ordem p , então ele se situa *após* todos os componentes de ordens 1 a $p-1$. Observe que o número total de tais componentes é dado pelo primeiro termo de $L(\cdot)$, à semelhança da expressão que fornece o número de coeficientes de um filtro de Volterra de ordem $p-1$ com extensão de memória n para todos os núcleos. Além disso, ele se situa exatamente $\Upsilon(r_2, \dots, r_n, n-1)$ posições após os componentes de ordens de ordens 1 a $p-1$, uma vez que os elementos de $\mathbf{x}^\otimes$ que começam com um fator $x_1^{r_1}$ são ordenados da mesma maneira que os elementos de ordem $p-r_1$ que compõem o vetor de estados $\tilde{\mathbf{x}}^\otimes$ construído a partir do vetor original $\mathbf{x} = [x_2 \ x_3 \ \dots \ x_n]^T$. Logo, aplica-se uma fórmula recursiva.

5.3.1.3 Algoritmo para a implementação eficiente

O Algoritmo 1 contém uma representação em pseudo-código do procedimento eficiente proposto. As entradas para o algoritmo são:

- (i) a ordem máxima considerada para a bilinearização, P ;
- (ii) vetores $\mathbf{f}_1, \dots, \mathbf{f}_n \in \mathbb{R}^{1 \times \overline{N}_P}$ contendo os coeficientes dos termos de ordens 1 a P das expansões em série de Taylor das funções $f_1, \dots, f_n$;
- (iii) analogamente, vetores $\mathbf{g}_1, \dots, \mathbf{g}_n \in \mathbb{R}^{1 \times \overline{N}_P}$ contendo os coeficientes dos termos de ordens 1 a P das expansões em série de Taylor das funções $g_1, \dots, g_n$, sendo todos os termos de ordem P necessariamente nulos, em

face do que foi discutido na subseção anterior;

- (iv) o vetor $\mathbf{G}_{1,0} \in \mathbb{R}^n$, o qual contém os termos de ordem zero das expansões de $g_1, \dots, g_n$.

As saídas do algoritmo são as estruturas que compõem a equação bilinear (5.109), porém sem redundâncias, ou seja, tais que $\mathbf{F}, \mathbf{G} \in \mathbb{R}^{\overline{N}_P \times \overline{N}_P}$ e $\mathbf{b} \in \mathbb{R}^{\overline{N}_P}$.

Algoritmo 1 Bilinearização de Carleman de ordem P

Entradas: $P \in \mathbb{N}^*$, $\mathbf{f}_1, \dots, \mathbf{f}_n, \mathbf{g}_1, \dots, \mathbf{g}_n \in \mathbb{R}^{1 \times \overline{N}_P}$, $\mathbf{G}_{1,0} \in \mathbb{R}^n$

Saídas: $\mathbf{F}, \mathbf{G} \in \mathbb{R}^{\overline{N}_P \times \overline{N}_P}$, $\mathbf{b} \in \mathbb{R}^{\overline{N}_P}$

```

1:  $\mathbf{b} \leftarrow [\mathbf{G}_{1,0}^T \ 0 \ \dots \ 0]^T \in \mathbb{R}^{\overline{N}_P}$ 
2: para  $k = n + 1, \dots, \overline{N}_P$  faça
3:   Seja  $x_{i_1} \dots x_{i_p}$  o  $k$ -ésimo componente de  $\mathbf{x}^\otimes$ 
4:    $q \leftarrow \lfloor p/2 \rfloor$ 
5:   Sejam  $x_{i_1} \dots x_{i_q} = x_1^{r_1} \dots x_n^{r_n}$  e  $x_{i_{q+1}} \dots x_{i_p} = x_1^{s_1} \dots x_n^{s_n}$ 
6:    $k_1 \leftarrow \Upsilon(r_1, \dots, r_n, n)$ 
7:    $k_2 \leftarrow \Upsilon(s_1, \dots, s_n, n)$ 
8:    $\mathbf{f}_k \leftarrow (i_1, \dots, i_q) \odot \mathbf{f}_{k_2} + (i_{q+1}, \dots, i_p) \odot \mathbf{f}_{k_1}$ 
9:    $\mathbf{g}_k \leftarrow (i_1, \dots, i_q) \odot \mathbf{g}_{k_2} + [\mathbf{b}]_{k_2} \mathbf{e}_{k_1} + (i_{q+1}, \dots, i_p) \odot \mathbf{g}_{k_1} + [\mathbf{b}]_{k_1} \mathbf{e}_{k_2}$ 
10: fim do para
11:  $\mathbf{F} \leftarrow \begin{bmatrix} \mathbf{f}_1^T & \mathbf{f}_2^T & \dots & \mathbf{f}_{\overline{N}_P}^T \end{bmatrix}^T$ 
12:  $\mathbf{G} \leftarrow \begin{bmatrix} \mathbf{g}_1^T & \mathbf{g}_2^T & \dots & \mathbf{g}_{\overline{N}_P}^T \end{bmatrix}^T$ 

```

No algoritmo, admite-se que a implementação do operador $\odot$ é feita por meio da função $\Upsilon(\cdot)$ definida na subseção anterior, assim como a busca dos vetores $\mathbf{f}_{k_1}$, $\mathbf{f}_{k_2}$, $\mathbf{g}_{k_1}$ e $\mathbf{g}_{k_2}$. Deve-se notar ainda que não é necessário elaborar uma função que forneça os índices $i_1, \dots, i_p$ do k -ésimo componente de $\mathbf{x}^\otimes$ considerado no laço, uma vez que se pode facilmente escrever um algoritmo que enumera tais componentes, i.e., que determina o k -ésimo com base no $(k - 1)$ -ésimo e no valor de n .

Para conveniência do leitor interessado, uma implementação em MATLAB do algoritmo eficiente para a bilinearização de Carleman, o qual é descrito acima, é fornecida em [87].

5.3.1.4 Comparação entre as complexidades do algoritmo eficiente e da implementação direta

Conforme discutido nas subseções anteriores, o aspecto chave para a redução da complexidade computacional obtida através da estratégia eficiente proposta é a eliminação de redundâncias no cômputo de $\mathbf{F}$ e $\mathbf{G}$. Para que se tenha uma ideia concreta da redução do número de elementos obtida com isso, na Tabela 5.1 são mostrados os números de elementos dos vetores de estado não-redundante $\overline{N}_P$, e redundante, N_P , para algumas combinações de valores de n e P . Os valores obtidos foram computados através das fórmulas

$$\overline{N}_P = \sum_{p=1}^P \frac{(n+p-1)!}{p!(n-1)!} = \frac{(n+P)!}{n!P!} - 1 \quad (5.161)$$

$$N_P = \sum_{p=1}^P n^p \quad (5.162)$$

onde a fórmula (5.161) é a da combinação com reposição. Deve-se notar que, como $\mathbf{F}, \mathbf{G} \in \mathbb{R}^{N_P \times N_P}$ no caso redundante e $\mathbf{F}, \mathbf{G} \in \mathbb{R}^{\overline{N}_P \times \overline{N}_P}$ no não-redundante, então o número total de elementos redundantes desconsiderados no cômputo eficiente é dado por $2(N_P^2 - \overline{N}_P^2)$. Assim, por exemplo, no caso em que $n = 3$ e $P = 4$, tal número é igual a 26.488 (o que corresponde a mais de 91% de $2N_P^2$); já no caso em que $n = 4$ e $P = 5$, tem-se um total de 3.594.990 elementos redundantes desconsiderados (o que representa mais de 96% de $2N_P^2$). Logo, visivelmente, com o aumento de n e P tem-se uma redução bastante significativa no tamanho das estruturas calculadas com o uso da estratégia eficiente.

	$n = 1$		$n = 2$		$n = 3$		$n = 4$		$n = 5$	
P	$\overline{N}_P$	N_P	$\overline{N}_P$	N_P	$\overline{N}_P$	N_P	$\overline{N}_P$	N_P	$\overline{N}_P$	N_P
1	1	1	2	2	3	3	4	4	5	5
2	2	2	5	6	9	12	14	20	20	30
3	3	3	9	14	19	39	34	84	55	155
4	4	4	14	30	34	120	69	340	125	780
5	5	5	20	62	55	363	125	1364	251	3905

Tabela 5.1: Comparação entre os números de elementos do vetor de estados não-redundante, $\overline{N}_P$, e do vetor de estados redundante, N_P , para diferentes valores de n e P .

Uma consequência direta da redução do número de elementos que compõem

as estruturas computadas é que, evidentemente, há um ganho em eficiência no que se refere ao uso de memória feito pelo algoritmo. Isto é, a quantidade de memória utilizada pela implementação eficiente cresce mais lentamente com n e P do que aquela que envolve redundâncias, o que pode constituir uma vantagem significativa para certos valores de n e P . Claramente, a economia resultante é proporcional ao total de elementos redundantes desconsiderados pela implementação eficiente.

Agora, consideremos uma comparação entre a implementação direta (i.e., com redundâncias) e a eficiente (i.e., sem redundâncias) do método de Carleman em termos do número de operações aritméticas realizadas. Na implementação direta, computam-se os blocos da matriz $\mathbf{F}$ com o uso da fórmula (5.104), dada por

$$\begin{aligned} \mathbf{F}_{p,i} = & \mathbf{F}_{i-p+1} \otimes \mathbf{I}_n \otimes \dots \otimes \mathbf{I}_n + \mathbf{I}_n \otimes \mathbf{F}_{i-p+1} \otimes \mathbf{I}_n \otimes \dots \otimes \mathbf{I}_n \\ & + \dots + \mathbf{I}_n \otimes \mathbf{I}_n \otimes \dots \otimes \mathbf{I}_n \otimes \mathbf{F}_{i-p+1}. \end{aligned} \quad (5.163)$$

Cada termo do somatório (5.163) envolve $p-1$ produtos de Kronecker, podendo ser escrito como

$$\mathbf{F}_{p,i} = \sum_{k=0}^{p-1} \mathbf{I}_{n,k} \otimes \mathbf{F}_{i-p+1} \otimes \mathbf{I}_{n,p-k-1} \quad (5.164)$$

sendo que

$$\mathbf{I}_{n,0} \otimes \mathbf{F}_{i-p+1} = \mathbf{F}_{i-p+1} \otimes \mathbf{I}_{n,0} = \mathbf{F}_{i-p+1} \quad (5.165)$$

e

$$\mathbf{I}_{n,1} = \mathbf{I}_n \quad (5.166)$$

$$\mathbf{I}_{n,k} = \mathbf{I}_n \otimes \mathbf{I}_{n,k-1}, \quad k > 1. \quad (5.167)$$

Observe que todas as matrizes $\mathbf{I}_{n,k}$ possuem apenas elementos iguais a zero ou um. Além disso, as matrizes $\mathbf{I}_{n,k}$ podem ser pré-computadas e armazenadas para uso. Assim, o resultado das operações do tipo $\mathbf{M} \otimes \mathbf{I}_{n,k}$ e $\mathbf{I}_{n,k} \otimes \mathbf{M}$, em que $\mathbf{M}$ é uma matriz qualquer, podem ser determinados eficientemente, sem o uso efetivo de multiplicações, apenas colocando-se “cópias” de $\mathbf{M}$ na matriz resultante de acordo com as posições dos elementos iguais a um de $\mathbf{I}_{n,k}$. Portanto, não

levaremos em consideração a formação dos termos de (5.164) na contagem das operações aritméticas realizadas pelo método.

No entanto, as somas dos elementos realizadas na expressão (5.164) são consideradas. Como $\mathbf{F}_{i-p+1}$ é uma matriz $n \times n^{i-p+1}$ e $\mathbf{I}_{n,k}$ é $n^k \times n^k$, então cada termo do somatório possui dimensão $n^p \times n^i$. Visto que a determinação de cada um de seus n^{p+i} elementos requer $p-1$ adições, então o número de adições envolvidas no cálculo de cada linha de blocos, i.e., $\mathbf{F}_{p,p}, \mathbf{F}_{p,p+1}, \dots, \mathbf{F}_{p,P}$, é

$$\sum_{i=p}^P (p-1)n^{p+i}. \quad (5.168)$$

Uma contagem de operações similar pode ser feita para o cômputo dos blocos $\mathbf{G}_{p,i}$, de maneira que para cada linha de blocos $\mathbf{G}_{p,p-1}, \mathbf{G}_{p,p}, \dots, \mathbf{G}_{p,P-1}$ tem-se um número de adições igual a

$$\sum_{i=p-1}^{P-1} (p-1)n^{p+i}. \quad (5.169)$$

Note que, embora cada parcela de tais somas de matrizes contenha diversos elementos nulos, sistematizá-la de forma que eles sejam desconsiderados não é uma tarefa simples. Logo, contabilizam-se as adições de todos os elementos das matrizes resultantes dos produtos de Kronecker. Assim, considerando-se todos os blocos não-nulos das matrizes $\mathbf{F}$ e $\mathbf{G}$, obtém-se o total de adições exigido pela implementação direta dado por

$$\Psi_d(n, P) = \sum_{p=2}^P \left[\sum_{i=p}^P (p-1)n^{p+i} + \sum_{i=p-1}^{P-1} (p-1)n^{p+i} \right] \quad (5.170)$$

$$= \sum_{p=2}^P (p-1) \left[n^{2p-1} + n^{P+p} + 2 \sum_{i=p}^{P-1} n^{p+i} \right]. \quad (5.171)$$

Quando o algoritmo eficiente é empregado, computam-se $\overline{N}_P - n$ linhas de $\mathbf{F}$ e $\mathbf{G}$, sendo $\overline{N}_P$ dado pela fórmula (5.161). Cada linha de $\mathbf{F}$ é calculada através de

$$\mathbf{f}_k = (i_1, \dots, i_q) \odot \mathbf{f}_{k_2} + (i_{q+1}, \dots, i_p) \odot \mathbf{f}_{k_1}. \quad (5.172)$$

Como discutido anteriormente, o resultado do operador $\odot$ pode ser determinado

eficientemente utilizando-se a função Υ para determinar as posições para as quais devem ser deslocados os componentes do segundo operando, de acordo com os índices envolvidos no primeiro. Em última análise, é possível ainda pré-computar uma estrutura de dados que forneça tais posições em função dos índices de seus respectivos componentes, usando-se a função Υ . Com tal estrutura armazenada previamente, o cômputo seria acelerado. Por conseguinte, de forma análoga ao que foi feito em relação ao cômputo da expressão (5.164), a qual é envolvida na implementação direta, levaremos em conta apenas as adições de (5.172). Neste caso, somam-se dois vetores resultantes de dimensão $\overline{N}_P$.

Para o cálculo das linhas de $\mathbf{G}$, emprega-se a fórmula $\mathbf{g}_k = \mathbf{g}_{k,1} + \mathbf{g}_{k,2}$, em que

$$\mathbf{g}_{k,1} = (i_1, \dots, i_q) \odot \mathbf{g}_{k_2} + [\mathbf{b}]_{k_2} \mathbf{e}_{k_1} \quad (5.173)$$

$$\mathbf{g}_{k,2} = (i_{q+1}, \dots, i_p) \odot \mathbf{g}_{k_1} + [\mathbf{b}]_{k_1} \mathbf{e}_{k_2}. \quad (5.174)$$

Semelhantemente ao que se mencionou no parágrafo anterior, a contagem de operações não considerará a aplicação do operador $\odot$. Adicionalmente, para se determinar o resultado dos produtos $[\mathbf{b}]_{k_2} \mathbf{e}_{k_1}$ e $[\mathbf{b}]_{k_1} \mathbf{e}_{k_2}$ não é necessário realizar multiplicações, bastando construir um vetor contendo $\overline{N}_P$ elementos em que apenas um é (possivelmente) não-nulo e dado pelo respectivo componente de $\mathbf{b}$ envolvido no produto. Ainda, as adições feitas em (5.173) e (5.174) envolvem apenas dois escalares cada, uma vez que ambos os vetores $[\mathbf{b}]_{k_2} \mathbf{e}_{k_1}$ e $[\mathbf{b}]_{k_1} \mathbf{e}_{k_2}$ podem possuir apenas um componente não-nulo. Além destas adições, restam apenas a adição de vetores de dimensão $\overline{N}_P$. Porém, como os últimos $\overline{n}_P$ componentes de cada linha de $\mathbf{G}$ sempre são nulos, em que

$$\overline{n}_P = \frac{(n + P - 1)!}{P!(n - 1)!} \quad (5.175)$$

então, na prática, pode-se considerar que os vetores envolvidos possuem dimensão $\overline{N}_P - \overline{n}_P$.

Com base nas operações contadas, o número total de operações da implementação eficiente é

$$\Psi_e(n, P) = (\overline{N}_P - n) (2\overline{N}_P - \overline{n}_P + 2). \quad (5.176)$$

Com o intuito de fornecer uma ideia sobre o que a expressão (5.176) significa em termos de redução do custo computacional, na Tabela 5.2 exibem-se resultados da razão $\Psi_{\text{direta}}(n, P)/\Psi_{\text{eficiente}}(n, P)$ para algumas combinações de valores de n e P . À exceção do caso em que $n = P = 2$, no qual a abordagem direta exige 24 operações enquanto a eficiente exige 27, o uso desta última de fato proporciona uma redução no número total de operações, a qual cresce rapidamente com P . No caso em que $n = P = 4$, por exemplo, a implementação direta requer 303.680 operações, enquanto a implementação proposta, apenas 6.825. Já para $n = P = 5$, o número de operações exigidas pela implementação direta é maior que 5×10^7 , enquanto a abordagem eficiente requer menos de 10^5 operações aritméticas.

	$n = 2$			$n = 3$			$n = 4$			$n = 5$		
P	Ψ_d	Ψ_e	Ψ_d/Ψ_e	Ψ_d	Ψ_e	Ψ_d/Ψ_e	Ψ_d	Ψ_e	Ψ_d/Ψ_e	Ψ_d	Ψ_e	Ψ_d/Ψ_e
2	24	27	0,89	108	84	1,29	320	200	1,60	750	405	1,85
3	264	112	2,36	2.376	480	4,95	11.840	1.500	7,89	42.000	3.850	10,91
4	1.896	300	6,32	35.424	1.705	20,78	303.680	6.825	44,50	1.654.500	21.840	75,76
5	11.304	648	17,44	449.496	4.732	94,99	6.713.920	23.716	283,10	56.592.000	92.988	608,59

Tabela 5.2: Números de operações aritméticas envolvidas na implementação direta, $\Psi_d(n, P)$, na implementação eficiente, $\Psi_e(n, P)$, e respectivas razões Ψ_d/Ψ_e , para alguns valores de n e P .

Para finalizar, duas observações são importantes. Primeiro, o número de adições efetivamente computadas pela abordagem eficiente é ainda menor na prática do que aquele considerado acima, o que se deve ao fato de que tanto em $\mathbf{F}$ quanto em $\mathbf{G}$ existem blocos nulos abaixo da diagonal principal, como se pode ver em (5.108). Sendo assim, na adição dos vetores pode-se levar tal fato em conta e somar apenas os componentes das colunas que não fazem parte de tais blocos. Visando simplificar a determinação do número de operações, contudo, não se considerou tal otimização na contagem feita acima.

Segundo, nesta contagem não se levaram em conta as operações necessárias para a determinação das n primeiras linhas de $\mathbf{F}$ e $\mathbf{G}$, i.e., de $\mathbf{f}_1, \dots, \mathbf{f}_n, \mathbf{g}_1, \dots, \mathbf{g}_n$, uma vez que tais vetores são admitidos como entradas para ambas as formas de implementação (direta e eficiente). Note que, como estes vetores contêm coeficientes das expansões de Taylor de equações de estado, eles podem ser obtidos computacionalmente se houver disponibilidade de um pacote de computação sim-

bólica.

5.3.2 Exatidão dos núcleos e das FRFGs resultantes

Uma propriedade muito importante do método de Carleman é que não há nenhuma aproximação envolvida na determinação dos núcleos cujas ordens são menores ou iguais à ordem máxima considerada na construção do sistema bilinear, P . Em outras palavras, tais núcleos são fornecidos pelo método de forma exata, como será estabelecido a seguir. Embora esta propriedade e a ideia básica que a explica sejam mencionadas em [32], não se fornece uma demonstração ou uma argumentação detalhada a respeito. Por este motivo, faremos uma demonstração deste fato na subseção presente.

Para simplificar a argumentação apresentada, consideraremos a formulação original do método de Carleman, i.e., aquela que envolve redundâncias. Porém, deve-se ter em mente que os resultados são válidos também para o cômputo eficiente, uma vez que as versões não-redundantes de tais estruturas compreendem as mesmas informações que as redundantes.

Lema 5.1 Seja $\mathbf{A}$ uma matriz $N \times N$ dada por

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \mathbf{A}_{1,3} & \dots & \mathbf{A}_{1,P} \\ \mathbf{0} & \mathbf{A}_{2,2} & \mathbf{A}_{2,3} & \dots & \mathbf{A}_{2,P} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{3,3} & \dots & \mathbf{A}_{3,P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{A}_{P,P} \end{bmatrix} \quad (5.177)$$

em que $\mathbf{A}_{i,j} \in \mathbb{R}^{p_i \times p_j}$, com $\sum_{i=1}^P p_i = N$. Então, $\mathbf{C} = \mathbf{A}^p = \mathbf{A} \dots \mathbf{A}$ também possui a forma

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{1,1} & \mathbf{C}_{1,2} & \mathbf{C}_{1,3} & \dots & \mathbf{C}_{1,P} \\ \mathbf{0} & \mathbf{C}_{2,2} & \mathbf{C}_{2,3} & \dots & \mathbf{C}_{2,P} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_{3,3} & \dots & \mathbf{C}_{3,P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{C}_{P,P} \end{bmatrix} \quad (5.178)$$

onde $\mathbf{C}_{i,j} \in \mathbb{R}^{p_i \times p_j}$. Além disso, cada bloco $\mathbf{C}_{i,j}$ pode ser escrito como um somatório de produtos de blocos $\mathbf{A}_{l,m}$ tais que $l \leq m \leq j$.

Demonstração. Inicialmente, deve-se notar que $\mathbf{A}$ pode ser escrita na forma

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \mathbf{A}_{1,3} & \dots & \mathbf{A}_{1,P} \\ \mathbf{A}_{2,1} & \mathbf{A}_{2,2} & \mathbf{A}_{2,3} & \dots & \mathbf{A}_{2,P} \\ \mathbf{A}_{3,1} & \mathbf{A}_{3,2} & \mathbf{A}_{3,3} & \dots & \mathbf{A}_{3,P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{P,1} & \mathbf{A}_{P,2} & \mathbf{A}_{P,3} & \dots & \mathbf{A}_{P,P} \end{bmatrix} \quad (5.179)$$

em que $j < i \implies \mathbf{A}_{i,j} = \mathbf{0} \in \mathbb{R}^{p_i \times p_j}$. Agora, procederemos à prova por indução. Primeiro, observe que $\mathbf{C} = \mathbf{A}^2$ é dada por

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{1,1} & \mathbf{C}_{1,2} & \mathbf{C}_{1,3} & \dots & \mathbf{C}_{1,P} \\ \mathbf{C}_{2,1} & \mathbf{C}_{2,2} & \mathbf{C}_{2,3} & \dots & \mathbf{C}_{2,P} \\ \mathbf{C}_{3,1} & \mathbf{C}_{3,2} & \mathbf{C}_{3,3} & \dots & \mathbf{C}_{3,P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{P,1} & \mathbf{C}_{P,2} & \mathbf{C}_{P,3} & \dots & \mathbf{C}_{P,P} \end{bmatrix} \quad (5.180)$$

em que

$$\mathbf{C}_{i,j} = \sum_{k=1}^P \mathbf{A}_{i,k} \mathbf{A}_{k,j} \in \mathbb{R}^{p_i \times p_j}. \quad (5.181)$$

Além disso, como $\mathbf{A}_{i,j}$ é não-nulo somente se $i \leq j$, então

$$\mathbf{C}_{i,j} = \begin{cases} \sum_{k=i}^j \mathbf{A}_{i,k} \mathbf{A}_{k,j}, & i \leq j \\ \mathbf{0} \in \mathbb{R}^{p_i \times p_j}, & \text{caso contrário} \end{cases}. \quad (5.182)$$

Com base em (5.182), vê-se claramente que a proposição do lema é válida para $\mathbf{C} = \mathbf{A}^2$.

Em seguida, vamos admitir que ela se aplica também a qualquer matriz $\mathbf{A}^{p'}$ em que $p' < p$, com $p \geq 3$. Em particular, ela é válida para $\mathbf{D} = \mathbf{A}^{p-1}$. Então,

temos que

$$\mathbf{C} = \mathbf{A}^p = \mathbf{A}\mathbf{D} = \begin{bmatrix} \mathbf{C}_{1,1} & \mathbf{C}_{1,2} & \mathbf{C}_{1,3} & \dots & \mathbf{C}_{1,P} \\ \mathbf{C}_{2,1} & \mathbf{C}_{2,2} & \mathbf{C}_{2,3} & \dots & \mathbf{C}_{2,P} \\ \mathbf{C}_{3,1} & \mathbf{C}_{3,2} & \mathbf{C}_{3,3} & \dots & \mathbf{C}_{3,P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{P,1} & \mathbf{C}_{P,2} & \mathbf{C}_{P,3} & \dots & \mathbf{C}_{P,P} \end{bmatrix} \quad (5.183)$$

onde

$$\mathbf{C}_{i,j} = \sum_{k=1}^P \mathbf{A}_{i,k} \mathbf{D}_{k,j} \in \mathbb{R}^{p_i \times p_j}. \quad (5.184)$$

Mas, por hipótese de indução, $\mathbf{D}$ possui uma forma análoga àquela mostrada em (5.180), sendo que cada bloco $\mathbf{D}_{i,j}$ é não-nulo somente se $i \leq j$, consistindo, nesse caso, em um somatório de produtos de blocos $\mathbf{A}_{l,m}$ tais que $l \leq m \leq j$. Sendo assim, da definição de $\mathbf{A}$ temos que

$$\mathbf{C}_{i,j} = \begin{cases} \sum_{k=i}^j \mathbf{A}_{i,k} \mathbf{D}_{k,j}, & i \leq j \\ \mathbf{0} \in \mathbb{R}^{p_i \times p_j}, & \text{caso contrário} \end{cases} \quad (5.185)$$

onde os blocos $\mathbf{C}_{i,j}$ não-nulos consistem em um somatório de produtos na forma $\mathbf{A}_{i,k} \mathbf{D}_{k,j}$ em que $i \leq k \leq j$. Como, também por hipótese de indução, $\mathbf{D}_{k,j}$ é dado por um somatório de produtos de blocos $\mathbf{A}_{l,m}$ cujos índices satisfazem $l \leq m \leq j$, então claramente o mesmo pode ser dito sobre $\mathbf{C}_{i,j}$. Em outras palavras, a proposição do lema também é válida para $\mathbf{C} = \mathbf{A}^p$, o que completa a prova. ■

Teorema 5.2 Sejam $\{h_p\}_{p \in \{1, \dots, P\}}$ os núcleos que compõem a série de Volterra truncada obtida a partir de um sistema linear-analítico na forma (5.48)–(5.49) empregando-se a fórmula (3.56) ao sistema bilinear derivado através da bilinearização de Carleman de ordem P , i.e.,

$$h_p(\tau_1, \tau_2, \dots, \tau_p) = \mathbf{c}_p^T e^{\mathbf{F}\tau_1} \mathbf{G} e^{\mathbf{F}(\tau_2 - \tau_1)} \mathbf{G} \dots \mathbf{G} e^{\mathbf{F}(\tau_p - \tau_{p-1})} \mathbf{b} \quad (5.186)$$

em que $\mathbf{F}$, $\mathbf{G}$, $\mathbf{b}$ e $\mathbf{c}_P$ são construídos de acordo com a descrição do método de Carleman dada. Então, todas as matrizes $\mathbf{F}_{i,j}$ que compõem $\mathbf{F}$ e contribuem para o núcleo de ordem p são tais que

$$j \in \{1, \dots, p\} \text{ e } i \in \{1, \dots, j\} \quad (5.187)$$

e todas as matrizes $\mathbf{G}_{i,j}$ que compõem $\mathbf{G}$ e contribuem para o núcleo h_p são tais que

$$j \in \{1, \dots, p-1\} \text{ e } i \in \{1, \dots, j+1\}. \quad (5.188)$$

Demonstração. Inicialmente, vê-se que, de acordo com a construção descrita ao início desta seção, o vetor $\mathbf{b}$ possui a forma

$$\mathbf{b} = [\mathbf{G}_{1,0}^T \ 0 \ 0 \ \dots \ 0 \ 0]^T \quad (5.189)$$

em que $\mathbf{G}_{1,0} \in \mathbb{R}^n$. Note ainda que cada forma exponencial de (5.186),

$$e^{\mathbf{F}(\tau_k - \tau_{k-1})} = e^{\tau_k - \tau_{k-1}} \left(\mathbf{I} + \mathbf{F} + \frac{1}{2!} \mathbf{F}^2 + \frac{1}{3!} \mathbf{F}^3 + \dots \right) \quad (5.190)$$

sendo $k \in \{1, \dots, p\}$ e $\tau_0 = 0$, é dada por uma série de potências em $\mathbf{F}$. Tal matriz, por construção, possui uma forma em blocos como aquela da matriz $\mathbf{A}$ do Lema 5.1, sendo seus blocos dados por $\mathbf{F}_{i,j} \in \mathbb{R}^{m_i \times m_j}$, com $m_i = n^i$ e $m_j = n^j$. Além disso, pode-se escrever também a matriz identidade da série nesta mesma forma em blocos $\mathbf{I}_{i,j}$ tais que $j < i \implies \mathbf{I}_{i,j} = \mathbf{0} \in \mathbb{R}^{m_i \times m_j}$. Então, como a soma de matrizes nesta forma resulta em outra matriz de mesma forma, utilizando o resultado do lema verifica-se que $e^{\mathbf{F}(\tau_k - \tau_{k-1})}$ também a possuirá. Portanto, no que segue, escreveremos tais matrizes como

$$e^{\mathbf{F}(\tau_k - \tau_{k-1})} = \begin{bmatrix} \mathbf{F}_{1,1}^{(k)} & \mathbf{F}_{1,2}^{(k)} & \mathbf{F}_{1,3}^{(k)} & \dots & \mathbf{F}_{1,P}^{(k)} \\ \mathbf{0} & \mathbf{F}_{2,2}^{(k)} & \mathbf{F}_{2,3}^{(k)} & \dots & \mathbf{F}_{2,P}^{(k)} \\ \mathbf{0} & \mathbf{0} & \mathbf{F}_{3,3}^{(k)} & \dots & \mathbf{F}_{3,P}^{(k)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{F}_{P,P}^{(k)} \end{bmatrix} \quad (5.191)$$

e utilizaremos o fato de que, por consequência do Lema 5.1, os blocos $\mathbf{F}_{l,m}$ de $\mathbf{F}$ que contribuem para o bloco $\mathbf{F}_{i,j}^{(k)}$ de $e^{\mathbf{F}(\tau_k - \tau_{k-1})}$ são todos tais que $l \leq m \leq j$. Note que os blocos da matriz em (5.191) dependem de τ_k e de τ_{k-1} , porém tal dependência é omitida por simplicidade de notação.

Agora, consideremos o resultado do produto mais à direita de (5.186), que pode ser escrito como

$$\mathbf{b}_1 = e^{\mathbf{F}(\tau_p - \tau_{p-1})} \mathbf{b} \quad (5.192)$$

$$= \begin{bmatrix} \mathbf{F}_{1,1}^{(p)} & \mathbf{F}_{1,2}^{(p)} & \mathbf{F}_{1,3}^{(p)} & \dots & \mathbf{F}_{1,P}^{(p)} \\ \mathbf{0} & \mathbf{F}_{2,2}^{(p)} & \mathbf{F}_{2,3}^{(p)} & \dots & \mathbf{F}_{2,P}^{(p)} \\ \mathbf{0} & \mathbf{0} & \mathbf{F}_{3,3}^{(p)} & \dots & \mathbf{F}_{3,P}^{(p)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{F}_{P,P}^{(p)} \end{bmatrix} \begin{bmatrix} \mathbf{G}_{1,0} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (5.193)$$

$$= \begin{bmatrix} \left(\mathbf{F}_{1,1}^{(p)} \mathbf{G}_{1,0} \right)^T & 0 & \dots & 0 \end{bmatrix}^T. \quad (5.194)$$

Mas, de acordo com o parágrafo anterior, o único bloco de $\mathbf{F}$ que contribui para $\mathbf{F}_{1,1}^{(p)}$ é $\mathbf{F}_{1,1}$. Logo, o mesmo se aplica a $\mathbf{b}_1$.

Continuando a avaliação da expressão, consideremos agora o vetor

$$\mathbf{b}_2 = e^{\mathbf{F}(\tau_{p-1} - \tau_{p-2})} \mathbf{G} \mathbf{b}_1. \quad (5.195)$$

Usando (5.194) e a definição da matriz $\mathbf{G}$ (cuja forma é mostrada em (5.108)), temos que

$$\mathbf{b}_2 = e^{\mathbf{F}(\tau_{p-1} - \tau_{p-2})} \begin{bmatrix} \left(\mathbf{G}_{1,1} \mathbf{F}_{1,1}^{(p)} \mathbf{G}_{1,0} \right)^T & \left(\mathbf{G}_{2,1} \mathbf{F}_{1,1}^{(p)} \mathbf{G}_{1,0} \right)^T & 0 & \dots & 0 \end{bmatrix}^T \quad (5.196)$$

em que $\mathbf{G}_{1,1} \mathbf{F}_{1,1}^{(p)} \mathbf{G}_{1,0} \in \mathbb{R}^n$ e $\mathbf{G}_{2,1} \mathbf{F}_{1,1}^{(p)} \mathbf{G}_{1,0} \in \mathbb{R}^{n^2}$. Então, de (5.196) e (5.191)

obtem-se

$$\mathbf{b}_2 = \begin{bmatrix} \mathbf{F}_{1,1}^{(p-1)} \mathbf{G}_{1,1} \mathbf{F}_{1,1}^{(p)} \mathbf{G}_{1,0} + \mathbf{F}_{1,2}^{(p-1)} \mathbf{G}_{2,1} \mathbf{F}_{1,1}^{(p)} \mathbf{G}_{1,0} \\ \mathbf{F}_{2,2}^{(p-1)} \mathbf{G}_{2,1} \mathbf{F}_{1,1}^{(p)} \mathbf{G}_{1,0} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (5.197)$$

em que ambos os blocos que compõem $\mathbf{b}_2$ são vetores coluna, possuindo dimensões n e n^2 , respectivamente.

Da expressão (5.197) e da propriedade de que apenas os blocos $\mathbf{F}_{l,m}$ de $\mathbf{F}$ tais que $l \leq m \leq j$ contribuem para $\mathbf{F}_{i,j}^{(k)}$, pode-se ver que os blocos de $\mathbf{F}$ e $\mathbf{G}$ cujos elementos contribuem para o vetor $\mathbf{b}_2$ satisfazem as condições (5.187) e (5.188) para $p = 2$.

Se estendermos o raciocínio acima para o caso geral, definindo os vetores $\mathbf{b}_k$ recursivamente como

$$\mathbf{b}_k = e^{\mathbf{F}(\tau_{p-k+1}-\tau_{p-k})} \mathbf{G} \mathbf{b}_{k-1} \quad (5.198)$$

para $k \in \{2, \dots, p\}$, em que $\mathbf{b}_1$ é dado pela equação (5.194) e $\tau_0 = 0$, então, como resultado, teremos que os blocos $\mathbf{F}_{i,j}$ e $\mathbf{G}_{l,m}$ de $\mathbf{F}$ e $\mathbf{G}$ cujos elementos contribuem para o vetor $\mathbf{b}_k$ satisfazem condições semelhantes a (5.187) e (5.188), i.e., eles são tais que

$$j \in \{1, \dots, k\} \text{ e } i \in \{1, \dots, j\} \quad (5.199)$$

e

$$m \in \{1, \dots, k-1\} \text{ e } l \in \{1, \dots, m+1\}. \quad (5.200)$$

Basicamente, o motivo disso é que, devido à forma de $\mathbf{G}$ e ao fato de que o vetor $\mathbf{b}_{k-1}$ é tal que apenas seus $n + n^2 + \dots + n^{k-1}$ primeiros componentes podem ser não nulos, o resultado de $\mathbf{G} \mathbf{b}_{k-1}$ envolve apenas blocos de $\mathbf{G}$ que satisfazem (5.200) e corresponde a um vetor em que apenas os $n + n^2 + \dots + n^k$ primeiros componentes podem ser não-nulos. Note que tal acréscimo de elementos não-nulos no resultado se deve a $\mathbf{G}$ possuir blocos não-nulos na “diagonal de blocos” abaixo da “diagonal principal de blocos”. Já $e^{\mathbf{F}(\tau_{p-k+1}-\tau_{p-k})}$, por sua vez, não possui blocos não-nulos abaixo de sua “diagonal principal de blocos”, de maneira que $\mathbf{b}_k$ também é tal

que apenas seus $n + n^2 + \dots + n^k$ primeiros componentes podem ser não-nulos. Adicionalmente, por causa da nulidade de todos os componentes de $\mathbf{G}\mathbf{b}_{k-1}$ após o $(n + n^2 + \dots + n^k)$ -ésimo, apenas elementos dos blocos $\mathbf{F}_{i,j}^{(p-k+1)}$ cujos índices satisfazem (5.199) contribuirão para $\mathbf{b}_k$. Consequentemente, apenas elementos de blocos $\mathbf{F}_{i,j}$ de $\mathbf{F}$ cujos índices satisfazem (5.199) contribuirão para $\mathbf{b}_k$.

Com base na argumentação dada e na relação

$$h_p(\tau_1, \tau_2, \dots, \tau_p) = \mathbf{c}_p^T \mathbf{b}_p \quad (5.201)$$

então, fazendo-se $k = p$ em (5.199) e (5.200), o resultado do teorema é estabelecido. ■

Deve-se observar que, se a expansão de Taylor das funções $\mathbf{f}$ e $\mathbf{g}$ usada no método de Carleman é truncada na ordem P , então, para todo $p \in \{1, \dots, P\}$, todos os blocos $\mathbf{F}_{i,j}$ e $\mathbf{G}_{i,j}$ possíveis cujos índices satisfazem (5.187) e (5.188) já fazem parte das matrizes $\mathbf{F}$ e $\mathbf{G}$ resultantes. Assim, por consequência do Teorema 5.2, a seguinte proposição é válida.

Corolário 5.1 Se os núcleos $\{h_p\}_{p \in \{1, \dots, P\}}$ associados a um sistema na forma (5.48)–(5.49) são obtidos através do método de Carleman e da fórmula (5.186) usando-se expansões de Taylor das funções $\mathbf{f}$ e $\mathbf{g}$ truncadas na ordem P , então para qualquer $\tilde{P} > P$, os núcleos $\{\tilde{h}_p\}_{p \in \{1, \dots, P\}}$ obtidos da mesma forma porém usando-se expansões truncadas na ordem $\tilde{P}$ são tais que

$$\tilde{h}_p(\tau_1, \dots, \tau_p) \equiv h_p(\tau_1, \dots, \tau_p) \quad (5.202)$$

para $p \in \{1, \dots, P\}$. Em outras palavras, a expansão truncada na ordem P já leva em conta todos os termos das séries de Taylor de $\mathbf{f}$ e $\mathbf{g}$ que contribuem para os núcleos $h_1, \dots, h_P$. Note que, para $\tilde{P} \rightarrow \infty$, a forma (5.91) tende à dinâmica exata do sistema, dada por (5.48), enquanto os núcleos $h_1, \dots, h_P$, como visto, não se alteram. Portanto, tais núcleos são fornecidos pela bilinearização de Carleman truncada na ordem P de forma exata.

Apesar da discussão acima ter sido centrada nos núcleos de ordens 1 a P , o mesmo é válido para suas respectivas FRFGs. Na realidade, o aspecto central do método de Carleman que resulta nesta exatidão é que ele permite derivar um sistema bilinear cuja resposta envolve todos os componentes de ordens 1 a P que compõem a resposta do sistema linear-analítico original. Sendo assim, tanto os seus núcleos quanto suas FRFGs de ordens até P podem ser determinadas exatamente a partir do sistema bilinear resultante.

5.3.3 Aproximação de dinâmica não-linear por uma equação diferencial bilinear

Quando ao menos uma das funções $\mathbf{f}$ e $\mathbf{g}$ da equação de estado associada a um sistema linear-analítico é não-linear em $\mathbf{x}$, então, em sua expansão em série de Taylor, aparecerão termos de ordem maior que um em $\mathbf{x}$. Se bilinearizarmos tal sistema através da abordagem de Carleman sem restringir a ordem máxima de não-linearidade, então uma nova equação diferencial deve ser adicionada para cada ordem apresentada por um ou mais termos da expansão. Logo, quando tal expansão possuir termos de todas as ordens, vê-se facilmente que o sistema bilinear possuirá dimensão infinita.

No entanto, ainda que as expansões de $\mathbf{f}$ e $\mathbf{g}$ possuam um grau de não-linearidade finito em $\mathbf{x}$, a bilinearização conduzirá a um sistema de dimensão infinita cujo vetor de estados contém como componentes produtos de todas as ordens dos componentes de $\mathbf{x}$, o que pode ser um tanto contra-intuitivo. Isto pode ser compreendido através do seguinte exemplo.

Exemplo 5.2

Considere um sistema linear-analítico com vetor de estados unidimensional, x , e funções f e g dadas por

$$f(x) = \alpha_1 x + \alpha_2 x^2 + \dots + \alpha_P x^P \quad (5.203)$$

$$g(x) = \beta_1 x + \beta_2 x^2 + \dots + \beta_{P-1} x^{P-1} \quad (5.204)$$

em que todos os coeficientes α_i e β_i são não-nulos. Claramente, f e g são iguais às suas expansões em série de Taylor truncadas nas ordens P e $P - 1$, respectivamente. Aplicando-se a bilinearização de Carleman, o próximo passo seria formar a equação diferencial

$$\frac{d}{dt}(x^2) = 2x \frac{dx}{dt} \quad (5.205)$$

$$= 2x [(\alpha_1 x + \dots + \alpha_P x^P) + u(\beta_1 x + \dots + \beta_{P-1} x^{P-1})]. \quad (5.206)$$

Observe que, muito embora o sistema original possua apenas termos de ordem menor ou igual a P , (5.206) contém termos de ordem até $P + 1$, o que se deve à aplicação da regra da derivada expressa em (5.205). Sendo assim, para construir o sistema bilinear completo, uma equação diferencial para x^{P+1} também deve ser incluída. Se construirmos agora a equação diferencial para o termo x^3 , aparecerão então termos de ordem $P+2$, e assim por diante. Assim, pode-se mostrar que, na realidade, o sistema bilinear completo (i.e., equivalente ao sistema original) é de dimensão infinita, uma vez que seu vetor de estados é

$$\mathbf{x}^\otimes = [x \ x^2 \ x^3 \ x^4 \ \dots]^T. \quad (5.207)$$

Correspondentemente, a representação em série de Volterra associada ao sistema linear-analítico é composta por um número infinito de núcleos.

□

O exemplo dado evidencia um aspecto bastante positivo do método de Carleman: o de que ele permite vislumbrar de uma forma simples a relação entre as equações de espaço de estados de um sistema linear-analítico e sua representação de Volterra. Ao estabelecer esta relação através da redução das equações de estado à forma bilinear, ela possibilita também compreender como sistemas bilineares (de dimensão finita) são capazes de aproximar relações entrada/saída mais gerais, conforme mencionado na Subseção 3.1.2. De acordo com o conteúdo desta subseção e da anterior, conclui-se que através de sistemas bilineares pode-se obter uma aproximação da dinâmica não-linear de um sistema no mesmo sentido

daquela obtida com uma série de Volterra truncada.

5.4 Outros métodos estudados para o cálculo dos núcleos

Antes do problema de derivação dos núcleos de referência ter sido formulado em função das FRFGs associadas ao modelo físico, como descrito na Seção 5.1, considerou-se a possibilidade de calculá-los diretamente no domínio do tempo. Para tanto, estudaram-se também dois métodos de cálculo baseados no conteúdo de [31, Cap. 3]. Todavia, o que se verificou é que, para um sistema linear-analítico qualquer, tais métodos não são apropriados do ponto de vista prático. No que segue, uma descrição breve destes métodos será dada, apontando-se as causas dos problemas envolvidos em seu uso.

5.4.1 Método da função de fluxo

Antes de apresentarmos as expressões para o cômputo dos núcleos que correspondem ao método baseado na função de fluxo, é importante introduzirmos algumas convenções e notações adotadas em [31], as quais serão usadas adiante. Em primeiro lugar, define-se o operador de derivação ao longo de um campo vetorial, $L_{\mathbf{f}}\lambda$, como a seguir. Seja $\lambda : X \rightarrow \mathbb{R}$ e $\mathbf{f} : X \rightarrow X$. Então, a derivada de λ ao longo de $\mathbf{f}$ é dada por

$$L_{\mathbf{f}}\lambda(\mathbf{x}) = (\nabla_{\mathbf{x}}\lambda)^T \mathbf{f}(\mathbf{x}) = \sum_{i=1}^n \frac{\partial \lambda}{\partial x_i} f_i(\mathbf{x}) \quad (5.208)$$

em que x_i e f_i , com $i \in \{1, \dots, n\}$, são os componentes de, respectivamente, $\mathbf{x}$ e $\mathbf{f}$.

Segundo, para cada $\mathbf{x} \in X$, definiremos a função fluxo $\Phi : I_{\mathbf{x}} \times X \rightarrow X$, em que $I_{\mathbf{x}}$ é um intervalo aberto de $\mathbb{R}$ dependente de $\mathbf{x}$ e tal que $0 \in I_{\mathbf{x}}$, a qual possui as seguintes propriedades [31]:

(i) $\Phi(0, \mathbf{x}) = \mathbf{x}$;

(ii) a função $\sigma_{\mathbf{x}} : I_{\mathbf{x}} \rightarrow X$ definida como $\sigma_{\mathbf{x}}(t) = \Phi(t, \mathbf{x})$ é uma curva de

integração de $\mathbf{f}$;

(iii) $\Phi(t, \Phi(r, \mathbf{x})) = \Phi(s, \mathbf{x})$, em que $s = t + r$, contanto que $t, r, s \in I_{\mathbf{x}}$;

(iv) existe uma vizinhança aberta N de $\mathbf{x}$ tal que o mapeamento $\Phi_t : N \rightarrow X$ definido como $\Phi_t(\mathbf{w}) = \Phi(t, \mathbf{w})$ é inversível, sendo sua inversa denotada por $\Phi_{-t}(\mathbf{w})$, com $\mathbf{w} \in N$ e $t \in I_{\mathbf{x}}$.

De forma semelhante ao fluxo mencionado na Subseção 3.1.2, $\Phi(t, \mathbf{x}')$ é basicamente uma função que dá a solução da equação diferencial não forçada $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ com condições iniciais $\mathbf{x}'$.

Terceiro e último, a Jacobiana do mapeamento inverso Φ_{-t} será denotada por [31]

$$(\Phi_{-t})_{\star} = \frac{\partial \Phi_{-t}}{\partial \mathbf{x}}. \quad (5.209)$$

5.4.1.1 Fórmulas analíticas

Agora, podemos fornecer as expressões dadas em [31] para os núcleos de um sistema linear-analítico. Consideremos as funções $\mathbf{f}$, $\mathbf{g}$ e λ do modelo físico (5.48)–(5.49). Sejam a função escalar $Q_t(\mathbf{x}) : X \rightarrow \mathbb{R}$ dada por

$$Q_t(\mathbf{x}) = \lambda(\Phi_t(\mathbf{x})) = \mathbf{c}^T \Phi_t(\mathbf{x}) \quad (5.210)$$

e o campo vetorial $\mathbf{P}_t(\mathbf{x}) : X \rightarrow X$ definido como

$$\mathbf{P}_t(\mathbf{x}) = (\Phi_{-t})_{\star} \mathbf{g}(\Phi_t(\mathbf{x})). \quad (5.211)$$

Em seguida, definindo-se as funções

$$\tilde{h}_0(t) = Q_t(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{x}_0} \quad (5.212)$$

$$\tilde{h}_p(t, \sigma_1, \dots, \sigma_p) = L_{\mathbf{P}_{\sigma_p}} \dots L_{\mathbf{P}_{\sigma_1}} Q_t(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{x}_0} \quad (5.213)$$

em que $\mathbf{x}_0$ é o estado inicial do sistema, os núcleos triangulares associados à série de Volterra desejada podem ser obtidos analiticamente através das expressões

$$h_0 = \tilde{h}_0 \quad (5.214)$$

$$h_p(\tau_1, \dots, \tau_p) = \tilde{h}_p(t, t - \tau_1, \dots, t - \tau_p). \quad (5.215)$$

Observe que, como o sistema é invariante no tempo e, por consequência, $\mathbf{x}_0$ é um ponto de equilíbrio de $\mathbf{f}$ (ver Subseção 5.1.3), então $\tilde{h}_0$ não depende de t , pois

$$\tilde{h}_0(t) = \mathbf{c}^T \Phi_t(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{x}_0} = \mathbf{c}^T \mathbf{x}_0. \quad (5.216)$$

Em particular, admitiu-se naquela subseção que $\mathbf{x}_0 = \mathbf{0}$, o que implica em $h_0 = 0$. Analogamente, devido ao sistema ser invariante no tempo, $\tilde{h}_p$ também não depende efetivamente de t , e sim apenas das diferenças $t - \sigma_1, \dots, t - \sigma_p$.

Como exemplo da aplicação das fórmulas apresentadas, considere o sistema bilinear invariante no tempo

$$\dot{\mathbf{x}}(t) = \mathbf{F}\mathbf{x}(t) + \mathbf{G}\mathbf{x}(t)u(t) + \mathbf{b}u(t) \quad (5.217)$$

$$y(t) = \mathbf{c}^T \mathbf{x} \quad (5.218)$$

com $\mathbf{x}(0) = \mathbf{x}_0 = \mathbf{0}$. Para tal sistema, como visto na Subseção 3.1.2, a função fluxo é dada por

$$\Phi_t(\mathbf{x}) = e^{\mathbf{F}t} \mathbf{x}. \quad (5.219)$$

Assim, de acordo com as expressões (5.210) e (5.211), temos que

$$Q_t(\mathbf{x}) = \mathbf{c}^T e^{\mathbf{F}t} \mathbf{x} \quad (5.220)$$

$$\mathbf{P}_t(\mathbf{x}) = e^{-\mathbf{F}t} (\mathbf{G}e^{\mathbf{F}t} \mathbf{x} + \mathbf{b}). \quad (5.221)$$

Logo,

$$\tilde{h}_1(t, \sigma_1) = (\nabla_{\mathbf{x}} Q_t)^T \mathbf{P}_{\sigma_1}(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{0}} \quad (5.222)$$

$$= \mathbf{c}^T e^{\mathbf{F}t} e^{-\mathbf{F}\sigma_1} (\mathbf{G} e^{\mathbf{F}\sigma_1} \mathbf{0} + \mathbf{b}) \quad (5.223)$$

$$= \mathbf{c}^T e^{\mathbf{F}(t-\sigma_1)} \mathbf{b} \quad (5.224)$$

$$\tilde{h}_2(t, \sigma_1, \sigma_2) = \nabla_{\mathbf{x}} [(\nabla_{\mathbf{x}} Q_t)^T \mathbf{P}_{\sigma_1}(\mathbf{x})]^T \mathbf{P}_{\sigma_2} \quad (5.225)$$

$$= \mathbf{c}^T e^{\mathbf{F}t} e^{-\mathbf{F}\sigma_1} \mathbf{G} e^{\mathbf{F}\sigma_1} e^{-\mathbf{F}\sigma_2} (\mathbf{G} e^{\mathbf{F}\sigma_2} \mathbf{x} + \mathbf{b}) \Big|_{\mathbf{x}=\mathbf{0}} \quad (5.226)$$

$$= \mathbf{c}^T e^{\mathbf{F}(t-\sigma_1)} \mathbf{G} e^{\mathbf{F}(\sigma_1-\sigma_2)} \mathbf{b} \quad (5.227)$$

e assim por diante. Consequentemente,

$$h_1(\tau_1) = \mathbf{c}^T e^{\mathbf{F}\tau_1} \mathbf{b} \quad (5.228)$$

$$h_2(\tau_1, \tau_2) = \mathbf{c}^T e^{\mathbf{F}\tau_1} \mathbf{G} e^{\mathbf{F}(\tau_2-\tau_1)} \mathbf{b} \quad (5.229)$$

e assim por diante, consistentemente com a expressão (3.56) dada na Subseção 3.1.2.

Embora as equações (5.214) e (5.215) sejam úteis para a obtenção analítica dos núcleos de sistemas que possuem certas formas específicas, como é o caso do sistema bilinear considerado acima, seu uso para sistemas lineares-analíticos quaisquer é, em geral, difícil. Isto decorre do fato de que, para usá-las, é preciso conhecer a solução Φ_t para o caso não-forçado da equação diferencial não-linear (5.48). Assim, dada a evidente dificuldade que pode surgir na determinação de tal solução para um sistema qualquer, essa abordagem analítica não foi adotada neste trabalho.

5.4.1.2 Cômputo numérico através das fórmulas

Ao invés do método analítico descrito na subseção anterior, tentou-se explorar as expressões dadas para se computar numericamente os pontos da hipersuperfície correspondente a cada núcleo, o que não é discutido em [31]. Para se calcular, por exemplo, o valor do núcleo de primeira ordem no instante τ_1 , é necessário

determinar

$$h_1(\tau_1) = \tilde{h}_1(t, t - \tau_1) = L_{\mathbf{P}_{t-\tau_1}} Q_t(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{x}_0} \quad (5.230)$$

$$= \sum_{i=1}^n \frac{\partial Q_t}{\partial x_i} [\mathbf{P}_{t-\tau_1}(\mathbf{x})]_i \Big|_{\mathbf{x}=\mathbf{x}_0} \quad (5.231)$$

para algum $t \in \mathbb{R}$ tal que Q_t seja definida. Assim, escolhendo-se $t = \tau_1$, como $\Phi_0(\mathbf{x}) = \mathbf{x}$, então

$$h_1(\tau_1) = \sum_{i=1}^n \frac{\partial Q_{\tau_1}}{\partial x_i} [\mathbf{g}(\mathbf{x})]_i \Big|_{\mathbf{x}=\mathbf{x}_0}. \quad (5.232)$$

Como Q_{τ_1} é definida em função do fluxo, podem-se estimar as derivadas parciais de Q_{τ_1} utilizando-se uma rotina de integração numérica, introduzindo-se, para tanto, uma pequena variação em cada uma das direções de $\mathbf{x}$ em torno do estado inicial $\mathbf{0}$.

Agora, considere o cálculo de $h_2(\tau_1, \tau_2)$, que envolve o cômputo numérico de

$$h_2(\tau_1, \tau_2) = \tilde{h}_2(t, t - \tau_1, t - \tau_2) \quad (5.233)$$

$$= \sum_{i=1}^n \left\{ \frac{\partial}{\partial x_i} [L_{\mathbf{P}_{t-\tau_1}} Q_t(\mathbf{x})] \right\} [\mathbf{P}_{t-\tau_2}(\mathbf{x})]_i \Big|_{\mathbf{x}=\mathbf{x}_0} \quad (5.234)$$

$$= \sum_{i=1}^n \left\{ \frac{\partial}{\partial x_i} \left[\sum_{j=1}^n \frac{\partial Q_t}{\partial x_j} [(\Phi_{\tau_1-t})_* \mathbf{g}(\Phi_{t-\tau_1}(\mathbf{x}))]_j \right] \right\} \\ \times [(\Phi_{\tau_2-t})_* \mathbf{g}(\Phi_{t-\tau_2}(\mathbf{x}))]_i \Big|_{\mathbf{x}=\mathbf{x}_0}. \quad (5.235)$$

Fazendo, novamente, $t = \tau_1$, temos

$$h_2(\tau_1, \tau_2) = \sum_{i=1}^n \left\{ \frac{\partial}{\partial x_i} \left[\sum_{j=1}^n \frac{\partial Q_{\tau_1}}{\partial x_j} [\mathbf{g}(\mathbf{x})]_j \right] \right\} [(\Phi_{\tau_2-\tau_1})_* \mathbf{g}(\Phi_{\tau_1-\tau_2}(\mathbf{x}))]_i \Big|_{\mathbf{x}=\mathbf{x}_0}. \quad (5.236)$$

Mas, como $\Phi_{\tau_1-\tau_2}(\mathbf{x}_0) = \mathbf{x}_0 = \mathbf{0}$, então

$$h_2(\tau_1, \tau_2) = \sum_{i=1}^n \left\{ \frac{\partial}{\partial x_i} \left[\sum_{j=1}^n \frac{\partial Q_{\tau_1}}{\partial x_j} [\mathbf{g}(\mathbf{x})]_j \right] \right\} \Big|_{\mathbf{x}=\mathbf{0}} [(\Phi_{\tau_2-\tau_1})_* \mathbf{g}(\mathbf{0})]_i. \quad (5.237)$$

Perceba que, como o núcleo a ser obtido é triangular nos índices τ_1, τ_2 , então $\tau_2 - \tau_1 \geq 0$, e portanto, novamente, pode-se recorrer a rotinas de integração numérica para se estimar todas as derivadas envolvidas em (5.237). Analogamente, para o núcleo de terceira ordem, tal estratégia resulta na fórmula

$$h_3(\tau_1, \tau_2, \tau_3) = \tilde{h}_3(t, t - \tau_1, t - \tau_2, t - \tau_3) \quad (5.238)$$

$$= \sum_{i=1}^n \left\{ \frac{\partial}{\partial x_i} [L_{\mathbf{P}_{t-\tau_2}} L_{\mathbf{P}_{t-\tau_1}} Q_t(\mathbf{x})] \right\} [\mathbf{P}_{t-\tau_3}(\mathbf{x})]_i \Big|_{\mathbf{x}=\mathbf{x}_0}. \quad (5.239)$$

Para $t = \tau_1$, obtém-se

$$\begin{aligned} h_3(\tau_1, \tau_2, \tau_3) = & \sum_{i=1}^n \left\{ \frac{\partial}{\partial x_i} \left[\sum_{j=1}^n \frac{\partial}{\partial x_j} \left(\sum_{k=1}^n \frac{\partial Q_{\tau_1}}{\partial x_k} [\mathbf{g}(\mathbf{x})]_k \right) [(\Phi_{\tau_2-\tau_1})_* \mathbf{g}(\Phi_{\tau_1-\tau_2}(\mathbf{x}))]_j \right] \right\} \Big|_{\mathbf{x}=\mathbf{0}} \\ & \times [(\Phi_{\tau_3-\tau_1})_* \mathbf{g}(\mathbf{0})]_i. \quad (5.240) \end{aligned}$$

Neste caso, surge um problema: a fórmula requer o cômputo de derivadas de expressões que envolvem $\Phi_{\tau_1-\tau_2}(\mathbf{x})$ em relação aos componentes de $\mathbf{x}$. Como podemos ter $\tau_1 < \tau_2$, então, isto significa que há a necessidade de se realizar uma integração numérica de $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ no “sentido reverso”, o que é um problema difícil. Uma reversão no tempo torna a equação diferencial instável (admitindo-se, evidentemente, que o sistema original é estável), causando a amplificação de erros numéricos cometidos no procedimento de integração e posterior estimação das derivadas. Algumas tentativas foram feitas utilizando-se diversas funções de integração numérica disponíveis no MATLAB, porém, nenhuma solução obtida se provou correta. Realizou-se também um levantamento bibliográfico em busca de uma abordagem adequada para tal problema, que não foi encontrada.

Note que, para $\tau_1 \neq \tau_2$, não há escolha de t que possibilite eliminar simultaneamente $\Phi_{t-\tau_1}(\mathbf{x})$, $\Phi_{t-\tau_2}(\mathbf{x})$, $(\Phi_{\tau_1-t})_*$ e $(\Phi_{\tau_2-t})_*$ das expressões que são derivadas em relação aos componentes de $\mathbf{x}$. Logo, independentemente da escolha de t , a dificuldade acima descrita surge na estimação numérica das derivadas. Por fim, deve-se ressaltar ainda que as fórmulas acima fornecidas são válidas contanto que

não existam indeterminações nas expressões resultantes – por exemplo, observe que na passagem de (5.236) para (5.237), tomam-se os valores numéricos dos dois fatores que compõem o produto relativo a cada termo do somatório –, o que em geral não é previamente sabido; trata-se, portanto, de mais um aspecto negativo relacionado ao uso desta abordagem numérica para o cômputo dos núcleos.

5.4.2 Cálculo pela série de Taylor dos núcleos

Em [31, Cáp. 3], são fornecidas também expansões em série de Taylor das fórmulas para os núcleos dadas na abordagem da função de fluxo. Em se tratando de um sistema invariante no tempo, elas podem ser escritas como

$$h_0 = \lambda(\mathbf{x}_0) \quad (5.241)$$

$$h_p(\tau_1, \dots, \tau_p) = \sum_{i_p=0}^{\infty} \dots \sum_{i_1=0}^{\infty} L_{\mathbf{g}} L_{\mathbf{f}}^{i_1} L_{\mathbf{g}} \dots L_{\mathbf{g}} L_{\mathbf{f}}^{i_p} \lambda(\mathbf{x}) \frac{\tau_1^{i_p} \dots (\tau_p - \tau_{p-1})^{i_1}}{i_1! \dots i_p!} \Big|_{\mathbf{x}=\mathbf{x}_0} \quad (5.242)$$

em que o operador $L_{\mathbf{f}}^i$ pode ser definido recursivamente como

$$L_{\mathbf{f}}^i \mathbf{g}(\mathbf{x}) = \begin{cases} L_{\mathbf{f}} [L_{\mathbf{f}}^{i-1} \mathbf{g}(\mathbf{x})], & \text{se } i > 0 \\ \mathbf{g}(\mathbf{x}), & \text{se } i = 0 \end{cases}. \quad (5.243)$$

Observe que os coeficientes da série de Taylor são dados por

$$\frac{L_{\mathbf{g}} L_{\mathbf{f}}^{i_1} L_{\mathbf{g}} \dots L_{\mathbf{g}} L_{\mathbf{f}}^{i_p} \lambda(\mathbf{x})}{i_1! \dots i_p!} \Big|_{\mathbf{x}=\mathbf{x}_0} \quad (5.244)$$

e podem ser pré-calculados e armazenados para cada núcleo. A partir deles, ao se substituírem diferentes combinações de $\tau_1, \dots, \tau_p$ na série e se computar o resultado do somatório, pode-se determinar completamente h_p . Obviamente, contudo, o cálculo numérico do somatório deve ser feito truncando-se a série, e portanto seu resultado é obtido apenas aproximadamente.

Enquanto a estratégia descrita acima é, em princípio, eficaz na determinação numérica aproximada de pontos dos núcleos, seu uso prático pode ser inviável devido à necessidade de se determinar um número elevado de termos da série

para se obter uma boa aproximação. Como se pode ver pelas expressões dadas, o cômputo do coeficiente associado a cada termo requer o cálculo de uma sequência de derivadas, o que pode ser feito computacionalmente com o uso de um pacote de computação simbólica. No entanto, com o aumento dos valores dos índices $i_1, \dots, i_p$, as expressões resultantes tornam-se consideravelmente complexas, exigindo muitos recursos computacionais para a avaliação dos coeficientes.

Fazendo-se uso do pacote de computação simbólica do MATLAB, implementou-se o cômputo dos núcleos através desta abordagem. Porém, ao se tentar utilizá-lo na obtenção dos núcleos associados a um modelo físico para alto-falantes, o qual será descrito na Seção 6.1, os problemas mencionados no parágrafo anterior impediram que resultados úteis fossem obtidos.

Para este modelo, de fato o consumo de recursos computacionais associado ao método é alto. Por exemplo, para que a série relativa ao núcleo de primeira ordem convirja, são necessários muitos termos, o que foi constatado comparando-se tal núcleo quando calculado pelo método descrito com a resposta impulsiva do modelo físico linearizado, a qual é mostrada na Figura 5.4. Note que o núcleo de primeira ordem deve corresponder exatamente a tal resposta impulsiva, dado que a linearização corresponde a uma versão do método de Carleman com $P = 1$.

Figura 5.4: Resposta impulsiva de uma versão linearizada de um modelo físico para alto-falantes.

Na Figura 5.5, mostra-se tal comparação para os casos em que série de Taylor do núcleo de primeira ordem é truncada de modo a conter apenas 6, 7 e 8 termos. Observe que o acréscimo no número de termos permite um aumento muito pequeno do limite superior do intervalo de tempo em que a aproximação

é válida (repare que a unidade do eixo de tempo é o milissegundo). Contudo, o crescimento da carga computacional é bastante elevado: com 6 termos, foram necessários 6,45 segundos para o cálculo, com 7 termos, 136,77 segundos, e com 8, o tempo de computação subiu para mais de 37.000 segundos (ou seja, mais de 10 horas).

Figura 5.5: Comparação entre a resposta impulsiva do modelo físico do alto-falante linearizado (—) e núcleos de primeira ordem calculados através da expansão em série de Taylor com N termos (---), para $N \in \{6, 7, 8\}$.

Capítulo 6

Avaliação do desempenho de formalismos não-lineares com o uso de núcleos de referência

No capítulo anterior, descreveu-se uma forma de se obter núcleos de Volterra de referência através do uso de um modelo físico do sistema de interesse. Neste capítulo, pretende-se fornecer uma ideia sobre como tais núcleos podem ser explorados para a avaliação do desempenho de diferentes estruturas no modelamento deste sistema. Para tanto, será apresentado um estudo de caso em que se considera o problema de modelamento de um alto-falante no contexto de aplicações de áudio. Os passos envolvidos neste estudo são descritos a seguir.

Inicialmente, descreveremos um modelo físico para alto-falantes encontrado na literatura e apresentaremos os resultados obtidos com a aplicação do procedimento de derivação dos núcleos proposto a este modelo. Subsequentemente, o desempenho de um filtro de Volterra convencional formado pelos núcleos calculados será avaliado no que se refere à complexidade computacional associada ao seu uso e também à sua capacidade de aproximar a saída do modelo físico, que se supõe constituir uma representação precisa do sistema de interesse. Em seguida, uma avaliação semelhante será feita para as estruturas de Volterra alternativas que foram apresentadas no Capítulo 4. Por fim, discutiremos os resultados alcançados com o uso da metodologia proposta.

6.1 Um modelo físico para alto-falantes

No estudo de caso realizado, utilizou-se o modelo físico para alto-falantes proposto em [30, 35] como referência. Este modelo foi escolhido pelo fato de que os resultados reportados em [30] mostram que ele é capaz de representar precisamente o comportamento de um alto-falante automotivo real. Além disso, trata-se de um exemplo interessante para ilustrar a metodologia proposta, visto que o sistema de equações de estado do modelo não é trivial.

Contudo, foi preciso formular uma versão aproximada de tal modelo para que fosse possível usá-lo no estudo de caso. Isto se deve ao fato de que, originalmente, há uma função não-analítica envolvida em suas equações de estado, ao passo que as equações fornecidas como entrada ao algoritmo de bilinearização de Carleman devem possuir apenas funções analíticas, como visto no Capítulo 5.

No que segue, os princípios em que se baseia o modelo físico original serão apresentados, de acordo com o conteúdo de [30, 35]. Em seguida, discutiremos brevemente os resultados obtidos com o uso deste modelo que são reportados em [30]. Por fim, descreveremos a aproximação que foi adotada para se obter uma versão linear-analítica do modelo.

6.1.1 Princípios do modelo físico

Basicamente, o alto-falante eletrodinâmico é um transdutor que produz variações no nível de pressão acústica a partir de um sinal elétrico que representa a onda sonora a ser reproduzida. Tipicamente, ele consiste em um arranjo de elementos como aquele mostrado na Figura 6.1. O funcionamento deste dispositivo é descrito a seguir, com base em [11].

A excitação de um alto-falante eletrodinâmico ocorre quando a corrente elétrica produzida por um amplificador flui através de sua bobina de voz, a qual é situada no interior do campo magnético de um ímã permanente e envolve um cilindro conectado ao cone. Na presença de tal corrente, a força de Lorentz provoca o deslocamento da bobina e, por consequência, também do cone. Este cone é preso à armação por um aro flexível e por suspensões com certa rigidez conhecidas

como “aranhas”, as quais impedem a saída da bobina de voz do campo magnético. Por fim, a proteção contra poeira é usada para prevenir a entrada de partículas no interior do dispositivo.

Figura 6.1: Esquema seccional de um alto-falante eletrodinâmico típico [11].

Um modelo bastante popular utilizado por diversos fabricantes da indústria de áudio para a especificação e a análise de alto-falantes é o modelo de Thiele-Small [88, 89]. Porém, trata-se de um modelo linear que permite representar o comportamento do transdutor somente quando se consideram entradas com baixas amplitudes, as quais não provocam deslocamentos do cone para além da região em que ele opera de modo aproximadamente linear. Assim, tal modelo fornece uma boa representação apenas sob condições que, para determinadas aplicações, podem ser muito restritas. Como mencionado na Subseção 1.1.1, a tendência de miniaturização de dispositivos eletrônicos faz com que, em diversas situações práticas, seja necessário excursionar o cone de um alto-falante além de seu limite de linearidade.

As não-linearidades envolvidas na operação do dispositivo ilustrado na Figura 6.1 possuem três causas básicas: (i) a heterogeneidade do campo magnético produzido pelo ímã permanente; (ii) a dependência da indutância da bobina em relação à excursão do cone, devido ao fato de que o comprimento de sua seção situada na região em que o campo magnético atua varia com seu deslocamento; (iii) a variação da rigidez da suspensão do cone em função de sua excursão [35]. Devido a estas características, a relação entre a onda acústica produzida pelo alto-falante (i.e., sua saída) e o sinal elétrico que lhe é fornecido (i.e., sua entrada) é, em geral, não-linear.

Analisando a composição de um alto-falante eletrodinâmico típico, os autores de [30, 35] inicialmente modelaram tal sistema através de um circuito elétrico. Para tanto, estudaram-se as leis físicas envolvidas no funcionamento das partes elétrica e mecânica, para que em seguida os componentes do circuito fossem dispostos de modo a apresentar relações matemáticas análogas àquelas que ocorrem entre os componentes reais. Assim, obteve-se o circuito com parâmetros concentrados mostrado na Figura 6.2, cuja relação entrada/saída fornece um modelo do alto-falante. Este modelo será descrito no que segue, sendo que, para simplificar a notação, omitiremos os argumentos relativos ao tempo das variáveis envolvidas.

Figura 6.2: Esquema de circuito elétrico análogo a um alto-falante eletrodinâmico, originalmente proposto em [30].

Para a obtenção das equações de espaço de estados, os autores propuseram utilizar como variáveis de estado a excursão do cone, denotada por x_1 , sua velocidade, $x_2 = \dot{x}_1$, e a corrente que flui na bobina de voz, $x_3 = i$. Na parte elétrica do modelo, tem-se um resistor R que concentra a resistência da fonte e e da bobina de voz, um indutor L_1 e dois transformadores cujas relações de espiras dependem das variáveis de estado. Este artifício foi usado pelos autores para introduzir as não-linearidades tipicamente presentes em um alto falante. Através do transformador situado na parte de cima da figura, por exemplo, modela-se a indutância característica da bobina de voz, $L(x_1)$, que depende da excursão do cone, uma vez que o movimento deste está associado a um deslocamento da bobina em re-

lação ao ímã permanente. Assim, definindo-se a função n que fornece a relação de espiras de tal transformador como

$$n(x_1) = \sqrt{\frac{L(x_1)}{L_1}} \quad (6.1)$$

tem-se que a energia armazenada pelo indutor L_1 é

$$E_{\text{mag}} = \frac{1}{2} L_1 [n(x_1)i]^2 = \frac{1}{2} L(x_1) i^2 \quad (6.2)$$

conforme desejado.

A relação entre a força que atua sobre o cone, f , e a corrente da bobina de voz, i , também é modelada através de um transformador ideal. Neste caso, a relação de espiras que o define envolve a função

$$m(x_1, i) = A(x_1) + \frac{1}{2} \frac{dL(x_1)}{dx_1} i^2 \quad (6.3)$$

onde $A(x_1)$ é o fator de força. O primeiro termo de (6.3) está associado à força de Lorentz que resulta da passagem da corrente na bobina, enquanto o segundo se refere à força de relutância, causada pela variação da indutância da bobina com o deslocamento do cone. Na parte mecânica, a aplicação da segunda lei de Newton resulta em

$$M\ddot{x}_1 = f - G\dot{x}_1 - s \quad (6.4)$$

sendo M a massa do cone, $G\dot{x}_1$ a força de fricção e s a força de resistência da suspensão, que age como uma mola. Observe que a massa e o coeficiente de fricção são modelados como, respectivamente, uma capacitância e uma condutância. A força exercida pela suspensão é modelada considerando-se que sua rigidez depende da excursão do cone, isto é,

$$s(x_1) = k(x_1)x_1 \quad (6.5)$$

sendo $k(x_1)$ a função de rigidez. Para modelar tal relação não-linear, emprega-se

o transformador na parte mecânica, com relação de espiras dada por

$$l(x_1) = \sqrt{\frac{2}{[k(x_1)x_1]^2 L_2} E_{\text{pot}}} \quad (6.6)$$

onde E_{pot} é a energia potencial da suspensão, i.e.,

$$E_{\text{pot}} = \int_0^{x_1} k(\xi) \xi d\xi. \quad (6.7)$$

Usando-se (6.6)–(6.7) e a relação $i_2 = l(x_1)s(x_1)$, obtém-se para $s(x_1)$ a expressão dada por (6.5).

A partir da formulação acima, tem-se que as diferenças de potencial v_1 e v_2 mostradas na Figura 6.2 são dadas por

$$v_1 = n(x_1) \frac{d[L_1 n(x_1) i]}{dt} = \frac{1}{2} \frac{dL(x_1)}{dx_1} i \dot{x}_1 + L(x_1) \frac{di}{dt} \quad (6.8)$$

$$v_2 = m(x_1, i) \dot{x}_1 = A(x_1) \dot{x}_1 + \frac{1}{2} \frac{dL(x_1)}{dx_1} i \dot{x}_1. \quad (6.9)$$

Assim, a equação da malha na parte elétrica, $e = Ri + v_1 + v_2$, resulta em

$$e = Ri + \frac{dL(x_1)}{dx_1} i \dot{x}_1 + L(x_1) \frac{di}{dt} + A(x_1) \dot{x}_1. \quad (6.10)$$

Agora, podemos proceder à derivação das equações de estado. O vetor de estados é definido como

$$\mathbf{x} \triangleq [x_1 \quad \dot{x}_1 \quad i]^T = [x_1 \quad x_2 \quad x_3]^T. \quad (6.11)$$

Logo, a equação diferencial para x_1 é trivialmente dada por

$$\dot{x}_1 = x_2. \quad (6.12)$$

Para x_2 , fazendo uso de (6.4)–(6.5), temos

$$\dot{x}_2 = \ddot{x}_1 = \frac{1}{M} (f - G\dot{x}_1 - s) \quad (6.13)$$

$$= \frac{1}{M} [f - Gx_2 - k(x_1)x_1]. \quad (6.14)$$

Mas, observando o diagrama, temos que

$$f = m(x_1, x_3)x_3. \quad (6.15)$$

Assim, utilizando a equação (6.3), resulta

$$f = A(x_1)x_3 + \frac{1}{2}L'(x_1)x_3^2 \quad (6.16)$$

em que ' denota a derivação em relação a x_1 . Então, substituindo-se (6.16) em (6.14), obtém-se a equação diferencial

$$\dot{x}_2 = \frac{1}{M} \left[A(x_1)x_3 + \frac{1}{2}L'(x_1)x_3^2 - Gx_2 - k(x_1)x_1 \right]. \quad (6.17)$$

Por fim, a equação diferencial para x_3 é obtida a partir de (6.10), que pode ser escrita como

$$\dot{x}_3 = \frac{1}{L(x_1)} [e - A(x_1)x_2 - Rx_3 - L'(x_1)x_2x_3]. \quad (6.18)$$

Com isso, a representação em espaço de estados resultante é

$$\dot{\mathbf{x}} = \begin{bmatrix} x_2 \\ \frac{1}{M} \left\{ A(x_1)x_3 + \frac{x_3^2}{2}L'(x_1) - k(x_1)x_1 - Gx_2 \right\} \\ -\frac{1}{L(x_1)} \{ Rx_3 + L'(x_1)x_2x_3 + A(x_1)x_2 \} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{L(x_1)} \end{bmatrix} e \quad (6.19)$$

$$y = [1 \ 0 \ 0]\mathbf{x}. \quad (6.20)$$

Note que a saída do sistema considerada é dada pelo próprio deslocamento do cone, x_1 .

Por fim, é necessário fornecer as funções $A(x_1)$, $k(x_1)$ e $L(x_1)$ que modelam o comportamento do fator de força, da rigidez da suspensão do cone e da indutância

não-linear da bobina de voz, respectivamente. Conforme discutido em [30], tais funções ser aproximadas pelas formas

$$A(x_1) = \frac{a_0}{1 + a_1(x_1 - x_a)^2} \quad (6.21)$$

$$k(x_1) = k_0 + k_1(x_1 - x_k)^2 \quad (6.22)$$

$$L(x_1) = l_0 + l_1 \left(1 - \frac{x_1 - x_l}{l_2 + |x_1 - x_l|} \right) \quad (6.23)$$

com coeficientes positivos apropriados a_0 , a_1 , k_0 , k_1 , l_0 , l_1 , l_2 , e parâmetros para o modelamento de possíveis assimetrias x_a , x_k , x_l .

Deve-se notar que, como a rede elétrica apresentada é composta apenas por componentes passivos (à exceção da fonte, logicamente), o modelo resultante é tal que o equilíbrio em seu estado inicial, $\mathbf{x}_0 = \mathbf{0}$ (i.e., considera-se que o alto-falante está inicialmente em repouso), é estável [35].

6.1.2 Resultados apresentados pelo modelo físico

Em [30], emprega-se o modelo físico descrito na representação de um alto-falante automotivo para graves relativamente pequeno (com diâmetro de 82 mm). Para tanto, os autores de tal artigo obtiveram uma estimativa inicial dos parâmetros que compõem o modelo a partir de uma análise para pequenos sinais e, após isso, um refinamento foi feito através de um processo de adaptação. Nele, buscou-se a minimização do erro entre a corrente medida e aquela fornecida pelo modelo. Tal variável foi escolhida por ser a mais facilmente observável. Com isso, foram obtidos os valores mostrados na Tabela 6.1.

Para a validação do modelo físico obtido, realizou-se em [30] uma comparação entre a saída fornecida por ele e a saída do alto-falante real. Como entrada, empregou-se um sinal com energia na faixa de frequências de 20 a 500 Hz contendo certa ênfase em torno da frequência ressonante, 100 Hz, para que as não-linearidades fossem significativamente excitadas. Os gráficos dados em [30] mostram que o modelo oferece, de fato, uma predição da saída muito próxima da real. Contudo, infelizmente, o erro quadrático médio obtido com o uso do modelo

Função de rigidez $k(x_1)$			Função do fator de força $A(x_1)$		
Parâmetro	Unidade	Valores	Parâmetro	Unidade	Valores
k_0	N/m	$6,753 \times 10^2$	a_0	N/A	2,6607
k_1	N/m ³	$6,75 \times 10^4$	a_1	1/m ²	$9,2 \times 10^4$
x_k	m	$7,618 \times 10^{-4}$	x_a	m	$-3,463 \times 10^{-4}$

Função de indutância $L(x_1)$			Características dos componentes		
Parâmetro	Unidade	Valores	Parâmetro	Unidade	Valores
l_0	H	$1,22 \times 10^{-5}$	M	kg	3×10^{-3}
l_1	H	$2,309 \times 10^{-4}$	G	Ns/m	0,8319
l_2	m	3×10^{-3}	R	Ω	4,6254
x_l	m	-1×10^{-3}			

Tabela 6.1: Valores numéricos dos parâmetros para um auto-falante de graves automotivo que foi usado no experimento reportado em [30].

físico não é quantificado no artigo. Ao invés disso, menciona-se que tal modelo possibilitou uma redução de 15,6 dB do erro quadrático médio em comparação com sua versão linearizada no que se refere à predição da excursão do cone.

6.1.3 Ajuste do modelo físico para colocá-lo na forma linear-analítica

Como visto na Subseção 6.1.1, o modelo físico proposto em [30] envolve uma função de indutância não-linear $L(x_1)$, dada por (6.23), que claramente não é analítica. Na Figura 6.3, o comportamento de tal função e de suas derivadas primeira e segunda é mostrado. Para a geração das curvas, foram empregados os mesmos parâmetros usados em [30].

Observe que a derivada segunda de L apresenta uma descontinuidade em $x_l = -0,001$ m, o que é causado pelo uso do operador módulo na definição da função. Essa propriedade de L introduz limitações na capacidade de representação de um modelo de Volterra calculado a partir das equações de espaço de estados. Especificamente, como $L(x_1)$ não é dada por sua série de Taylor em torno da origem para valores de x_1 tais que $x_1 < x_l$, a série de Volterra formada por núcleos calculados através da bilinearização em torno de $\mathbf{0}$ não fornecerá uma representação adequada da dinâmica do sistema.

Para contornar as dificuldades introduzidas por tal comportamento, um ajuste foi feito ao modelo através da substituição da função $L(x)$ por uma aproximação

Figura 6.3: Comportamento da função de indutância não-linear (a), suas derivadas primeira (b) e segunda (c).

na forma

$$\tilde{L}(x_1) = b_0 + b_1 \tan^{-1}(\omega_l x_1 + \phi_l). \quad (6.24)$$

Usando um algoritmo de otimização não-linear e o critério de mínimos quadrados, os parâmetros de (6.24) foram estimados, resultando nos valores mostrados na Tabela 6.2. Com tais parâmetros, as curvas de $\tilde{L}(x_1)$ e suas derivadas são como mostradas na Figura 6.4. Por conveniência, as curvas correspondentes de $L(x_1)$ são também exibidas para comparação. Observe que L e $\tilde{L}$ são praticamente indistinguíveis, enquanto as derivadas da função aproximada se comportam de forma semelhante às de L , sendo, contudo, mais suaves na região em torno de x_l .

Para verificar o efeito que o uso de tal aproximação teve nas saídas produzidas pelo modelo, calculou-se a resposta do MF adaptado para a entrada mostrada na Figura 6.5, a qual foi empregada na validação do modelo físico em [30]. O resultado é mostrado na Figura 6.6, juntamente com a resposta do MF original, para comparação. O que se observa pelas curvas mostradas é que as saídas dos

Parâmetro	Valor estimado	Limites de confiança de 95%
b_0	$0,2423 \times 10^{-4}$	$(0,2421 \times 10^{-4}, 0,2425 \times 10^{-4})$
b_1	$-0,1303 \times 10^{-4}$	$(-0,1300 \times 10^{-4}, -0,1306 \times 10^{-4})$
ω_l	419,2	(416,7 , 421,7)
ϕ_l	0,4089	(0,4050 , 0,4128)

Tabela 6.2: Valores estimados para a aproximação de $L(x_1)$ dada por (6.24).

Figura 6.4: (a) Comparação entre a função de indutância não-linear (—) e sua aproximação (---). Em (b) e (c), a mesma comparação é feita para as derivadas primeira e segunda, respectivamente.

dois modelos são extremamente próximas, sendo bem difícil distingui-las visualmente. De fato, o erro quadrático médio normalizado apresentado pelo modelo ajustado, definido como

$$\text{EQMN}(\hat{y}) = \frac{\text{E}[(y - \hat{y})^2]}{\sigma_y^2} \quad (6.25)$$

é de -35,4 dB,¹ sendo que $\text{E}[\cdot]$ denota o operador de esperança matemática, $\hat{y}$ é

¹A definição para decibéis aqui empregada é $\text{dB}(x) = 10 \log_{10} x$.

a saída do modelo físico aproximado, y é a saída do modelo físico original e σ_y^2 é sua variância.

Figura 6.5: Entrada aproximadamente igual à usada em [30] para a validação do modelo físico do alto-falante.

Figura 6.6: Comparação entre as respostas (visualmente indistinguíveis) produzidas pelo MF original (—) e pelo MF ajustado (—) para a entrada da Figura 6.5.

Deve-se notar ainda que a aproximação introduzida é razoável uma vez que, como estamos lidando com um modelo de um sistema físico, não é de se esperar que haja variações abruptas nas variáveis envolvidas em sua operação. Assim, no restante do texto, quando o modelo físico do alto-falante for mencionado, estaremos nos referindo à versão adaptada aqui descrita, a qual conta com uma alteração apenas na função $L(x_1)$, já que nenhuma das demais funções envolvidas apresenta o mesmo problema anteriormente mencionado.

6.2 Derivação dos núcleos de referência

Como ponto de partida do estudo de caso, realizou-se a derivação de núcleos de referência com base no modelo físico do alto-falante descrito na seção anterior. Para o restante da discussão, é conveniente reproduzirmos aqui as equações que o definem, as quais são dadas por

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})e \quad (6.26)$$

$$= \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ f_3(\mathbf{x}) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ g_3(\mathbf{x}) \end{bmatrix} e \quad (6.27)$$

$$= \begin{bmatrix} x_2 \\ \frac{1}{M} \left\{ A(x_1)x_3 + \frac{x_2^2}{2} \tilde{L}'(x_1) - k(x_1)x_1 - Gx_2 \right\} \\ -\frac{1}{\tilde{L}(x_1)} \left\{ Rx_3 + \tilde{L}'(x_1)x_2x_3 + A(x_1)x_2 \right\} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{\tilde{L}(x_1)} \end{bmatrix} e \quad (6.28)$$

$$y = x_1 \quad (6.29)$$

em que $\mathbf{x} = [x_1 \ x_2 \ x_3]^T$, e as funções que compõem (6.28) são definidas como

$$A(x_1) = \frac{a_0}{1 + a_1(x_1 - x_a)^2} \quad (6.30)$$

$$k(x_1) = k_0 + k_1(x_1 - x_k)^2 \quad (6.31)$$

$$\tilde{L}(x_1) = b_0 + b_1 \tan^{-1}(\omega_l x_1 + \phi_l). \quad (6.32)$$

Os valores numéricos dos parâmetros adotados como referência são os mesmos do exemplo apresentado em [30], os quais foram mostrados na Tabela 6.1, à exceção daqueles que descrevem a função de indutância não-linear, $L(x_1)$, que foi substituída pela aproximação $\tilde{L}(x_1)$. Os valores relativos aos parâmetros que compõem esta aproximação são mostrados na Tabela 6.2.

Para a taxa de amostragem, adotou-se $F_s = 5$ kHz, o que corresponde a uma frequência de amostragem $\Omega_s = 10\pi \times 10^3$ rad/s. Tal valor é suficiente para a representação das características do alto-falante estudado, visto que ele é usado na reprodução de graves.

6.2.1 Obtenção das FRFGs associadas ao modelo físico

Com base nas funções $\mathbf{f}$ e $\mathbf{g}$ do modelo (6.26)–(6.29) e no estado inicial $\mathbf{x}_0 = \mathbf{0}$, aplicou-se o algoritmo descrito na Subseção 5.3.1.3 para se obter as matrizes $\mathbf{F}$ e $\mathbf{G}$ e o vetor $\mathbf{b}$ que compõem o sistema bilinearizado. Como ordem máxima de não-linearidade, definiu-se $P = 4$. As expansões de Taylor dos componentes de $\mathbf{f}$ e $\mathbf{g}$ foram determinadas empregando-se o pacote de computação simbólica do MATLAB.

Dado que $n = 3$ e $P = 4$, então, de acordo com a Tabela 5.1 (dada na Subseção 5.3.1.4), a equação de estados do sistema bilinear resultante possui dimensão $\overline{N}_P = 34$. Deve-se notar que, se uma implementação direta do método de Carleman fosse utilizada, este sistema teria dimensão $N_P = 120$.

Uma vez obtidas as estruturas que compõem o sistema bilinearizado, procedeu-se então ao cômputo numérico de pontos das FRFGs H_p simétricas associadas ao modelo físico, com $p \in \{1, 2, 3, 4\}$. Isto foi feito empregando-se as expressões que fornecem as FRFGs para sistemas bilineares, as quais foram dadas na Subseção 5.2.2, em um esquema de cômputo que explora a recursividade de tais expressões.

Como a finalidade da obtenção das FRFGs é a determinação dos núcleos de referência através da estratégia apresentada na Subseção 5.1.1, a qual se apoia na segunda abordagem de discretização descrita na Subseção 3.1.10.2, então, para cada uma delas computou-se um número de pontos idêntico àquele que compõe seu respectivo núcleo de tempo discreto. Logo, foi necessário definir de antemão qual seria a extensão de memória admitida para os núcleos. Inicialmente, considerou-se que as extensões de memória dos núcleos são dadas por $N_1 = N_2 = N_3 = N_4 = 100$, o que se mostrou um bom limite superior em face da frequência de amostragem adotada.

Devido à simetria de cada FRFG H_p , seu cômputo requer a determinação de um número de pontos dado por

$$\binom{N_p + p - 1}{p} \quad (6.33)$$

que é justamente o número de coeficientes que caracteriza um núcleo de Volterra

triangular de tempo discreto. Observe que, ao se levar em conta tais simetrias, pode-se reduzir consideravelmente o esforço computacional envolvido no cálculo, uma vez que (6.33) é significativamente menor que N_p^p para $p > 1$. Contudo, ainda que se use tal estratégia, o cálculo do núcleo de quarta ordem envolve um número elevado de pontos, sendo portanto a etapa mais custosa, em termos computacionais, da derivação dos núcleos de referência.

6.2.2 Cálculo dos núcleos de referência

De posse das FRFGs relativas ao modelo físico do alto-falante, H_p , o passo seguinte consistiu em determinar amostras das FRFGs associadas aos núcleos de referência de tempo discreto, V_p . Para tanto, de acordo com o conteúdo da Subseção 5.1.1, empregam-se as expressões (5.2) e (5.12), o que requer o conhecimento das respostas em frequência dos filtros de reconstrução, G_1 , e anti-rebatimento, G_2 . Este aspecto será tratado a seguir e, posteriormente, a obtenção dos núcleos a partir deles será discutida.

6.2.2.1 Definição dos filtros G_1 e G_2

Como será visto nas próximas seções deste capítulo, a avaliação experimental do desempenho de estruturas não-lineares é feita comparando-se, para um certo conjunto de entradas, as saídas produzidas por tais estruturas com aquelas produzidas pela estrutura de tempo discreto considerada na Subseção 5.1.1 (ver Figura 5.1), a qual inclui o modelo físico (6.26)–(6.29). As saídas deste modelo são calculadas com o uso de um procedimento de integração numérica, enquanto os filtros $G_1(j\Omega)$ e $G_2(j\Omega)$ foram aproximados por filtros de tempo discreto $\tilde{G}_1(e^{j\omega})$ e $\tilde{G}_2(e^{j\omega})$. Desta forma, para as amostras das respostas em frequência $G_1(j\Omega)$ e $G_2(j\Omega)$ necessárias na determinação dos núcleos, empregaram-se simplesmente amostras das respostas em frequência de tais filtros de tempo discreto.

Para que esta aproximação dos filtros de tempo contínuo por filtros de tempo discreto não incorresse em uma perda de precisão significativa, na simulação da estrutura da Figura 5.1 trabalhou-se com uma taxa de amostragem de RF_s , em

que R é um número inteiro tal que $R > 1$. Este esquema adotado será melhor explicado na próxima seção, em que se descreve o procedimento adotado para os experimentos. Em decorrência do uso desta estratégia, os filtros de tempo discreto utilizados na simulação, $\tilde{G}_1(e^{j\omega})$ e $\tilde{G}_2(e^{j\omega})$, são definidos com a taxa de amostragem de RF_s , de maneira que a relação entre as respostas em frequência dos filtros incorporados nos núcleos, $G_1(j\Omega)$ e $G_2(j\Omega)$, e os filtros de tempo discreto utilizados na simulação, $\tilde{G}_1(e^{j\omega})$ e $\tilde{G}_2(e^{j\omega})$, é dada por

$$G_i(j\Omega) = \tilde{G}_i(e^{j\Omega/RF_s}) \quad (6.34)$$

com $i \in \{1, 2\}$ e $\Omega \leq |\pi RF_s|$.

Considerando-se as características passa-baixas do modelo físico utilizado, definiu-se como banda de interesse o intervalo de frequências de $\omega_0 = 0$ até $\omega_f = \pi/4R$. Note que tal escolha, juntamente com (6.34), corresponde a definir a banda de interesse em tempo contínuo de $\Omega_0 = 0$ até $\Omega_f = 2\pi \times 625 = \Omega_N/4$, em que Ω_N é a frequência de Nyquist,² ou seja, $\Omega_N = \Omega_s/2$. Os filtros $\tilde{G}_1(e^{j\omega})$ e $\tilde{G}_2(e^{j\omega})$ foram projetados de modo a possuir uma frequência de corte ω_c um pouco menor que $2\omega_f$ e um comportamento suave na banda de transição. Mais especificamente, definiu-se $R = 20$ e $\omega_c = \pi/50$.

Uma mesma especificação foi usada para os filtros, isto é, adotou-se $\tilde{G}_1(e^{j\omega}) = \tilde{G}_2(e^{j\omega}) = \tilde{G}(e^{j\omega})$. O filtro $\tilde{G}(e^{j\omega})$ foi definido como um filtro FIR do tipo I (isto é, cuja resposta impulsiva possui simetria par e comprimento ímpar), tendo sido utilizado em seu projeto a implementação do algoritmo de Parks-McClellan [58] disponível no MATLAB. Escolheram-se como frequências limite da banda de passagem e da banda de rejeição $\omega_p = \omega_c = \pi/50$ e $\omega_r = 12\pi/250$, respectivamente. Além disso, fixou-se a ordem do filtro em $N_G = 1000$, o que permitiu uma atenuação de $\delta \approx -160$ dB. Os pesos empregados para a banda de passagem e para a banda de atenuação foram ambos iguais a um.

A resposta em frequência do filtro resultante é mostrada na Figura 6.7. Nos gráficos, exibe-se apenas a região de frequências de $\omega = 0$ até $\omega = \pi/10$, por

²É importante que se observe que se tal frequência de Nyquist se refere à taxa de amostragem original adotada para os experimentos, i.e., $F_s = 5$ kHz, e não à taxa usada para a simulação da estrutura que envolve o modelo físico, que é de RF_s .

questões de clareza. Como se vê em (a), devido a imprecisões numéricas, a ondulação do módulo desta resposta na banda de rejeição não é exatamente constante, apesar do algoritmo de Parks-McClellan ter sido empregado (o qual, em princípio, fornece respostas em frequência cujos módulos possuem ondulação constante tanto na banda de rejeição quanto na de passagem [58]). Isto é uma consequência da ordem elevada escolhida na especificação do filtro. Contudo, para os propósitos com que ele será empregado, tais imprecisões não constituem um problema.

Figura 6.7: Resposta em frequência do filtro $G(e^{j\omega})$: (a) módulo; (b) fase.

6.2.2.2 Núcleos resultantes

Tomando-se as TFDs inversas multidimensionais dos tensores formados pelas amostras das FRFGs (de tempo discreto) V_p calculadas, obtiveram-se, finalmente,

os núcleos de referência desejados. Em seguida, tais núcleos foram triangularizados, com o uso da fórmula (3.17) dada na Subseção 3.1.1.1, para compor os FVs convencionais empregados nos experimentos. Nas Figuras 6.8 e 6.9 exibem-se, respectivamente, os núcleos resultantes de primeira e de segunda ordem. Para facilitar a visualização deste último, foram mostrados apenas os coeficientes $h_2(k_1, k_2)$ relativos a índices $k_1, k_2 \leq 75$.

Figura 6.8: Núcleo de primeira ordem de referência, $h_1(k_1)$, obtido a partir do modelo (6.28)–(6.29).

Figura 6.9: Núcleo de segunda ordem de referência, $h_2(k_1, k_2)$, obtido a partir do modelo (6.28)–(6.29).

É importante observar que, como foi discutido ao final da Subseção 3.1.10.2, uma consequência do método de discretização aqui empregado é que os núcleos resultantes estão sujeitos ao efeito de aparecimento de componentes não-causais para índices próximos de N_p (i.e., ocorre um rebatimento no domínio do tempo). Para ilustrar este fenômeno, na Figura 6.10 mostram-se os valores de $h_1(k_1)$ para índices $k_1 \geq 70$. Observe que, no final da curva, há uma oscilação causada pelo fato de que o método de discretização usado corresponde a uma convolução circular com a função sinc no domínio do tempo.

Embora apareça de forma mais visível nas amostras de índices próximos ao comprimento de h_1 , a rigor, este efeito influencia todas as demais amostras, visto que a função sinc não possui suporte limitado. Além disso, ele influencia também os demais núcleos calculados. Porém, no caso deles, não é tão simples analisar a intensidade do efeito por inspeção visual.

Conforme visto na Subseção 3.1.10.2, uma estratégia para a atenuação deste efeito consiste no uso de uma janela senoidal. Deve-se notar que, devido às características da resposta em frequência do filtro passa-baixas $G(j\Omega)$ incorporado na derivação dos núcleos, ele já atua de maneira semelhante a uma janela senoidal para frequências acima da frequência de corte. De fato, a aplicação adicional de uma janela senoidal nas FRFGs antes do cômputo das TFDs inversas não oferece uma redução significativa deste efeito, como se mostra na Figura 6.11, em que o núcleo de referência obtido é comparado com um outro núcleo calculado da mesma forma, porém aplicando-se uma janela de von Hann (ver Subseção 3.1.10.2).

Como será visto adiante, tais imprecisões introduzidas pela discretização dos núcleos exerceram certa influência sobre os resultados experimentais obtidos. Contudo, uma compreensão completa sobre os efeitos observados nestes resultados e de como seria possível contorná-los ainda não foi alcançada, constituindo um tópico importante do prosseguimento deste trabalho. Para o propósito do estudo de caso desenvolvido neste capítulo, consideraremos os núcleos de tempo discreto obtidos como descrito acima (e sem o janelamento senoidal) e discutiremos posteriormente os efeitos observados que são consequências da discretização adotada.

Figura 6.10: Valores do núcleo de primeira ordem de referência, $h_1(k_1)$, para índices $k_1 \geq 70$.

Figura 6.11: Valores do núcleo de primeira ordem de referência, $h_1(k_1)$, para índices $k_1 \geq 70$, sem janelamento senoidal (—) e com janelamento senoidal (---).

6.2.3 Inadequação de modelos modulares simples

Tomando por base os núcleos de referência obtidos, pode-se concluir que alguns modelos modulares descritos na Seção 3.4 não são capazes de representar com precisão o comportamento do modelo físico. A seguir, apresenta-se uma argumentação sucinta que leva a tal conclusão para as estruturas modulares de Hammerstein e L-N.

Observando-se o núcleo de segunda ordem, o qual é mostrado na Figura 6.8, vê-se que ele não pode ser bem aproximado caso seja utilizado um modelo de Hammerstein ou de Hammerstein generalizado, visto que há uma parcela con-

siderável de energia distribuída sobre os coeficientes que não residem em sua diagonal principal. Quantitativamente, pode-se verificar tal fato para qualquer núcleo de ordem p (com $p > 1$) através da expressão

$$\eta_p = \frac{\sum_{k=1}^{N_2} h_p^2(k, \dots, k)}{\sum_{k_1=1}^{N_2} \dots \sum_{k_p=k_{p-1}}^{N_p} h_p^2(k_1, \dots, k_p)} \quad (6.35)$$

a qual fornece um indicativo da concentração da energia do núcleo na diagonal principal.

Note que, se $\eta_p \ll 1$, então o núcleo de referência de ordem p não pode ser aproximado com precisão por uma estrutura de Hammerstein ou de Hammerstein generalizada. Para os núcleos de referência calculados, verificou-se que $\eta_2 \approx 3,57 \times 10^{-2}$, $\eta_3 \approx 1,26 \times 10^{-4}$ e $\eta_4 = 1,91 \times 10^{-6}$, o que nos leva à conclusão de que, de fato, modelos modulares do tipo Hammerstein não são apropriados neste caso.

Já a estrutura L-N, como visto na Subseção 3.4.3, é equivalente a um filtro de Volterra cuja representação matricial do núcleo de segunda ordem é simétrica e possui posto unitário. Sendo assim, conclui-se que tal estrutura possibilita uma boa aproximação do núcleo de ordem dois, no sentido do erro quadrático médio, somente se a melhor aproximação de posto unitário do núcleo neste sentido for tal que

$$\frac{\|\mathbf{H}_2^* - \mathbf{H}_2\|_F^2}{\|\mathbf{H}_2\|_F^2} \approx 0 \quad (6.36)$$

em que $\mathbf{H}_2^*$ é a aproximação e $\mathbf{H}_2$ é a representação matricial do núcleo de segunda ordem de referência na forma simétrica. Com o uso da decomposição por valores singulares de $\mathbf{H}_2$, pode-se calcular $\mathbf{H}_2^*$ pela fórmula

$$\mathbf{H}_2^* = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T \quad (6.37)$$

em que σ_1 é o primeiro valor singular de $\mathbf{H}_2$ e $\mathbf{u}_1$ e $\mathbf{v}_1$ são os vetores singulares esquerdo e direito, respectivamente, associados a tal valor singular. Para o núcleo

de referência calculado, temos que

$$\frac{\|\mathbf{H}_2^* - \mathbf{H}_2\|_F^2}{\|\mathbf{H}_2\|_F^2} \approx 0,21. \quad (6.38)$$

Logo, a melhor aproximação de posto unitário apresenta um erro quadrático médio significativo, de modo que o modelo L-N não se apresenta como um bom compromisso entre compacidade e precisão. Mais adiante, veremos que aumentando um pouco o posto de $\mathbf{H}_2^*$, é possível reduzir significativamente o erro quadrático médio de tal aproximação.

6.3 Avaliação do desempenho de um FV composto pelos núcleos de referência

O primeiro aspecto que se pode avaliar com base nos núcleos de referência é, naturalmente, qual seria o desempenho de um filtro de Volterra por eles formado. Nesta seção, mostraremos como esta avaliação pode ser feita através de simulações computacionais, levando-se em conta aspectos específicos da aplicação de interesse, como as características típicas das entradas envolvidas.

6.3.1 Aproximação fornecida pelo FV

De posse dos núcleos de referência calculados conforme descrito na seção anterior, realizou-se a avaliação do desempenho de um filtro de Volterra convencional na aproximação da saída do modelo físico (6.28)–(6.29). Basicamente, tal avaliação foi feita por meio de um experimento Monte Carlo, no qual as respostas do FV para diversas sequências de entrada são comparadas às respectivas respostas produzidas pela estrutura de tempo discreto por ele modelada (a qual é descrita na Subseção 5.1.1). Tais entradas são geradas de modo a possuírem características espectrais como as daquelas encontradas na aplicação considerada e potências distribuídas sobre uma determinada faixa de interesse. Além disso, visando uma análise do ganho em precisão obtido com a inclusão de cada um

dos núcleos calculados, esta comparação é feita variando-se também o número de núcleos utilizados no FV.

Mais detalhadamente, o procedimento para a avaliação do FV é composto pelos passos descritos a seguir.

- 1) Especificam-se as entradas de interesse em termos de suas características espectrais e da faixa de potências (variâncias) considerada, a qual é definida em termos de seus limites, $\sigma_{\min}^2$ e $\sigma_{\max}^2$.
- 2) Gera-se um conjunto de M entradas aleatórias contendo as características espectrais definidas no passo anterior, média nula e variância unitária.
- 3) Para cada entrada $u(n)$ do conjunto gerado, geram-se L sequências na forma $\sigma_l u(n)$, com $l \in \{0, \dots, L-1\}$ e

$$\sigma_l = \sqrt{\sigma_{\min}^2 + \frac{l}{L-1} (\sigma_{\max}^2 - \sigma_{\min}^2)} \quad (6.39)$$

de maneira que suas variâncias distribuem-se uniformemente sobre a faixa estabelecida no primeiro passo.

- 4) Computa-se a saída da estrutura da Figura 5.1 para cada uma das sequências mencionadas no item anterior, empregando-se na simulação do modelo físico uma rotina de integração numérica.
- 5) Para cada sequência produzida no passo 3, computam-se também as aproximações de ordem $p \in \{1, \dots, P\}$ fornecidas pelo filtro de Volterra convencional composto pelos núcleos de referência. Em outras palavras, para cada $p \in \{1, \dots, P\}$, a saída do FV quando excitado por tal sequência é calculada empregando-se apenas os núcleos de ordens $1, \dots, p$.
- 6) Comparam-se as saídas do modelo físico e do filtro de Volterra convencional através do índice de ajuste (do inglês *fit*) dado por

$$\text{FIT}(\hat{y}) = 100 \times \left(1 - \frac{\|y - \hat{y}\|_2}{\|y - \mu_y\|_2} \right) \quad (6.40)$$

em que $\hat{y}$ é a saída do filtro de Volterra, y é a saída do modelo físico e μ_y é seu valor médio. Este índice fornece um valor que indica o percentual de similaridade entre duas sequências.

- 7) Para cada par (σ_l, p) , calcula-se a média dos índices de ajuste obtidos para todas as entradas geradas no segundo passo.

Em relação ao passo 4 do procedimento acima, como mencionado na seção anterior, emprega-se um esquema em que os filtros de tempo contínuo $G_1(j\Omega)$ e $G_2(j\Omega)$ são aproximados por filtros de tempo discreto $\tilde{G}_1(e^{j\omega})$ e $\tilde{G}_2(e^{j\omega})$. Assim, para que esta aproximação não introduza imprecisões que deteriore os resultados significativamente, a simulação é realizada trabalhando-se com uma taxa de amostragem de RF_s , em que R é um fator inteiro maior que 1 e F_s é a taxa de amostragem original. Em outras palavras, interpola-se a entrada do modelo físico por um fator de R e realiza-se o cômputo de sua saída de forma a se obter um sinal amostrado à taxa de RF_s . Em seguida, o resultado deste cômputo é decimado por um fator de R , de modo que a resposta final da estrutura é obtida na taxa de amostragem original. Tanto na interpolação quanto na decimação, utiliza-se o filtro $\tilde{G}(e^{j\omega})$ mostrado na Subseção 6.2.2.1.

Note que, para $R \gg 1$, o uso do esquema discutido no parágrafo anterior faz com que os resultados, em princípio, se aproximem daqueles que seriam obtidos com o uso de um sistema real descrito pela estrutura da Figura 5.1, utilizando-se filtros de reconstrução e anti-rebatimento cujas respostas em frequência são análogas à do filtro digital $\tilde{G}(e^{j\omega})$ no sentido da expressão (6.34). Observe que, procedendo-se desta forma, a simulação da estrutura de interesse é feita de forma muito mais simples do que se os filtros analógicos também fossem simulados.

Com o uso do procedimento apresentado, avaliou-se o desempenho de um FV convencional composto pelos núcleos de referência calculados. Para construir o conjunto de entradas, foram geradas $M = 100$ sequências de $Q = 500$ amostras de ruído Gaussiano contendo energia apenas na banda de interesse definida na Subseção 6.2.2.1, isto é, de 0 a $\pi/4$, o que corresponde ao intervalo de 0 a 625 Hz para a taxa de amostragem adotada ($F_s = 5$ kHz). Este limite superior foi escolhido pelo fato de que o alto-falante considerado no estudo se destina à repro-

dução de sinais de baixa frequência. Em relação às demais variáveis envolvidas na definição do experimento, adotou-se $L = 50$, $\sigma_{\min} = 0,02$ e $\sigma_{\max} = 5$. No que se refere ao cômputo da saída da estrutura simulada, definiu-se $R = 20$ para o esquema de interpolação/decimação descrito e utilizou-se a rotina de integração numérica `ode113` do MATLAB, a qual corresponde ao método de ordem variável de Adams-Bashforth-Moulton [60].

Com isso, foram obtidos os resultados mostrados na Figura 6.12. No gráfico, cada curva se refere a uma ordem de aproximação p , com $p \in \{1, \dots, 4\}$. Uma discussão dos resultados obtidos é feita no que segue.

Figura 6.12: Índices de ajuste das aproximações de ordem p dadas por um filtro de Volterra convencional composto pelos núcleos de referência.

6.3.1.1 Discussão dos resultados

Inicialmente, repare que a precisão de cada aproximação oferecida pelo filtro de Volterra de referência decai com o aumento da potência da entrada, como esperado. Isto se deve, em primeiro lugar, ao fato de que a representação de Volterra do modelo físico possui núcleos de todas as ordens, ao passo que cada aproximação de ordem p oferecida pelo filtro de Volterra é computada empregando-se um número finito de núcleos, p . Segundo, por se tratar de um sistema físico, espera-se que as contribuições dadas pelos núcleos de alta ordem sejam baixas para entradas de baixa potência e aumentem à medida em que tal potência aumenta.

Em particular, a aproximação de primeira ordem se degrada rapidamente, conforme se vê no gráfico. Admitindo-se que o modelo físico oferece de fato uma representação precisa do alto-falante estudado, isso evidencia a necessidade de se empregar um modelo não-linear quando o nível de precisão desejado é alto e as entradas pertinentes possuem potências elevadas.

Outro fato que se observa é que os índices de ajuste calculados não chegaram ao valor de 100%. Isto se deve à aproximação dos filtros de tempo contínuo da estrutura simulada por filtros de tempo discreto e também a imprecisões nos núcleos, as quais correspondem ao efeito discutido na Subseção 6.2.2.2. Em relação à primeira causa, verificou-se que ela influencia os resultados uma vez que um aumento nos índices de ajuste ocorre quando o fator R é aumentado. Um outro aspecto que afeta os resultados, embora provavelmente exerça menor influência sobre eles, são os erros numéricos cometidos pelo procedimento de integração empregado. Na realidade, um esclarecimento completo sobre a contribuição de cada um dos aspectos mencionados para os erros constatados nas aproximações ainda não foi alcançado, fazendo parte dos pontos previstos para o prosseguimento do estudo aqui realizado.

Os resultados obtidos com o experimento Monte Carlo são úteis para se avaliar qual é a ordem mínima necessária, p , para que um FV convencional atinja o nível de precisão desejado, dado um valor máximo de potência que as entradas podem possuir, σ_0^2 . Basicamente, p será dada pela menor ordem capaz de fornecer índices de ajuste superiores ou iguais a tal nível para entradas com potências menores ou iguais a σ_0^2 . Por exemplo, caso um nível de precisão correspondente a 97% de ajuste seja desejado e $\sigma_0^2 = 3$, então, de acordo com a Figura 6.12, deve-se escolher um FV de ordem $p = 3$.

Para finalizar, é importante ter em mente que, por se tratar de uma simulação Monte Carlo, em princípio, os resultados obtidos convergem para os valores reais assintoticamente, ou seja, apenas quando $M \rightarrow \infty$ (admitindo-se que o experimento é consistente e não induz viés). Na prática, contudo, pode-se monitorar a variância dos índices de ajuste que compõem as curvas à medida em que novas realizações de entradas são consideradas e estabelecer um critério de parada.

No experimento acima descrito, verificou-se que, para $M > 40$, cada variável aleatória $\phi_{l,p}$ relativa ao índice de ajuste do par (σ_l, p) distribui-se de forma que

$$\gamma_{l,p} = 2 \frac{\sigma_{\phi_{l,p}}}{\mu_{\phi_{l,p}}} < 0,1 \quad (6.41)$$

onde $\sigma_{\phi_{l,p}}$ é seu desvio padrão e $\mu_{\phi_{l,p}}$ sua média. Na Figura 6.13 mostra-se, por exemplo, o comportamento de $\gamma_{l,p}$ em função do número de realizações, M , para $M \in \{11, 12, \dots, 100\}$, $l \in \{30, 40, 50\}$ e $p \in \{1, 2, 3, 4\}$.

Figura 6.13: Comportamento de $\gamma_{l,p}$ para $l = 30$ (—), $l = 40$ (---) e $l = 50$ (-•). Os valores de p são: (a) $p = 1$, (b) $p = 2$, (c) $p = 3$ e (d) $p = 4$.

6.3.2 Comportamento das aproximações em um intervalo de potências maior

Observando-se as curvas da Figura 6.12, percebe-se que, com o aumento do número de núcleos usados no cômputo da saída do FV, a aproximação sempre melhora na faixa de potências considerada. Além disso, para este conjunto de entradas, o ganho em precisão trazido pelo acréscimo de um núcleo sempre aumenta com a potência da entrada. Contudo, analogamente ao que ocorre com aproximações polinomiais de funções que crescem mais lentamente que polinômios, este comportamento é válido somente até uma determinada potência de entrada, σ_d^2 . A partir de tal potência, espera-se que, ao contrário do que se constata na Figura 6.12, haja uma tendência de deterioração da aproximação com o aumento da ordem.

Na Figura 6.14 este fato é mostrado, exibindo-se os resultados de um experimento semelhante àquele que resultou na Figura 6.12, porém desta vez com $\sigma_{\min}^2 = 5$ e $\sigma_{\max}^2 = 50$. Observe que, considerando-se os núcleos de referência obtidos, a aproximação fornecida pelo FV de referência piora com o aumento de p para entradas com potência maior do que $\sigma^2 = 25$.

Figura 6.14: Comportamento das aproximações fornecidas pelo FV convencional de referência em um intervalo de potências maior.

6.3.3 Avaliação de complexidades estrutural e computacional do FV

Com a informação relativa à ordem mínima necessária para o modelamento de um sistema, obtida através da análise mostrada na Subseção 6.3.1, tem-se uma estimativa do número de coeficientes que um filtro de Volterra convencional necessita para que o nível de precisão desejado seja atingido. Tal estimativa, por sua vez, fornece um indicativo da complexidade computacional associada ao uso deste filtro, como foi discutido no Capítulo 4.

No experimento realizado, dado que a extensão de memória escolhida para os núcleos foi relativamente alta, tem-se um FV resultante com um total de coeficientes bastante elevado. Este fato é demonstrado pelos dados mostrados na Tabela 6.3, os quais foram calculados considerando-se que os núcleos são representados na forma triangular.

De acordo com tais dados, o núcleo de quarta ordem compreende mais de quatro milhões e quatrocentos mil coeficientes (o que corresponde a mais de 95% do total), e portanto seu uso é extremamente dispendioso do ponto de vista computacional. No entanto, ainda que se considerem apenas os núcleos de ordens $p \leq 3$, o uso de um FV convencional composto por tais núcleos implica na implementação de uma estrutura composta por 176.850 coeficientes, o que ainda pode ser um número elevado para aplicações que necessitam do cômputo de predições em tempo real. Note que, de acordo com o que foi discutido na Seção 4.1, uma implementação direta de tal filtro necessitaria de um número de operações aritméticas da mesma ordem que o total de coeficientes do filtro.

Ordem do núcleo (p)	Memória	Coeficientes
1	100	100
2	100	5.050
3	100	171.700
4	100	4.421.275
Total	—	4.598.125

Tabela 6.3: Número de coeficientes de cada um dos núcleos de referência calculados.

Como uma primeira tentativa de se reduzir a complexidade estrutural do FV, uma alternativa seria simplesmente reduzir as extensões das memórias dos núcleos obtidos, i.e., considerar para cada núcleo de referência h_p apenas os coeficientes

que se referem a índices $k_1, \dots, k_p \leq \tilde{N}_p \leq N_p$. Evidentemente, espera-se que a precisão da aproximação decaia com esta redução, de modo que seria necessário realizar novamente o procedimento de avaliação do desempenho para o novo FV obtido.

Para demonstrar os efeitos de uma redução do número de coeficientes feita dessa forma, considerou-se um segundo filtro de Volterra convencional de quarta ordem, $\mathcal{F}_2$, cujos núcleos foram obtidos reduzindo-se as extensões da memória dos núcleos de referência para $N_1 = \dots = N_4 = 80$, ou seja, descartaram-se alguns coeficientes dos núcleos originais. Na Tabela 6.4, mostram-se os números de coeficientes de cada um destes novos núcleos. De acordo com os dados mostrados, tais núcleos totalizam 1.929.500 coeficientes, contra 4.598.125 do filtro original, o qual denotaremos por $\mathcal{F}_1$. Ou seja, esta redução na extensão da memória permitiu eliminar 2.668.625 coeficientes, o que corresponde a mais de 58% do total de $\mathcal{F}_1$.

Ordem do núcleo (p)	Memória	Coeficientes
1	80	80
2	80	3.240
3	80	88.560
4	80	1.837.620
Total	—	1.929.500

Tabela 6.4: Número de coeficientes dos núcleos do filtro de Volterra convencional $\mathcal{F}_2$.

Na Figura 6.15, comparam-se os desempenhos de $\mathcal{F}_1$ e de $\mathcal{F}_2$ na aproximação das saídas do modelo físico. Os resultados para $\mathcal{F}_2$ foram obtidos utilizando-se as mesmas $M = 100$ entradas e demais parâmetros do procedimento que haviam sido empregados na avaliação de $\mathcal{F}_1$. Observe que os filtros possuem desempenhos muito próximos em termos de precisão, o que indica que $N = 80$ é uma melhor escolha para as extensões de memória dos núcleos.

Já quando se considera um terceiro filtro de Volterra em cujos núcleos são preservadas apenas as amostras relativas a índices $k_1, \dots, k_p < 70$, há uma visível degradação no desempenho, como mostrado na Figura 6.16. Nela, compara-se o desempenho deste novo filtro de Volterra, $\mathcal{F}_3$, o qual possui extensão de memória de $N = 70$, com o de $\mathcal{F}_1$. Por outro lado, o descarte de coeficientes feito na

Figura 6.15: Índices de ajuste das aproximações de ordem p dadas por $\mathcal{F}_1$ (—) e por $\mathcal{F}_2$ (---).

definição de $\mathcal{F}_3$ conduz a uma redução ainda maior da complexidade estrutural em relação ao FV original, como se vê na Tabela 6.5. De acordo com os dados, $\mathcal{F}_3$ possui apenas cerca de 25% dos coeficientes de $\mathcal{F}_1$.

Figura 6.16: Índices de ajuste das aproximações de ordem p dadas por $\mathcal{F}_1$ (—) e por $\mathcal{F}_3$ (---).

Como forma de se ter uma ideia concreta da complexidade computacional dos filtros $\mathcal{F}_1$, $\mathcal{F}_2$ e $\mathcal{F}_3$, na Tabela 6.6 mostram-se os tempos médios gastos no cômputo de suas respostas feito nos experimentos. Tais filtros de Volterra foram implementados de maneira *offline* e empregando-se a representação diagonal dos

Ordem do núcleo (p)	Memória	Coeficientes
1	70	70
2	70	2.485
3	70	59.640
4	70	1.088.430
Total	—	1.150.625

Tabela 6.5: Número de coeficientes dos núcleos do filtro de Volterra convencional $\mathcal{F}_3$.

núcleos, a qual é descrita na Subseção 4.2.1. Contudo, o cálculo das convoluções envolvidas foi feito no domínio do tempo, visto que, dado o comprimento reduzido dos filtros que correspondem às diagonais dos núcleos, não há vantagem em fazê-lo no domínio da frequência.

O computador utilizado possui um processador Intel Core i7 2600 de 3.4 GHz e 8 gigabytes de memória RAM DDR3 de 1333 MHz. Para reduzir a influência de certos aspectos específicos do hardware empregado, o cálculo das respostas não foi implementado explorando-se a disponibilidade dos múltiplos núcleos do processador, i.e., utilizou-se uma implementação sequencial. Além disso, durante o cômputo das respostas, verificou-se que o sistema operacional não realizou operações de paginação; portanto, não houve influência deste aspecto nos tempos de execução medidos.

Os valores dados na Tabela 6.6 foram agregados por ordem de aproximação considerada, i.e., de acordo com o número de núcleos envolvido na estrutura utilizada. Com isso, tem-se uma ideia do que a inclusão de um núcleo representa em termos do aumento da carga computacional. Adicionalmente, como são mostrados os tempos médios gastos no uso de $\mathcal{F}_1$, $\mathcal{F}_2$ e $\mathcal{F}_3$, pode-se ver como a redução das extensões de memória dos núcleos afetou o tempo de computação esperado.

Com base nos dados fornecidos, observa-se que tal redução, de fato, causa uma diminuição no tempo médio gasto para o cômputo das aproximações. Contudo, esta diminuição é apenas marginal quando comparada com o acréscimo de tempo trazido pelo aumento da ordem da aproximação. Em particular, um tempo de computação elevado é consumido no cômputo das aproximações de quarta ordem, considerando-se que as sequências de excitação contêm apenas $Q = 500$ amostras cada.

Note ainda que, caso uma aproximação de terceira ordem fosse necessária, então o tempo médio gasto para o cômputo de um bloco de $Q = 500$ amostras da saída seria de, no mínimo, pouco mais 0,36 segundos. Como a taxa de amostragem empregada é de $F_s = 5$ kHz, isto significa que a determinação das amostras relativas a um trecho de 0,1 segundo da saída necessitaria de mais de três vezes este tempo, o que mostra que o uso dos filtros de Volterra convencionais obtidos para fornecer aproximações de terceira ordem seria problemático em aplicações de tempo real.

Filtro	Tempo médio de cálculo (seg)			
	$p = 1$	$p = 2$	$p = 3$	$p = 4$
$\mathcal{F}_1$	$1,72 \times 10^{-4}$	$1,55 \times 10^{-2}$	$7,46 \times 10^{-1}$	$2,56 \times 10^1$
$\mathcal{F}_2$	$1,12 \times 10^{-4}$	$1,18 \times 10^{-2}$	$4,84 \times 10^{-1}$	$1,34 \times 10^1$
$\mathcal{F}_3$	$7,95 \times 10^{-5}$	$1,01 \times 10^{-2}$	$3,67 \times 10^{-1}$	8,92

Tabela 6.6: Tempo médio de cálculo das aproximações de ordem p fornecidas pelos FVs convencionais $\mathcal{F}_1$, $\mathcal{F}_2$ e $\mathcal{F}_3$.

6.4 Avaliação do desempenho de estruturas de Volterra alternativas

Os resultados obtidos na seção anterior mostram que, quando um modelamento de alta ordem ($p = 3$ ou $p = 4$) é necessário, então a complexidade estrutural de um filtro de Volterra convencional que aproxima a saída do modelo físico (6.28)–(6.29) é elevada. Além das consequências que isto tem no que se refere à complexidade computacional associada ao uso de tal filtro, um grande número de coeficientes faz com que o dispêndio de recursos computacionais necessário para a adaptação do modelo também seja alto.

Conforme se discutiu no Capítulo 4, em tais situações uma alternativa seria recorrer a estruturas que foram propostas na literatura com o intuito de aproximar filtros de Volterra empregando, para tanto, um número de parâmetros bastante reduzido. Nesta seção, mostraremos como os núcleos de referência podem ser explorados para se avaliar o desempenho das estruturas apresentadas no Capítulo

4 quando utilizadas na aproximação das saídas do modelo físico do alto-falante.

6.4.1 Filtro de Volterra diagonal aproximado

Nesta subseção, faremos uma avaliação semelhante à apresentada na Seção 6.3, considerando o desempenho de filtros de Volterra diagonais aproximados, os quais foram discutidos na Seção 4.2.

Com base nos núcleos de referência obtidos, é fácil construir um filtro de Volterra cujos núcleos são decompostos em diagonais, como descrito na Subseção 4.2.1. Uma vez que tal estrutura é construída, pode-se empregar a estratégia de aproximação descrita na Subseção 4.2.2, a qual consiste em descartar os coeficientes de cada núcleo que residem em diagonais afastadas de sua diagonal principal. Como visto naquela subseção, para tanto, deve-se estabelecer uma “distância” máxima, W_p , para cada núcleo de ordem p , a qual define que diagonais serão descartadas.

Ao se observar, por exemplo, o núcleo de referência de segunda ordem (mostrado na Figura 6.9), vê-se que, de fato, sua energia concentra-se em torno da diagonal principal. Esta constatação corrobora a discussão feita na Subseção 4.2.2, segundo a qual núcleos associados a sistemas físicos tendem a possuir tal característica. Logo, espera-se que uma aproximação obtida com o descarte de coeficientes distantes da diagonal principal afete pouco a precisão da aproximação.

Em relação aos núcleos de alta ordem ($p > 2$), a avaliação deste aspecto por inspeção visual é, obviamente, mais difícil. Variando-se os valores de W_p , é possível, contudo, avaliar como o descarte de coeficientes afeta a precisão das aproximações que envolvem estes núcleos e também a complexidade computacional associada ao uso da estrutura resultante. Estas questões serão tratadas na sequência através de simulações computacionais.

6.4.1.1 Avaliação das aproximações obtidas

Na avaliação experimental realizada, empregaram-se dois filtros de Volterra diagonais aproximados, $\mathcal{D}_1$ e $\mathcal{D}_2$, cada um deles construído a partir dos quatro

núcleos de referência calculados. Em face dos resultados obtidos na seção anterior, foram utilizadas apenas as amostras de tais núcleos cujos índices associados satisfazem $k_1, \dots, k_p < 80$, ou seja, empregaram-se os núcleos de $\mathcal{F}_2$.

Tal escolha foi feita pelo fato de que $\mathcal{F}_2$ é aproximadamente tão preciso quanto $\mathcal{F}_1$ possuindo um número bastante menor de coeficientes, e portanto é mais apropriado para a comparação com as estruturas alternativas. Já $\mathcal{F}_3$, por sua vez, teve um desempenho inferior aos demais filtros convencionais, não constituindo, por conseguinte, uma boa referência para construção das estruturas alternativas de Volterra.

Para $\mathcal{D}_1$, utilizou-se $W_2 = W_3 = W_4 = 50$, uma vez que tal valor trouxe pouca perda em precisão, como será visto. Já para $\mathcal{D}_2$, definiu-se $W_2 = W_3 = W_4 = 30$.

Uma vez construídas as estruturas dos filtros diagonais aproximados, realizou-se um experimento de Monte Carlo como aquele descrito na Subseção 6.3.1, utilizando-se as mesmas $M = 100$ sequências de entrada e também os mesmos valores para os parâmetros R , L , $\sigma_{\min}^2$ e $\sigma_{\max}^2$. Contudo, desta vez, as saídas da estrutura que contém o modelo físico são comparadas, evidentemente, com aquelas produzidas pelos filtros $\mathcal{D}_1$ e $\mathcal{D}_2$.

Nas Figuras 6.17 e 6.18, mostram-se os desempenhos dos filtros diagonais aproximados $\mathcal{D}_1$ e $\mathcal{D}_2$, respectivamente. Para fins de comparação, exibem-se também as curvas relativas às aproximações fornecidas pelo filtro de Volterra convencional $\mathcal{F}_2$, cujos núcleos serviram como base para $\mathcal{D}_1$ e $\mathcal{D}_2$.

Observe que, mesmo com a definição de $W_p = 50$ para todos os núcleos, o filtro $\mathcal{D}_1$ teve um desempenho muito próximo de $\mathcal{F}_2$, sendo impossível distinguir as curvas de mesma ordem visualmente. Logo, o descarte de coeficientes implicado por esta definição de W_p causa, de fato, uma perda não-significativa de precisão, sendo recomendável para a diminuição da complexidade estrutural. Como visto na seção anterior, uma simples redução da extensão das memórias dos núcleos de $\mathcal{F}_2$ já implica em uma perda significativa, ao passo que, quando se descartam apenas coeficientes afastados das diagonais principais de tais núcleos, o mesmo não ocorre.

Já o desempenho de $\mathcal{D}_2$ não se mostrou tão bom para as aproximações de

Figura 6.17: Índices de ajuste das aproximações de ordem p dadas por $\mathcal{F}_2$ (—) e por $\mathcal{D}_1$ (---).

ordens $p = 3$ e $p = 4$, devido ao maior descarte de coeficientes implicado pela definição $W_p = 30$ para todos os núcleos. É evidente que, mesmo com a perda de desempenho observada para os núcleos de alta ordem, a aproximação oferecida por $\mathcal{D}_2$ pode ser suficiente, o que dependerá da precisão desejada e da potência máxima admitida para as entradas.

Figura 6.18: Índices de ajuste das aproximações de ordem p dadas por $\mathcal{F}_2$ (—) e por $\mathcal{D}_2$ (---).

6.4.1.2 Avaliação da redução de complexidades estrutural e computacional

Vejamos, agora, qual é o ganho em compacidade trazido pelo uso dos filtros diagonais aproximados $\mathcal{D}_1$ e $\mathcal{D}_2$. Na Tabela 6.7, quantificam-se as complexidades estruturais destes dois filtros e também de $\mathcal{F}_2$, para comparação. Pelos dados fornecidos, vê-se que o filtro $\mathcal{D}_1$ possui menos da metade dos parâmetros de $\mathcal{F}_2$, ainda que seu desempenho tenha se mostrado tão bom quanto o deste último em termos de precisão. Logo, seu uso é totalmente justificável. Já no que se refere a $\mathcal{D}_2$, caso a aproximação por ele fornecida seja suficiente para os requisitos da aplicação de interesse, então seu uso é ainda mais vantajoso no que se refere à compacidade. Observe que ele possui, no total, menos de 1/6 do número de coeficientes de $\mathcal{F}_2$.

Ordem do núcleo (p)	Número de coeficientes		
	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{F}_2$
1	80	80	80
2	2.775	1.965	3.240
3	60.350	28.210	88.560
4	955.825	288.920	1.837.620
Total	1.019.030	319.175	1.929.500

Tabela 6.7: Número de coeficientes dos núcleos do FV convencional $\mathcal{F}_2$ e dos filtros diagonais aproximados $\mathcal{D}_1$ e $\mathcal{D}_2$.

Analogamente ao que foi feito para os filtros de Volterra convencionais, calculou-se a média de tempo gasto no cômputo das aproximações fornecidas pelos filtros diagonais aproximados. Os resultados obtidos são dados na Tabela 6.8. A economia computacional obtida é evidente. No caso de quarta ordem, por exemplo, o uso de $\mathcal{D}_1$ requer um tempo de computação quase quatro vezes menor que o de $\mathcal{F}_2$, enquanto que o uso de $\mathcal{D}_2$ requer um tempo de computação quase dezesseis vezes menor que o do FV convencional.

Se considerarmos os requisitos citados como exemplo na Subseção 6.3.1, de precisão mínima de 97% e de potência máxima da entrada de $\sigma_0 = 3$ então, dentre as três estruturas consideradas na análise acima, $\mathcal{D}_2$ é a mais apropriada.

Filtro	Tempo médio de cálculo (seg)			
	$p = 1$	$p = 2$	$p = 3$	$p = 4$
$\mathcal{D}_1$	$5,65 \times 10^{-4}$	$7,58 \times 10^{-3}$	$2,00 \times 10^{-1}$	3,55
$\mathcal{D}_2$	$7,60 \times 10^{-5}$	$4,62 \times 10^{-3}$	$7,55 \times 10^{-2}$	$8,43 \times 10^{-1}$
$\mathcal{F}_2$	$1,12 \times 10^{-4}$	$1,18 \times 10^{-2}$	$4,84 \times 10^{-1}$	$1,34 \times 10^1$

Tabela 6.8: Tempo médio de cálculo das aproximações de ordem p fornecidas pelo FV convencional $\mathcal{F}_2$ e pelos filtros diagonais aproximados $\mathcal{D}_1$ e $\mathcal{D}_2$.

6.4.2 Filtro de Volterra esparsos com interpolação

Observando-se os núcleos de primeira e segunda ordem mostrados nas Figuras 6.8 e 6.9, respectivamente, vê-se que eles variam suavemente. Em outras palavras, a correlação entre coeficientes vizinhos dos núcleos é relativamente alta. Sob tais condições, o uso de um esquema esparsos com interpolação para a implementação do filtro de Volterra pode conduzir a bons resultados, como discutido na Seção 4.3.

Nesta subseção, analogamente ao que foi feito em relação aos filtros diagonais aproximados, apresentaremos uma avaliação do desempenho de filtros de Volterra esparsos com interpolação quando empregados para a aproximação do modelo físico (6.28)–(6.29).

Para a avaliação experimental, foram consideradas duas estruturas esparsas com interpolação, ambas construídas a partir dos núcleos do filtro $\mathcal{F}_2$. Na primeira delas, que denotaremos por $\mathcal{S}_1$, empregou-se um fator de interpolação $L_1 = 3$ e um filtro interpolador cuja resposta impulsiva é representada pelo vetor

$$\mathbf{g}_1 = \frac{1}{3}[1 \ 2 \ 3 \ 2 \ 1]^T. \quad (6.42)$$

Já para a segunda, $\mathcal{S}_2$, definiu-se o fator de interpolação $L_2 = 4$ e o filtro interpolador usado foi

$$\mathbf{g}_2 = \frac{1}{4}[1 \ 2 \ 3 \ 4 \ 3 \ 2 \ 1]^T. \quad (6.43)$$

Para possibilitar equidade na comparação com as demais estruturas, adotou-se a implementação parcialmente interpolada que é proposta em [28]. Basicamente, trata-se de uma variante da estratégia apresentada na Seção 4.3 em que o sub-sistema de primeira ordem que compõe a estrutura não é aproximado, ou seja,

sua resposta impulsiva é empregada na totalidade para o cômputo da saída. Isto foi feito com o intuito de prevenir uma degradação muito grande da aproximação em relação aos resultados das demais EVAs, visto que o componente linear não é aproximado nelas e trata-se de uma porção que impacta significativamente a precisão das estimativas. Por fim, deve-se notar que o acréscimo de complexidade estrutural causado por esta escolha não é grande, visto que a contribuição do subsistema de primeira ordem para o total de coeficientes da estrutura é muito menor do que aquelas dadas pelos de mais alta ordem.

6.4.2.1 Avaliação das aproximações obtidas

Recorrendo ao mesmo procedimento empregado anteriormente, computaram-se experimentalmente os índices de ajuste médios atingidos pelas aproximações de ordens $p \in \{1, 2, 3, 4\}$ oferecidas por tais filtros. Nas Figuras 6.19 e 6.20 mostram-se, respectivamente, os índices relativos a $\mathcal{S}_1$ e $\mathcal{S}_2$, juntamente com os respectivos índices de $\mathcal{F}_2$, para comparação.

Figura 6.19: Índices de ajuste das aproximações de ordem p dadas por $\mathcal{F}_2$ (—) e por $\mathcal{S}_1$ (---).

Com base nos dados obtidos, observou-se que as aproximações dadas pelos filtros com interpolação são inferiores para baixas potências, porém este comportamento vai gradualmente se invertendo com o aumento da potência da entrada, à exceção da aproximação de quarta ordem (e também da de primeira ordem, evi-

Figura 6.20: Índices de ajuste das aproximações de ordem p dadas por $\mathcal{F}_2$ (—) e por $\mathcal{S}_2$ (---).

dentemente, por não haver aproximação, devido ao uso da estrutura parcialmente interpolada). O fenômeno observado decorre basicamente de erros introduzidos pela discretização dos núcleos, os quais, no caso do filtro com interpolação, são suavizados devido à característica passa-baixas dos filtros interpoladores.

De fato, quando se aumenta a taxa de amostragem usada nos experimentos, verificou-se que há uma melhora consistente em todos os índices das aproximações fornecidas pelo FV convencional, ao passo que as aproximações fornecidas pelo FV interpolado pioram justamente na região em que são superiores no experimento original. Em particular, o aumento da taxa causa uma redução do limiar de potência a partir do qual o FV interpolado passa a oferecer uma aproximação melhor.

Deste modo, não se pode concluir, a partir dos gráficos dados, que $\mathcal{S}_1$ oferece uma aproximação melhor do que $\mathcal{F}_2$, ainda que seu desempenho tenha sido superior no experimento realizado para entradas a partir de um certo nível de potência. Na realidade, trata-se apenas de um efeito observado que é causado por imprecisões envolvidas na discretização dos núcleos. Por motivos que ainda não foram totalmente esclarecidos, tal efeito é mais importante para potências de entrada mais altas do que para potências próximas de zero. Uma investigação aprofundada sobre este aspecto e sobre como contorná-lo é um dos tópicos que se

pretende abordar no prosseguimento deste estudo.

Mesmo diante das dificuldades trazidas para a análise causadas pelo que foi discutido acima, pode-se ter uma ideia da perda de precisão envolvida na aproximação esparsa com interpolação observando-se os índices de ajuste para as baixas potências (e, no caso de quarta ordem, para todas as potências). A partir de tal observação, percebe-se que a perda evidenciada em ambos os gráficos é baixa, sendo ligeiramente maior no caso de $\mathcal{S}_2$. Em particular, ambas as estruturas obtidas também possibilitam atender aos requisitos utilizados como exemplo anteriormente, de precisão mínima de 97% e de potência máxima da entrada de $\sigma_0 = 3$. Confirma-se, portanto, a hipótese de que, devido à característica passa-baixas do sistema de interesse, o esquema esparsa com interpolação produz boas aproximações. Na próxima subseção, discutiremos a redução de complexidade alcançada.

6.4.2.2 Avaliação da redução de complexidades estrutural e computacional

Assim como foi feito para os filtros diagonais aproximados, é interessante compararmos quantitativamente o ganho em compacidade trazido com a implementação esparsa com interpolação, bem como a redução do esforço computacional resultante. Na Tabela 6.9, mostram-se os números de coeficientes que compõem os núcleos de $\mathcal{S}_1$, de $\mathcal{S}_2$ e também de $\mathcal{F}_2$, para comparação. Tais números foram computados de acordo com as fórmulas dadas no Capítulo 4.

Nota-se a partir dos dados que a redução da complexidade estrutural obtida para os núcleos de alta ordem é imensa. Em relação ao total de coeficientes dos filtros, tem-se que o uso de $\mathcal{S}_1$ requer um número de coeficientes um pouco maior do que 1,6% do número que compõe $\mathcal{F}_2$. Logo, $\mathcal{S}_1$ traz um bom compromisso entre precisão e compacidade, visto que as diferenças entre as aproximações por ele fornecidas e aquelas dadas pelo FV convencional são pequenas, como visto na subseção anterior. Já $\mathcal{S}_2$ requer pouco mais de 0,5% dos total de parâmetros do FV convencional $\mathcal{F}_2$, possibilitando uma redução ainda mais significativa, a qual será bastante vantajosa caso o desempenho mostrado na Figura 6.19 seja

adequado para a aplicação de interesse.

Ordem do núcleo (p)	Número de coeficientes		
	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{F}_2$
1	80	80	80
2	378	210	3.240
3	3.654	1.540	88.560
4	27.405	8.855	1.837.620
Total	31.517	10.685	1.929.500

Tabela 6.9: Número de coeficientes dos núcleos do FV convencional $\mathcal{F}_2$ e dos filtros esparsos com interpolação $\mathcal{S}_1$ e $\mathcal{S}_2$.

Na Tabela 6.10, mostram-se os tempos médios consumidos para o cálculo das saídas dos filtros de Volterra esparsos com interpolação, de acordo com a ordem da aproximação computada. Para comparação, exibem-se também os respectivos tempos gastos no uso de $\mathcal{F}_2$. Observe que, apesar das estruturas $\mathcal{S}_1$ e $\mathcal{S}_2$ serem bastante compactas, o tempo computacional associado ao uso delas não foi muito reduzido em relação a $\mathcal{F}_2$. Em particular, não houve redução importante no tempo de computação dos componentes de segunda e terceira ordem, o que se deve ao *overhead* adicional trazido pela implementação utilizada. Já para a aproximação de quarta ordem, vê-se que o descarte de coeficientes traz uma economia que supera significativamente tal *overhead*, de modo que os tempos associados a $\mathcal{S}_1$ e $\mathcal{S}_2$ foram uma ordem de grandeza inferiores aos de $\mathcal{F}_2$. Mesmo nesse caso, nota-se que tal economia computacional não acompanha a redução de complexidade estrutural: o tempo de computação gasto por cada estrutura interpolada é de pouco menos da metade daquele gasto por $\mathcal{F}_2$, ao passo que os números de coeficientes de tais estruturas são mais de 50 vezes menores do que o de $\mathcal{F}_2$.

Filtro	Tempo médio de cálculo (seg)			
	$p = 1$	$p = 2$	$p = 3$	$p = 4$
$\mathcal{S}_1$	$3,50 \times 10^{-4}$	$1,26 \times 10^{-2}$	$2,82 \times 10^{-1}$	5,60
$\mathcal{S}_2$	$6,91 \times 10^{-5}$	$1,10 \times 10^{-2}$	$2,37 \times 10^{-1}$	4,94
$\mathcal{F}_2$	$1,12 \times 10^{-4}$	$1,18 \times 10^{-2}$	$4,84 \times 10^{-1}$	$1,34 \times 10^1$

Tabela 6.10: Tempo médio de cálculo das aproximações de ordem p fornecidas pelo FV convencional $\mathcal{F}_2$ e pelos filtros esparsos com interpolação $\mathcal{S}_1$ e $\mathcal{S}_2$.

6.4.3 Modelo PARAFAC-Volterra

Na Subseção 6.2.3, viu-se que a melhor aproximação de posto unitário (no sentido da norma de Frobenius) da matriz que representa o núcleo de segunda ordem apresenta um erro quadrático médio normalizado significativo. Então, seria interessante verificarmos se, aumentando-se um pouco o posto da aproximação, tal erro pode ser reduzido consideravelmente.

De forma relacionada a esta questão, avaliaremos nesta subseção se uma estrutura do tipo PARAFAC-Volterra, na qual os núcleos são aproximados por tensores de posto simétrico reduzido,³ pode conduzir a um bom compromisso entre precisão e compacidade. No que segue, isto será feito experimentalmente, à semelhança do que foi feito para as estruturas anteriormente avaliadas.

Com base nos tensores correspondentes aos núcleos de referência de tempo discreto que compõem $\mathcal{F}_2$, foram construídos dois modelos PARAFAC-Volterra, $\mathcal{P}_1$ e $\mathcal{P}_2$. Na definição de $\mathcal{P}_1$, os postos simétricos escolhidos para as aproximações dos núcleos de segunda, terceira e quarta ordem foram, respectivamente, $R_2 = 5$, $R_3 = 12$ e $R_4 = 15$. Já para $\mathcal{P}_2$, definiu-se $R_2 = 3$, $R_3 = 3$ e $R_4 = 3$. Tais escolhas foram feitas por tentativa e erro, buscando-se dois exemplos que oferecessem compromissos distintos entre compacidade e precisão.

Para o núcleo de segunda ordem, de acordo com o conteúdo da Subseção 4.4.1, os vetores que compõem um modelo PARAFAC podem ser calculados através da decomposição por autovalores da matriz que representa sua forma simétrica, $\mathbf{H}_2$. Como $\mathbf{H}_2$ é também simétrica, podemos inicialmente computar sua decomposição por valores singulares e posteriormente escrevê-la como

$$\mathbf{H}_2 = \sum_{m=1}^{80} v_m \sigma_m \mathbf{u}_m \mathbf{u}_m^T \quad (6.44)$$

³É importante salientar que, ao contrário do caso de segunda ordem, as aproximações para os núcleos de ordem três e quatro aqui empregadas não são garantidamente ótimas no sentido da norma de Frobenius.

onde $\mathbf{u}_m$ é o m -ésimo vetor singular esquerdo, σ_m é seu valor singular associado e

$$v_m = \begin{cases} 1, & \mathbf{u}_m = \mathbf{v}_m \\ -1, & \mathbf{u}_m = -\mathbf{v}_m \end{cases} \quad (6.45)$$

sendo $\mathbf{v}_m$ o m -ésimo vetor singular direito da decomposição. Note que, devido à simetria da matriz, os vetores singulares são tais que necessariamente uma das condições de (6.45) é verdadeira. Adicionalmente, pode-se mostrar que (6.44) é uma decomposição de $\mathbf{H}_2$ por autovalores dados por $\lambda_m = v_m \sigma_m$. Sendo assim, utilizando-se a ideia descrita na Subseção 4.4.1, uma decomposição PARAFAC do núcleo de segunda ordem de posto R_2 é obtida escrevendo-se

$$\mathbf{H}_2 \approx \sum_{m=1}^{R_2} \mathbf{a}_m \mathbf{a}_m^T \quad (6.46)$$

onde $\mathbf{a}_m = \sqrt{v_m \sigma_m} \mathbf{u}_m \in \mathbb{C}^{80}$.

Note que, devido ao uso da decomposição por valores singulares para a obtenção dos fatores paralelos de (6.46), tem-se que, se o somatório desta expressão for reduzido para incluir somente os R_2 primeiros termos, então a aproximação de posto R_2 resultante é ótima no sentido da norma de Frobenius. Assim, empregou-se tal estratégia para se obter os fatores paralelos de segunda ordem.

Em relação aos núcleos de terceira e quarta ordem, como visto na Subseção 4.4.1.1, o problema de determinação dos fatores paralelos não é tão bem resolvido como ocorre no caso de segunda ordem. Em particular, a decomposição por valores singulares de alta ordem em geral resulta em uma forma na qual o tensor núcleo $\mathcal{S}$ não é diagonal (ver Subseções A.1.2 e A.1.3 do Apêndice A).

Em princípio, sabe-se que, dada a simetria das representações tensoriais das formas simétricas dos núcleos de terceira e quarta ordem, $\mathcal{H}_3$ e $\mathcal{H}_4$, pode-se

decompô-las em modelos PARAFAC, i.e., nas formas

$$\mathcal{H}_3 = \sum_{r=1}^{S_3} \mathbf{v}_r^{(3)} \circ \mathbf{v}_r^{(3)} \circ \mathbf{v}_r^{(3)} \quad (6.47)$$

$$\mathcal{H}_4 = \sum_{r=1}^{S_4} \mathbf{v}_r^{(4)} \circ \mathbf{v}_r^{(4)} \circ \mathbf{v}_r^{(4)} \circ \mathbf{v}_r^{(4)} \quad (6.48)$$

em que S_3 é o posto simétrico de $\mathcal{H}_3$ e S_4 é o posto simétrico de $\mathcal{H}_4$. Porém, na prática, tais postos simétricos são, em geral, desconhecidos. Além disso, não é conhecido também um algoritmo que forneça tais decomposições para quaisquer tensores.

Assim, como alternativa para o cômputo dos fatores paralelos de ordens três e quatro, empregou-se uma implementação do algoritmo de mínimos quadrados alternados (ALS), o qual foi discutido na Subseção 4.4.1.1, oferecida pelo pacote para MATLAB *N-way* [90]. Como visto, tal algoritmo se aplica a tensores quaisquer (i.e., não-necessariamente simétricos), de modo que no caso geral resulta em uma decomposição na forma

$$\mathcal{T} \approx \sum_{r=1}^{R_T} \mathbf{v}_{1,r} \circ \mathbf{v}_{2,r} \circ \dots \circ \mathbf{v}_{p,r} \quad (6.49)$$

em que $\mathcal{T}$ é um tensor de ordem p e R_T é o posto escolhido para a decomposição. Contudo, aplicando-se o pacote *N-way* para decompor os tensores $\mathcal{H}_3$ e $\mathcal{H}_4$ com os valores de R_T dados por R_3 e R_4 , respectivamente, observou-se que as formas obtidas são tais que cada termo de (6.49) é composto por produtos externos que envolvem vetores aproximadamente iguais. Em outras palavras, obtiveram-se decomposições nas formas

$$\mathcal{H}_3 \approx \sum_{r=1}^{R_3} \mathbf{a}_r^{(3)} \circ \mathbf{a}_r^{(3)} \circ \mathbf{a}_r^{(3)} \quad (6.50)$$

$$\mathcal{H}_4 \approx \sum_{r=1}^{R_4} \mathbf{a}_r^{(4)} \circ \mathbf{a}_r^{(4)} \circ \mathbf{a}_r^{(4)} \circ \mathbf{a}_r^{(4)} \quad (6.51)$$

as quais podem ser empregadas na construção de um modelo PARAFAC.

Para termos uma ideia da qualidade das aproximações dos núcleos obtidas, é

interessante que utilizemos como índice quantitativo o EQMN envolvido em cada uma delas. Em relação a $\mathcal{P}_1$, tem-se que

$$\frac{\|\mathcal{H}_2^* - \mathcal{H}_2\|_F^2}{\|\mathcal{H}_2\|_F^2} = 0,0025 \quad (6.52)$$

$$\frac{\|\mathcal{H}_3^* - \mathcal{H}_3\|_F^2}{\|\mathcal{H}_3\|_F^2} = 0,0013 \quad (6.53)$$

$$\frac{\|\mathcal{H}_4^* - \mathcal{H}_4\|_F^2}{\|\mathcal{H}_4\|_F^2} = 0,0025 \quad (6.54)$$

em que $\mathcal{H}_p^*$ é a aproximação do núcleo de ordem p relativa a $\mathcal{P}_1$. Analogamente, para $\mathcal{P}_2$ verificou-se que

$$\frac{\|\mathcal{H}_2^* - \mathcal{H}_2\|_F^2}{\|\mathcal{H}_2\|_F^2} = 0,0121 \quad (6.55)$$

$$\frac{\|\mathcal{H}_3^* - \mathcal{H}_3\|_F^2}{\|\mathcal{H}_3\|_F^2} = 0,0253 \quad (6.56)$$

$$\frac{\|\mathcal{H}_4^* - \mathcal{H}_4\|_F^2}{\|\mathcal{H}_4\|_F^2} = 0,0159. \quad (6.57)$$

Pelos valores mostrados acima, nota-se que as aproximações tensoriais obtidas são, de fato, bem próximas aos núcleos a que se referem. A seguir, o desempenho das estruturas PARAFAC resultantes serão avaliados experimentalmente.

6.4.3.1 Avaliação das aproximações obtidas

Os modelos PARAFAC-Volterra $\mathcal{P}_1$ e $\mathcal{P}_2$ construídos foram avaliados de forma análoga às estruturas consideradas anteriormente, i.e., empregando-se um procedimento como aquele descrito na Subseção 6.3.1. Com isso, obtiveram-se os resultados que são mostrados nas Figuras 6.21 e 6.22, os quais se referem, respectivamente, a $\mathcal{P}_1$ e $\mathcal{P}_2$. Para comparação, mostram-se também os respectivos índices de ajuste obtidos com o FV convencional $\mathcal{F}_2$.

De acordo com as curvas exibidas, vê-se que mesmo com o número reduzido de vetores utilizados na estrutura $\mathcal{P}_1$, as aproximações por ela fornecidas são muito próximas àsquelas dadas por $\mathcal{F}_2$, sendo portanto muito vantajoso utilizá-la. Já nas aproximações de ordens três e quatro produzidas por $\mathcal{P}_2$, percebe-se que há uma

Figura 6.21: Índices de ajuste das aproximações de ordem p dadas por $\mathcal{F}_2$ (—) e por $\mathcal{P}_1$ (---).

Figura 6.22: Índices de ajuste das aproximações de ordem p dadas por $\mathcal{F}_2$ (—) e por $\mathcal{P}_2$ (---).

degradação mais significativa, especialmente na curva de quarta ordem. Contudo, trata-se de uma estrutura ainda mais compacta, sendo que, caso o desempenho apresentado pela aproximação de terceira ordem seja suficiente, então tem-se um compromisso excelente entre precisão e compacidade.

Um aspecto que se nota nos gráficos dados é que, assim como ocorreu no caso dos filtros esparsos com interpolação, algumas curvas de ajuste associadas às estruturas PARAFAC são tais que seus valores superam aqueles das curvas relativas a $\mathcal{F}_2$ a partir de um determinado patamar de potência da entrada. Novamente,

trata-se de um efeito causado pelos erros de discretização contidos nos núcleos. Por definição, $\mathcal{P}_1$ e $\mathcal{P}_2$ são estruturas que aproximam os núcleos por tensores de posto reduzido, os quais estão, portanto, menos sujeitos às influências de tais erros, uma vez que no cálculo dos fatores paralelos buscam-se aqueles que explicam apenas variações essenciais dos núcleos, e não pequenas oscilações causadas por imprecisões. Por um motivo que ainda não foi totalmente esclarecido, os efeitos observados são mais importantes para entradas de alta potência. Deste modo, para tais entradas o erro envolvido na aproximação por um tensor de posto reduzido é inferior àquele causado pelos motivos mencionados. Note que isto sugere que, de fato, a decomposição PARAFAC conduziu a aproximações coerentes dos núcleos do ponto de vista físico, possibilitando inclusive uma redução da importância relativa dos erros de discretização mencionados em comparação com os núcleos de referência.

De modo geral, pode-se concluir que modelos do tipo PARAFAC-Volterra são bastante eficientes na aproximação do comportamento do modelo físico considerado. Na subseção a seguir, descreveremos a redução de complexidade obtida com o uso de $\mathcal{P}_1$ e $\mathcal{P}_2$.

6.4.3.2 Avaliação da redução de complexidades estrutural e computacional

Na Tabela 6.11, quantificam-se os totais de coeficientes envolvidos nas estruturas $\mathcal{P}_1$ e $\mathcal{P}_2$ e também no FV $\mathcal{F}_2$, para comparação. Os dados corroboram a afirmação feita no final da subseção anterior, mostrando que as estruturas PARAFAC são extremamente compactas quando comparadas ao FV convencional de referência. Note que $\mathcal{P}_1$ possui apenas 0,13% dos coeficientes de $\mathcal{F}_2$, enquanto o total de coeficientes de $\mathcal{P}_2$ corresponde a pouco mais de 0,042% do total que compõe $\mathcal{F}_2$.

Retomando novamente os requisitos dados como exemplo na Subseção 6.3.1, de precisão mínima de 97% e de potência máxima da entrada de $\sigma_0 = 3$, vê-se que com o uso de $\mathcal{P}_2$, é possível atendê-los utilizando apenas 560 coeficientes, os quais são necessários para computar a aproximação de terceira ordem.

Ordem do núcleo (p)	Número de coeficientes		
	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{F}_2$
1	80	80	80
2	400	240	3.240
3	960	240	88.560
4	1.200	240	1.837.620
Total	2.640	800	1.929.500

Tabela 6.11: Número de coeficientes dos núcleos do FV convencional $\mathcal{F}_2$ e das estruturas PARAFAC-Volterra $\mathcal{P}_1$ e $\mathcal{P}_2$.

Como feito nas avaliações anteriores, mediu-se também o tempo médio envolvido no cômputo das aproximações de ordem p dadas pelos dos modelos PARAFAC-Volterra empregados. Os resultados são dispostos na Tabela 6.12. A partir deles, conclui-se que as estruturas PARAFAC são, de fato, muito menos custosas computacionalmente para uso do que o FV convencional $\mathcal{F}_2$. Ainda, como esperado, esta diferença se acentua com a ordem. Em particular, para $p = 4$, tem-se uma redução do tempo de computação de mesma ordem de grandeza que a redução de complexidade paramétrica obtida, ao contrário do que se observou, por exemplo, em relação aos FVs esparsos com interpolação.

Filtro	Tempo médio de cálculo (seg)			
	$p = 1$	$p = 2$	$p = 3$	$p = 4$
$\mathcal{P}_1$	$9,33 \times 10^{-4}$	$6,55 \times 10^{-3}$	$1,71 \times 10^{-2}$	$3,21 \times 10^{-2}$
$\mathcal{P}_2$	$1,47 \times 10^{-4}$	$1,84 \times 10^{-3}$	$5,21 \times 10^{-3}$	$1,02 \times 10^{-2}$
$\mathcal{F}_2$	$1,12 \times 10^{-4}$	$1,18 \times 10^{-2}$	$4,84 \times 10^{-1}$	$1,34 \times 10^1$

Tabela 6.12: Tempo médio de cálculo das aproximações de ordem p fornecidas pelo FV convencional $\mathcal{F}_2$ e pelas estruturas PARAFAC-Volterra $\mathcal{P}_1$ e $\mathcal{P}_2$.

6.4.4 Modelo de modos dinâmicos principais

A última estrutura avaliada experimentalmente é o modelo de modos dinâmicos principais, o qual foi apresentado na Subseção 4.4.2.

Para tanto, a partir dos núcleos de referência construíram-se duas estruturas MDP, $\mathcal{M}_1$ e $\mathcal{M}_2$. O número de modos principais utilizados para compor os núcleos de segunda, terceira e quarta ordens de $\mathcal{M}_1$ foram definidos como, respectivamente, $M_2 = 5$, $M_3 = 12$ e $M_4 = 15$. No tocante à estrutura $\mathcal{M}_2$, adotou-se

$M_2 = 3$, $M_3 = 3$ e $M_4 = 3$. Note que tais números são iguais aos de fatores paralelos utilizados nos modelos PARAFAC-Volterra avaliados, tendo sido escolhidos desta forma para que se possa comparar também os desempenhos de tais estruturas de Wiener.

Conforme discutido na Subseção 4.4.2.1, uma forma de se obter filtros que podem ser empregados como modos dinâmicos principais é através da decomposição por valores singulares, para o núcleo de segunda ordem, e da decomposição por valores singulares de alta ordem, para os núcleos de mais alta ordem. Neste estudo de caso, seguiu-se tal ideia, sendo que as decomposições de alta ordem foram calculadas com base nos princípios discutidos na Subseção A.1.2 do Apêndice A.

Para a determinação dos modos que representam o núcleo de segunda ordem, procedeu-se de forma idêntica ao que foi feito para o modelo PARAFAC, obtendo-se, assim, M_2 vetores que compõem uma aproximação da matriz $\mathbf{H}_2$. Em seguida, determinaram-se os tensores $\mathcal{S}_3$ e $\mathcal{S}_4$ e as matrizes $\mathbf{U}_3$ e $\mathbf{U}_4$ que permitem escrever os elementos de $\mathcal{H}_3$ e $\mathcal{H}_4$ como

$$[\mathcal{H}_3]_{i_1, i_2, i_3} = \sum_{r_1=1}^{80} \sum_{r_2=1}^{80} \sum_{r_3=1}^{80} [\mathcal{S}_3]_{r_1, r_2, r_3} [\mathbf{U}_3]_{i_1, r_1} [\mathbf{U}_3]_{i_2, r_2} [\mathbf{U}_3]_{i_3, r_3} \quad (6.58)$$

e

$$[\mathcal{H}_4]_{i_1, i_2, i_3, i_4} = \sum_{r_1=1}^{80} \sum_{r_2=1}^{80} \sum_{r_3=1}^{80} \sum_{r_4=1}^{80} [\mathcal{S}_4]_{r_1, r_2, r_3, r_4} [\mathbf{U}_4]_{i_1, r_1} [\mathbf{U}_4]_{i_2, r_2} [\mathbf{U}_4]_{i_3, r_3} [\mathbf{U}_4]_{i_4, r_4} . \quad (6.59)$$

Assim, os modos dinâmicos principais que compõem a aproximação do núcleo de ordem três são dados pelas M_3 primeiras colunas de $\mathbf{U}_3$ e aqueles que formam a aproximação do de quarta ordem correspondem às M_4 primeiras colunas de $\mathbf{U}_4$.

Note que, para aproximar os núcleos de ordem $p > 2$, é necessário também incorporar nos modelos os coeficientes dos tensores núcleo $\mathcal{S}_3$ e $\mathcal{S}_4$ que combinam os elementos dos modos principais nas expressões (6.58) e (6.59). Como discutido na Subseção 4.4.2.1, tais coeficientes compõem a não-linearidade sem memória polinomial do modelo MDP, cujas entradas são as saídas dos filtros dados pelos modos dinâmicos.

À semelhança do que foi feito para a estrutura PARAFAC-Volterra, carac-

terizou-se a qualidade das aproximações fornecidas pelos modelos MDP para os núcleos de referência através do cálculo do EQMN de cada aproximação. Para $\mathcal{M}_1$, obteve-se

$$\frac{\|\mathcal{H}_2^* - \mathcal{H}_2\|_F}{\|\mathcal{H}_2\|_F} = 0,0025 \quad (6.60)$$

$$\frac{\|\mathcal{H}_3^* - \mathcal{H}_3\|_F}{\|\mathcal{H}_3\|_F} = 0,0008 \quad (6.61)$$

$$\frac{\|\mathcal{H}_4^* - \mathcal{H}_4\|_F}{\|\mathcal{H}_4\|_F} = 0,0003 \quad (6.62)$$

e, para $\mathcal{M}_2$,

$$\frac{\|\mathcal{H}_2^* - \mathcal{H}_2\|_F}{\|\mathcal{H}_2\|_F} = 0,0121 \quad (6.63)$$

$$\frac{\|\mathcal{H}_3^* - \mathcal{H}_3\|_F}{\|\mathcal{H}_3\|_F} = 0,0253 \quad (6.64)$$

$$\frac{\|\mathcal{H}_4^* - \mathcal{H}_4\|_F}{\|\mathcal{H}_4\|_F} = 0,0162. \quad (6.65)$$

Novamente, vê-se que os erros envolvidos são pequenos, especialmente para as aproximações associadas a $\mathcal{M}_1$. Em particular, os erros envolvidos nas aproximações dos núcleos de terceira e quarta ordem são menores do que aqueles associados ao modelo PARAFAC $\mathcal{P}_1$, o que se deve ao fato de que, mesmo tendo sido utilizados o mesmo número de vetores, no modelo MDP o produto externo de cada combinação possível dos vetores contribui para a aproximação, ao passo que no somatório da decomposição PARAFAC há apenas um termo para cada vetor utilizado (ver expressões (6.50) e (6.51)).

Fazendo-se a mesma comparação para $\mathcal{M}_2$ e $\mathcal{P}_2$, vê-se que, neste caso, as aproximações dos núcleos associadas ao modelo PARAFAC-Volterra são tão boas quanto as do modelo MDP ou, no caso de quarta ordem, até melhores. Isto mostra que, com o número reduzido de vetores utilizado para ambos estes modelos, os fatores paralelos encontrados com o uso do algoritmo do Toolbox N-way são mais eficientes do que os modos dinâmicos obtidos algebricamente através da HOSVD.

No que segue, os resultados da avaliação feita com as estruturas MDP será apresentado e, após isso, serão discutidos aspectos relativos à complexidade com-

putacional de uso dos modelos.

6.4.4.1 Avaliação das aproximações obtidas

Novamente, por meio de um procedimento como aquele descrito na Subseção 6.3.1, avaliaram-se os modelos de modos dinâmicos principais $\mathcal{M}_1$ e $\mathcal{M}_2$. Os resultados desta avaliação são exibidos nas Figuras 6.23 e 6.24, referindo-se, respectivamente, a $\mathcal{M}_1$ e $\mathcal{M}_2$. Como feito para as outras estruturas, exibem-se também os respectivos índices de ajuste associados às aproximações dadas pelo FV convencional $\mathcal{F}_2$.

Figura 6.23: Índices de ajuste das aproximações de ordem p dadas por $\mathcal{F}_2$ (—) e por $\mathcal{M}_1$ (---).

Com base nos gráficos, nota-se que o desempenho das estruturas MDP se assemelha muito àquele demonstrado pelas estruturas PARAFAC. Isto é, ainda que as aproximações dos tensores de terceira e quarta ordem sejam obtidas de maneiras diferentes para os dois modelos, os resultados apontam que elas são comparáveis em termos de precisão.

Novamente, observa-se a superioridade de algumas das curvas de ajuste obtidas com $\mathcal{M}_1$ e $\mathcal{M}_2$ em relação às obtidas com $\mathcal{F}_2$, a partir de um certo nível de potência da entrada. De fato, a mesma discussão feita na Subseção 6.4.3.1 em relação a este aspecto se aplica também no caso das estruturas MDP.

Figura 6.24: Índices de ajuste das aproximações de ordem p dadas por $\mathcal{F}_2$ (—) e por $\mathcal{M}_2$ (---).

6.4.4.2 Avaliação da redução de complexidades estrutural e computacional

A Tabela 6.13 contém uma comparação entre os números totais de coeficientes envolvidos nas estruturas $\mathcal{M}_1$ e $\mathcal{M}_2$ e no FV convencional $\mathcal{F}_2$. De acordo com ela, percebe-se que, assim como no caso das estruturas PARAFAC, a grande compacidade de $\mathcal{M}_1$ e $\mathcal{M}_2$ é evidente: $\mathcal{M}_1$ possui apenas 0,31% dos coeficientes de $\mathcal{F}_2$, ao passo que $\mathcal{M}_2$ contém um total de coeficientes que corresponde a pouco mais de 0,043% do total de $\mathcal{F}_2$.

Contudo, devido ao fato de que a não-linearidade polinomial sem memória do modelo MDP é tal que há produtos cruzados entre suas entradas, eles normalmente são menos compactos que modelos PARAFAC com o mesmo número de filtros na parte linear. Isto pode ser constatado comparando-se os dados da Tabela 6.13 com aqueles que foram mostrados na Tabela 6.11.

Por fim, o tempo médio consumido no uso dos modelos MDP para se obter cada aproximação de ordem p é dado na Tabela 6.14. Observando-se tais dados, conclui-se as estruturas de modos dinâmicos principais possibilitam redução significativa deste tempo em relação a um FV convencional, especialmente quando aproximações de alta ordem são desejadas. Em comparação com os modelos PARAFAC, modelos MDP constituídos pelos mesmos números de filtros tendem a

Ordem do núcleo (p)	Número de coeficientes		
	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{F}_2$
1	80	80	80
2	400	240	3.240
3	1.324	250	88.560
4	4.260	255	1.837.620
Total	6.064	825	1.929.500

Tabela 6.13: Número de coeficientes dos núcleos do FV convencional $\mathcal{F}_2$ e das estruturas MDP $\mathcal{M}_1$ e $\mathcal{M}_2$.

ser um pouco mais custosos computacionalmente, devido à necessidade do cálculo de produtos cruzados entre as saídas dos filtros na não-linearidade polinomial. Na estrutura PARAFAC, não se incluem, por definição, tais produtos cruzados.

Filtro	Tempo médio de cálculo (seg)			
	$p = 1$	$p = 2$	$p = 3$	$p = 4$
$\mathcal{P}_1$	$6,55 \times 10^{-4}$	$2,17 \times 10^{-3}$	$4,96 \times 10^{-2}$	$4,56 \times 10^{-1}$
$\mathcal{P}_2$	$1,75 \times 10^{-4}$	$7,02 \times 10^{-4}$	$2,99 \times 10^{-3}$	$7,60 \times 10^{-3}$
$\mathcal{F}_2$	$1,12 \times 10^{-4}$	$1,18 \times 10^{-2}$	$4,84 \times 10^{-1}$	$1,34 \times 10^1$

Tabela 6.14: Tempo médio de cálculo das aproximações de ordem p fornecidas pelo FV convencional $\mathcal{F}_2$ e pelas estruturas MDP $\mathcal{M}_1$ e $\mathcal{M}_2$.

6.5 Discussão dos resultados

De maneira geral, os resultados dispostos nas seções anteriores mostram que

- (i) em um exemplo concreto envolvendo um sistema não-linear (alto-falante), o uso de um FV convencional para uma aproximação de alta ordem da saída do modelo físico (6.26)–(6.29) deste sistema implica em um custo computacional elevado, devido à grande complexidade estrutural deste filtro;
- (ii) a dinâmica do modelo físico é suscetível a uma aproximação precisa através das estruturas de Volterra alternativas consideradas;
- (iii) o uso destas EVAs possibilita um bom desempenho com um número reduzido de coeficientes em comparação com o FV convencional.

Nas Figuras 6.25 e 6.26 e na Tabela 6.15, resumam-se os resultados obtidos. Os gráficos mostram, de forma comparativa, o desempenho de cada tipo de estrutura testada para cada ordem de aproximação. Para a geração deles, empregaram-se os modelos de cada tipo de estrutura que apresentaram melhor desempenho: $\mathcal{F}_2$, $\mathcal{D}_1$, $\mathcal{S}_1$, $\mathcal{P}_1$ e $\mathcal{M}_1$. Na tabela, os números totais de coeficientes que compõem estes modelos são comparados.

Figura 6.25: Comparação entre os desempenhos das estruturas avaliadas nas aproximações de primeira (a) e de segunda (b) ordens: FV convencional (◆), FV diagonal aproximado (*), FV esparsa com interpolação (●), estrutura PARAFAC-Volterra (■) e estrutura MDP (▲).

As curvas mostradas são bastante próximas, em muitos casos sendo impossível distingui-las visualmente. Já a comparação entre os totais coeficientes dada na tabela evidencia a superioridade das abordagens baseadas em aproximação ten-

Figura 6.26: Comparação entre os desempenhos das estruturas avaliadas nas aproximações de terceira (a) e de quarta (b) ordens: FV convencional (◆), FV diagonal aproximado (*), FV esparsa com interpolação (●), estrutura PARAFAC-Volterra (■) e estrutura MDP (▲).

sorial para aproximação do modelo físico do alto-falante. Assim, em suma, vê-se que os modelos baseados na estrutura de Wiener foram aqueles que possibilitaram o melhor compromisso entre precisão e compacidade, possuindo um número extremamente reduzido de parâmetros.

No que se refere aos objetivos buscados na realização do estudo de caso, em linhas gerais, mostrou-se que o uso dos núcleos de referência é útil em diversos aspectos. Primeiro, ele possibilita ter uma ideia do comportamento de aproximações de diversas ordens dadas pelas estruturas estudadas. Segundo, com o uso deles é possível definir de forma apropriada diversas características importantes

Ordem do núcleo (p)	Número de coeficientes				
	$\mathcal{F}_2$	$\mathcal{D}_1$	$\mathcal{S}_1$	$\mathcal{P}_1$	$\mathcal{M}_1$
1	80	80	80	80	80
2	3.240	2.775	378	400	400
3	88.560	60.350	3.654	960	1.324
4	1.837.620	955.825	27.464	1.200	4.260
Total	1.929.500	1.019.030	31.517	2.640	6.064

Tabela 6.15: Comparação entre os números de coeficientes que compõem os melhores modelos das estruturas avaliadas.

das estruturas estudadas:

- (i) a extensão de memória considerada para todos os filtros;
- (ii) as distâncias máximas admitidas para as diagonais dos núcleos que compõem os filtros diagonais aproximados;
- (iii) o fator de interpolação empregado em um FV esparsos com interpolação;
- (iv) o número de fatores paralelos empregados para aproximar cada núcleo em um modelo PARAFAC-Volterra;
- (v) o número de modos dinâmicos principais que compõem a aproximação de cada núcleo em um modelo MDP.

Terceiro, pode-se ter uma ideia quantitativa do custo computacional exigido por cada estrutura para um modelamento com determinado nível de precisão, dado um valor máximo de potência desejado para as entradas. Quarto, as estruturas consideradas podem ser avaliadas comparativamente para se decidir por uma delas.

Apesar do potencial da abordagem para os propósitos listados acima, durante a descrição dos resultados ficou claro que ainda é necessário investigar certos aspectos envolvidos na obtenção dos núcleos e na realização dos experimentos, dados os efeitos inesperados que foram observados. Mais especificamente, viu-se que as estruturas aproximadas foram superiores ao FV convencional em alguns casos e a partir de um determinado nível de potência. Como discutido, trata-se de um efeito causado por erros de discretização dos núcleos, além de existir uma certa influência dos erros introduzidos pelo uso de filtros de tempo discreto na

simulação com o intuito de aproximar os filtros de tempo contínuo incorporados nos núcleos.

Logicamente, tais erros de discretização não estão ligados a características físicas do dispositivo em questão. Assim, é interessante notar que o efeito mencionado no parágrafo anterior evidencia que as aproximações adotadas pelas estruturas de Volterra alternativas estudadas são razoáveis do ponto de vista físico, visto que a importância relativa de tais erros é menor nos núcleos que as compõem do que nos núcleos de referência originais. Isto pode ser constatado pelo fato de que, a partir de uma determinada potência, as aproximações das estruturas alternativas permitem reduzir a influência dos erros de discretização nos resultados obtidos.

Em relação às medições de tempo computacional apresentadas, verificou-se uma certa variabilidade dos resultados na prática. Adicionalmente, alguns aspectos envolvidos nas implementações feitas exercem também influência sobre os resultados. Por tais motivos, os números apresentados são válidos para comparação apenas em termos de ordem de grandeza.

Para finalizar, é fundamental também que se esclareça que as estimativas dos números de coeficientes necessários para que cada EVA atinja um certo nível de precisão obtidas com a metodologia proposta constituem condição suficiente (supondo, evidentemente, que o modelo físico representa com precisão o sistema de interesse), porém não necessária. Como a construção das EVAs empregadas nas simulações foi feita diretamente a partir dos núcleos de referência obtidos, não tendo sido levadas em conta características estatísticas das entradas típicas consideradas, na prática é possível que um processo de estimação ou adaptação dos parâmetros conduza a uma estrutura mais compacta que apresente desempenho semelhante diante das estatísticas das entradas admitidas – ou, ainda, que tal processo resulte em um modelo com complexidade estrutural semelhante capaz de aproximar com maior precisão as saídas do sistema de interesse.

Um comentário análogo é válido também para o próprio FV convencional composto pelos núcleos de referência. De fato, do ponto de vista teórico, o Teorema 3.5 garante que, para qualquer conjunto de entradas limitadas considerado (as

quais devem possuir crescimento também limitado), um sistema com memória desvanescente sempre pode ser aproximado com o nível de precisão desejado por uma série de Volterra truncada, não existindo uma garantia de que o filtro de Volterra convencional composto pelos núcleos de referência corresponda a uma versão de tempo discreto e com núcleos truncados desta série. Entretanto, trata-se de uma estrutura baseada em considerações físicas; logo, espera-se que seu desempenho seja razoável como referência.

Capítulo 7

Conclusões e perspectivas

Neste trabalho, estudaram-se os problemas da escolha de um formalismo matemático e da definição de uma estrutura específica para a representação de um sistema físico não-linear no contexto de aplicações de processamento digital de sinais. Idealmente, os modelos empregados nesse contexto devem oferecer aproximações precisas sem exigir recursos computacionais em demasia para uso. Logo, trata-se de decisões determinantes para a utilidade prática do modelo resultante, uma vez que elas impactam diretamente sua capacidade de representação e sua complexidade estrutural.

Apesar da importância destes problemas, com frequência, eles não são tratados sistematicamente, e sim de forma *ad hoc*, com base em uma argumentação informal. Adicionalmente, o comportamento da aproximação fornecida por um determinado modelo em função da potência das entradas, o que é um aspecto fundamental quando se trata de sistemas não-lineares, usualmente não é discutido.

A contribuição central deste trabalho é a proposta de uma metodologia de avaliação da adequação de formalismos não-lineares ao modelamento de um sistema de interesse com base em características físicas de tal sistema. Para tanto, é necessário que um modelo físico seja conhecido *a priori*, o que diversas vezes ocorre em aplicações de engenharia. Com base neste modelo e na definição de valores de parâmetros típicos, obtém-se um conjunto de núcleos de Volterra de referência, os quais proveem informações úteis na escolha de um formalismo matemático mais conveniente para uso prático e adaptação. Além disso, pode-se explorar tais núcleos para a definição de características estruturais do modelo, como sua extensão

de memória e sua ordem de não-linearidade máxima, e também para a realização de uma avaliação quantitativa do desempenho de estruturas deriváveis a partir dos núcleos de referência.

Para a implementação desta estratégia, investigaram-se quatro métodos de cálculo de núcleos de Volterra a partir de equações de estado na forma linear-analítica. Embora sejam interessantes do ponto de vista teórico, três deles não se mostraram apropriados quando se tem interesse no cômputo numérico de núcleos associados a um sistema linear-analítico qualquer. Já o quarto método estudado, chamado de bilinearização de Carleman, pode ser implementado convenientemente com o uso de um algoritmo proposto neste trabalho, o qual é significativamente mais eficiente do que uma implementação direta. Além disso, demonstrou-se que com o uso dele é possível obter expressões analíticas exatas para os núcleos de Volterra, ou, alternativamente, para as funções de resposta em frequência generalizadas a eles associadas.

A possibilidade de cômputo dos núcleos no domínio da frequência se mostrou interessante para os fins estabelecidos, uma vez que permite incorporar facilmente características de filtros tipicamente presentes em aplicações de processamento digital de sinais e que são de particular importância quando se lida com sistemas não-lineares. Assim, com base no método de Carleman e nas funções de resposta em frequência que definem tais filtros, propôs-se uma forma de obtenção dos núcleos de referência de tempo discreto adequada ao contexto das aplicações consideradas. Mostrou-se também que, mesmo na presença de fatores ambientais como a reverberação acústica, os núcleos de referência obtidos permanecem válidos. Se desejado, pode-se ainda levar tais fatores em conta na derivação dos núcleos, o que permite estudar a influência que eles exercem nas características do modelo resultante.

Realizou-se uma revisão bibliográfica crítica de diversos formalismos e estruturas utilizadas em trabalhos de processamento digital de sinais, focando a capacidade de representação de cada uma delas e também aspectos que determinam a viabilidade de seu uso prático, como estabilidade e eficiência estrutural. Dentre as estruturas estudadas, incluem-se estruturas de Volterra alternativas que foram

propostas recentemente na literatura com o intuito de oferecer um compromisso favorável entre precisão e compacidade, visto que filtros de Volterra convencionais não costumam ser compactos. Contudo, cada uma delas envolve alguma estratégia de aproximação dos núcleos de Volterra, sendo aplicável somente quando eles possuem determinadas características, as quais foram apontadas no texto. Diante disto, vê-se que o conhecimento de núcleos de referência pode ser de grande auxílio na escolha de uma estrutura para o modelo.

Com a finalidade de demonstrar o potencial da metodologia proposta, conduziu-se um estudo de caso em que núcleos de referência foram calculados a partir de um modelo físico para alto-falantes disponível na literatura. Com o uso de simulações Monte Carlo, avaliou-se o desempenho de algumas estruturas na aproximação da saída do modelo físico. As realizações consideradas para a sequência de entrada foram geradas com características espectrais especificadas à semelhança de uma situação prática. Para a quantificação do desempenho de cada estrutura considerada, variou-se a potência das entradas dentro de uma faixa de interesse estabelecida e computou-se a média dos índices de ajuste obtidos para cada valor de potência. Adicionalmente, calculou-se também o número de parâmetros que a compõem e mediu-se o tempo de computação médio associado ao seu uso.

Os resultados obtidos mostram que, se um modelamento preciso é desejado sob a presença de entradas com alta potência, o uso de um filtro de Volterra convencional composto pelos núcleos de referência implica em um custo computacional elevado, podendo ser até inviável para implementação em tempo real. Já com o uso de estruturas de Volterra alternativas, demonstrou-se que é possível obter modelos muito mais compactos, com uma perda de precisão pouco significativa. Em particular, as estruturas baseadas no modelo de Wiener tiveram um desempenho excelente na aproximação do modelo físico necessitando, para tanto, de um número extremamente reduzido de coeficientes em comparação com um FV convencional.

7.1 Perspectivas de trabalhos futuros

Em primeiro lugar, como discutido no Capítulo 6, é importante investigar ainda certos aspectos envolvidos na metodologia proposta, especialmente no que se refere à discretização dos núcleos e à avaliação experimental das estruturas obtidas a partir deles. Mais especificamente, há certas imprecisões envolvidas nesta discretização e na simulação da estrutura que envolve o modelo físico, as quais causam efeitos inesperados nos resultados da avaliação, especialmente para entradas de alta potência. Uma melhor compreensão de tais efeitos e sobre como seria possível contorná-los é fundamental para consolidar a metodologia proposta.

Adicionalmente, de modo geral, acreditamos que um aprofundamento do estudo de formas de se explorar informações físicas previamente conhecidas sobre um sistema de interesse pode conduzir a estratégias que auxiliem significativamente na obtenção de modelos precisos e eficientes. Desta forma, o estudo realizado neste trabalho constitui apenas um primeiro passo de uma investigação pautada por uma linha de pesquisa que parece pouco explorada na literatura.

Dentre os aspectos que podem ser considerados neste aprofundamento, inclui-se a avaliação do desempenho de estruturas não-lineares em um contexto não-estacionário, levando-se em conta suas propriedades no que concerne à adaptação dos parâmetros. Outro ponto que se deseja estudar se refere a como se pode extrair informações dos núcleos de referência visando a construção de estruturas ótimas (no sentido de eficiência estrutural) para a aproximação de um sistema, admitindo-se que as estatísticas das entradas que ocorrem na prática são conhecidas.

Além disso, seria interessante também estudar extensões das técnicas desenvolvidas para que elas se tornem mais abrangentes. Por exemplo, a forma admitida para os modelos físicos poderia ser generalizada para além da classe linear-analítica. Ainda, o uso dos núcleos de referência na avaliação de outros formalismos importantes, como o modelo NARX, poderia ser investigada.

Por fim, a realização de estudos de caso envolvendo sistemas reais também seria importante, uma vez que permitiria validar as conclusões obtidas com o uso

da metodologia proposta.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] MATHEWS, V. J.; SICURANZA, G. L. *Polynomial Signal Processing*. [S.l.]: Wiley-Interscience, 2000. ISBN 0471034142.
- [2] AZPICUETA-RUIZ, L. et al. Adaptive Combination of Volterra Kernels and its Application to Nonlinear Acoustic Echo Cancellation. *IEEE Transactions on Audio, Speech and Language Processing*, IEEE, v. 19, n. 1, p. 97–110, jan. 2011.
- [3] CRESPO-CADENAS, C. et al. A new approach to pruning volterra models for power amplifiers. *IEEE Transactions on Signal Processing*, IEEE, v. 58, n. 4, p. 2113–2120, abr. 2010.
- [4] GUERIN, A. et al. Nonlinear acoustic echo cancellation based on Volterra filters. *IEEE Transactions on Speech and Audio Processing*, IEEE, v. 11, n. 6, p. 672–683, nov. 2003. ISSN 1063-6676.
- [5] BEIDAS, B. Intermodulation Distortion in Multicarrier Satellite Systems: Analysis and Turbo Volterra Equalization. *IEEE Transactions on Communications*, IEEE, v. 59, n. 6, p. 1580–1590, jun. 2011.
- [6] MORGAN, D. et al. A Generalized Memory Polynomial Model for Digital Predistortion of RF Power Amplifiers. *IEEE Transactions on Signal Processing*, IEEE, v. 54, n. 10, p. 3852–3860, out. 2006. ISSN 1053-587X.
- [7] KUNTZE, S. B. et al. Nonlinear State–Space Model of Semiconductor Optical Amplifiers With Gain Compression for System Design and Analysis. *Journal of Lightwave Technology*, OSA, v. 26, n. 14, p. 2274–2281, jul. 2008.
- [8] SATO, C.; OHTANI, T.; NISHITANI, H. Modeling, simulation and nonlinear control of a gas-phase polymerization process. *Computers and Chemical Engineering*, Elsevier, v. 24, n. 2, p. 945–951, jul. 2000.
- [9] MARMARELIS, V. *Nonlinear Dynamic Modeling of Physiological Systems*. Hoboken, NJ, USA: John Wiley & Sons, 2004. ISBN 0471469602.
- [10] SJÖBERG, J. et al. Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, Elsevier, v. 31, n. 12, p. 1691–1724, dez. 1995.
- [11] SCHURER, H. *Linearization of electroacoustic transducers*. Tese (Doutorado) — Twente University, 1997.
- [12] FRANK, W. A. An efficient approximation to the quadratic Volterra filter and its application in real-time loudspeaker linearization. *Signal Processing*, Elsevier, v. 45, n. 1, p. 97–113, jul. 1995.

- [13] GAO, F.; SNELGROVE, W. Adaptive linearization of a loudspeaker. In: *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. [S.l.]: IEEE, 1991. p. 3589–3592. ISBN 0780300033. ISSN 1520-6149.
- [14] TAN, L.; JIANG, J. Adaptive Volterra filters for active control of nonlinear noise processes. *IEEE Transactions on Signal Processing*, IEEE, v. 49, n. 8, p. 1667–1676, 2001. ISSN 1053587X.
- [15] YEH, D. T.; ABEL, J. S.; SMITH, J. O. Automated Physical Modeling of Nonlinear Audio Circuits For Real-Time Audio Effects - Part I: Theoretical Development. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE, v. 18, n. 4, p. 728–737, maio 2010. ISSN 1558-7916.
- [16] YEH, D. T. et al. Numerical Methods for Simulation of Guitar Distortion Circuits. *Computer Music Journal*, MIT Press, v. 32, n. 2, p. 23–42, jun. 2008. ISSN 0148-9267.
- [17] PAKARINEN, J.; KARJALAINEN, M. Enhanced Wave Digital Triode Model for Real-Time Tube Amplifier Emulation. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE, v. 18, n. 4, p. 738–746, maio 2010. ISSN 1558-7916.
- [18] PAKARINEN, J.; YEH, D. T. A Review of Digital Techniques for Modeling Vacuum-Tube Guitar Amplifiers. *Computer Music Journal*, MIT Press, v. 33, n. 2, p. 85–100, maio 2009. ISSN 0148-9267.
- [19] KARJALAINEN, M.; PAKARINEN, J. Wave Digital Simulation of a Vacuum-Tube Amplifier. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, v. 5, p. 153–156, 2006.
- [20] DANG, Y.; HAN, C.; FANG, Y. Robust nonlinear MPC based on simplified second-order Volterra series model. In: *3rd World Congress on Intelligent Control and Automation*. [S.l.]: IEEE, 2000. v. 5, p. 3269–3273.
- [21] GANTT, J.; GATZKE, E. Control of a granulation process using a nonlinear MPC formulation. In: *American Control Conference*. [S.l.]: IEEE, 2004. v. 5, p. 4231–4236. ISSN 0743-1619.
- [22] NOVAK, A. et al. A new method for identification of nonlinear systems using miso model with swept-sine technique: Application to loudspeaker analysis. In: *Audio Engineering Society Convention 124*. [S.l.: s.n.], 2008.
- [23] NOVAK, A. et al. Nonlinear system identification using exponential swept-sine signal. *IEEE Transactions on Instrumentation and Measurement*, IEEE, v. 59, n. 8, p. 2220–2229, ago. 2010. ISSN 0018-9456.
- [24] NOVAK, A.; SIMON, L.; KADLEC, F. Modeling of Nonlinear Audio Systems using Swept-Sine signals: Application to audio effects. In: *12th International Conference on Digital Audio Effects (DAFx)*. [S.l.: s.n.], 2009.

- [25] STENGER, A.; TRAUTMANN, L.; RABENSTEIN, R. Nonlinear acoustic echo cancellation with 2nd order adaptive Volterra filters. In: *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. [S.l.]: IEEE, 1999. v. 2, p. 877–880. ISBN 0-7803-5041-3.
- [26] SEO, J. B.; KIM, K. J.; NAM, S. W. Nonlinear Acoustic Echo Cancellation Using Volterra Filtering with a Variable Step-Size GS-PAP Algorithm. In: *World Academy of Science, Engineering and Technology*. [S.l.: s.n.], 2009. p. 59–62.
- [27] ZELLER, M.; KELLERMANN, W. Fast and Robust Adaptation of DFT-Domain Volterra Filters in Diagonal Coordinates Using Iterated Coefficient Updates. *IEEE Transactions on Signal Processing*, IEEE, v. 58, n. 3, p. 1589–1604, mar. 2010.
- [28] BATISTA, E.; TOBIAS, O.; SEARA, R. A sparse-interpolated scheme for implementing adaptive volterra filters. *IEEE Transactions on Signal Processing*, IEEE, v. 58, n. 4, p. 2022–2035, abr. 2010.
- [29] FAVIER, G.; KIBANGOU, A.; BOUILLOC, T. Nonlinear system modeling and identification using Volterra-PARAFAC models. *International Journal of Adaptive Control and Signal Processing*, John Wiley & Sons, v. 26, n. 1, p. 30–53, 2011.
- [30] FRANKEN, D.; MEERKOTTER, K.; MUTH, J. W. Passive parametric modeling of dynamic loudspeakers. *IEEE Transactions on Speech and Audio Processing*, IEEE, v. 9, n. 8, p. 885–891, nov. 2002. ISSN 1063-6676.
- [31] ISIDORI, A. *Nonlinear Control Systems*. 3rd. ed. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1995. ISBN 3540199160.
- [32] RUGH, W. J. *Nonlinear System Theory: The Volterra/Wiener Approach*. Baltimore, MD: Johns Hopkins University Press, 1981.
- [33] GOULART, J. H. M.; BURT, P. M. S. Efficient kernel computation for Volterra filter structure evaluation. *IEEE Signal Processing Letters*, v. 19, n. 3, p. 135–138, mar. 2012. ISSN 1070-9908.
- [34] DUNIPACE, K. R.; LYSHEVSKI, S. E.; COLGREN, R. D. State-space nonlinear identification and tracking control of aircraft: a motion control problem. In: *IEEE Conference on Decision and Control*. [S.l.: s.n.], 1998. p. 2277–2282.
- [35] FRANKEN, D.; MEERKOTTER, K.; WASSMUTH, J. Observer-based feedback linearization of dynamic loudspeakers with Ac amplifiers. *IEEE Transactions on Speech and Audio Processing*, v. 13, n. 2, p. 233–242, mar. 2005. ISSN 1063-6676.
- [36] BROCKETT, R. Volterra series and geometric control theory. *Automatica*, Elsevier, v. 12, n. 2, p. 167–176, mar. 1976.

- [37] SCHETZEN, M. *The Volterra and Wiener Theories of Nonlinear Systems*. [S.l.]: John Wiley & Sons, 1980. ISBN 0471044555.
- [38] AGUIRRE, L. A. *Introdução à Identificação de Sistemas*. 3a. ed. Belo Horizonte, MG: Editora UFMG, 2007. ISBN 978-85-7041-584-4.
- [39] LJUNG, L. *System identification: theory for the user*. 2nd. ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1999. ISBN 0136566952.
- [40] LJUNG, L. Perspectives on system identification. *Annual Reviews in Control*, v. 34, n. 1, p. 1–12, abr. 2010. ISSN 13675788.
- [41] CHEN, S.; BILLINGS, S. A.; LUO, W. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, Taylor & Francis, v. 50, n. 5, p. 1873–1896, nov. 1989.
- [42] LEONTARITIS, I. J.; BILLINGS, S. A. Model selection and validation methods for non-linear systems. *International Journal of Control*, Taylor & Francis, v. 45, n. 1, p. 311–341, jan. 1987. ISSN 0020-7179.
- [43] BENDAT, J. S. *Nonlinear System Analysis and Identification from Random Data*. [S.l.]: Wiley-Interscience, 1990. ISBN 0471606235.
- [44] DING, F.; LIU, X. P.; LIU, G. Identification methods for Hammerstein non-linear systems. *Digital Signal Processing*, Elsevier, v. 21, p. 215–238, mar. 2011. ISSN 10512004.
- [45] HAYKIN, S. *Adaptive Filter Theory*. 4th. ed. Upper Saddle River, NJ, USA: Prentice Hall, 2001. ISBN 0130901261.
- [46] PINTELON, R.; SCHOUKENS, J. *System Identification: A Frequency Domain Approach*. 2nd. ed. NJ, USA: Wiley-IEEE Press, 2012. ISBN 0470640375.
- [47] HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. 2nd. ed. Upper Saddle River, NJ, USA: Prentice Hall, 1998. ISBN 0132733501.
- [48] GILBERT, E. Functional expansions for the response of nonlinear differential systems. *IEEE Transactions on Automatic Control*, IEEE, v. 22, n. 6, p. 909–921, jun. 1977. ISSN 0018-9286.
- [49] BOYD, S.; CHUA, L.; DESOER, C. Analytical foundations of Volterra series. *IMA Journal of Mathematical Control and Information*, IMA, v. 1, n. 3, p. 243–282, 1984.
- [50] HELIE, T.; LAROCHE, B. Computation of Convergence Bounds for Volterra Series of Linear-Analytic Single-Input Systems. *IEEE Transactions on Automatic Control*, IEEE, v. 56, n. 9, p. 2062–2072, set. 2011. ISSN 0018-9286.
- [51] SONTAG, E. D. Input Classes for Identifiability of Bilinear Systems. *IEEE Transactions on Automatic Control*, IEEE, v. 54, n. 2, p. 195–207, fev. 2009.

- [52] CHUA, L.; NG, C.-Y. Frequency domain analysis of nonlinear systems: general theory. *IEE Journal on Electronic Circuits and Systems*, IET, v. 3, n. 4, p. 165–185, 1979. ISSN 03086984.
- [53] BILLINGS, S.; TSANG, K. Spectral Analysis for Nonlinear Systems, Part II: Interpretation of Nonlinear Frequency Response Functions. *Mechanical Systems and Signal Processing*, Elsevier, v. 3, n. 4, p. 341–359, 1989. ISSN 0888-3270.
- [54] HELIE, T.; LAROCHEY, B. Convergence radius and guaranteed error bound for the Volterra series expansion of finite dimensional quadratic systems. In: *16th Mediterranean Conference on Control and Automation*. [S.l.]: IEEE, 2008. p. 741–746. ISBN 978-1-4244-2504-4.
- [55] BOYD, S.; CHUA, L. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, CAS-32, n. 11, p. 1150–1160, 1985.
- [56] GALLMAN, P. G.; NARENDRA, K. Representations of nonlinear systems via the Stone-Weierstrass theorem. *Automatica*, v. 12, n. 6, p. 619–622, 1976. ISSN 0005-1098.
- [57] KOEPPL, H.; SCHWINGSHACKL, D. Comparison of discrete-time approximations for continuous-time nonlinear systems. In: *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. [S.l.]: IEEE, 2004. v. 2, p. 881–884.
- [58] OPPENHEIM, A. V.; SCHAFER, R. W.; BUCK, J. R. *Discrete-time signal processing*. 2nd. ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1999. ISBN 0137549202.
- [59] BILBAO, S. *Wave and Scattering Methods for Numerical Simulation*. Chichester, UK: John Wiley & Sons, 2004. ISBN 0470870176.
- [60] CELLIER, F. E.; KOFMAN, E. *Continuous System Simulation*. New York, NY, USA: Springer, 2006. ISBN 144193863X.
- [61] CHEN, S.; BILLINGS, S. A. Representation of non-linear systems: the NARMAX model. *International Journal of Control*, Taylor & Francis, v. 49, n. 3, p. 1012–1032, 1989.
- [62] JING, X. J.; LANG, Z. Q.; BILLINGS, S. A. Magnitude bounds of generalized frequency response functions for nonlinear Volterra systems described by NARX model. *Automatica*, v. 44, n. 3, p. 838–845, 2008. ISSN 0005-1098.
- [63] BILLINGS, S. A.; TSANG, K. M. Spectral analysis for non-linear systems, Part I: Parametric non-linear spectral analysis. *Mechanical Systems and Signal Processing*, Elsevier, v. 3, n. 4, p. 319–339, 1989.
- [64] GÓMEZ, J. Identification of block-oriented nonlinear systems using orthonormal bases. *Journal of Process Control*, v. 14, n. 6, p. 685–697, set. 2004.

- [65] BIAGIOLA, S.; FIGUEROA, J. Wiener and Hammerstein uncertain models identification. *Mathematics and Computers in Simulation*, v. 79, n. 11, p. 3296–3313, jul. 2009.
- [66] ESKINAT, E.; JOHNSON, S. H.; LUYBEN, W. L. Use of Hammerstein models in identification of nonlinear systems. *AIChE Journal*, v. 37, n. 2, p. 255–268, fev. 1991. ISSN 0001-1541.
- [67] HABER, R.; UNBEHAUEN, H. Structure identification of nonlinear dynamic systems-a survey on input/output approaches. *Automatica*, Elsevier, v. 26, n. 4, p. 651–677, jul. 1990. ISSN 00051098.
- [68] BAI, E.; LI, K. Convergence of the iterative algorithm for a general hammerstein system identification. *Automatica*, Elsevier, v. 46, n. 11, p. 1891–1896, nov. 2010.
- [69] NARENDRA, K.; GALLMAN, P. An iterative method for the identification of nonlinear systems using a Hammerstein model. *IEEE Transactions on Automatic Control*, IEEE, v. 11, n. 3, p. 546–550, jul. 1966. ISSN 0018-9286.
- [70] GALLMAN, P. An iterative method for the identification of nonlinear systems using a Uryson model. *IEEE Transactions on Automatic Control*, IEEE, v. 20, n. 6, p. 771–775, 1975.
- [71] KUECH, F.; MITNACHT, A.; KELLERMANN, W. Nonlinear Acoustic Echo Cancellation Using Adaptive Orthogonalized Power Filters. In: *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. [S.l.]: IEEE, 2005. v. 3, p. 105–108. ISBN 0-7803-8874-7.
- [72] BAI, E. Frequency domain identification of Wiener models. *Automatica*, Elsevier, v. 39, n. 9, p. 1521–1530, set. 2003.
- [73] BILLINGS, S.; FAKHOURI, S. Identification of systems containing linear dynamic and static nonlinear elements. *Automatica*, Elsevier, v. 18, n. 1, p. 15–26, jan. 1982.
- [74] NOLLETT, B.; JONES, D. Nonlinear echo cancellation for hands-free speakerphones. In: *IEEE Workshop on Nonlinear Signal and Image Processing (NSIP)*. [S.l.]: IEEE, 1997.
- [75] VANDERSTEEN, G. Non-parametric Estimation of the Frequency-response Functions of the Linear Blocks of a Wiener-Hammerstein Model. *Automatica*, Elsevier, v. 33, n. 7, p. 1351–1355, jul. 1997.
- [76] ZELLER, M.; KELLERMANN, W. Coefficient pruning for higher-order diagonals of Volterra filters representing Wiener-Hammerstein models. In: *International Workshop on Acoustics Echo and Noise Control (IWAENC)*. [S.l.: s.n.], 2008.

- [77] SEAGRAVES, E.; WALCOTT, B.; FEINAUER, D. Efficient implementation of Volterra systems using a multilinear SVD. In: *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*. [S.l.]: IEEE, 2007. p. 762–765.
- [78] DUDGEON, D. E.; MERSEREAU, R. M. *Multidimensional Digital Signal Processing*. NJ, USA: Prentice Hall, 1984. ISBN 0136049591.
- [79] MORHAC, M. A fast algorithm of nonlinear Volterra filtering. *IEEE Transactions on Signal Processing*, IEEE, v. 39, n. 10, p. 2353–2356, out. 1991.
- [80] BATISTA, E. L. O.; SEARA, R. A fully LMS/NLMS adaptive scheme applied to sparse-interpolated Volterra filters with removed boundary effect. *Signal Processing*, Elsevier, v. 92, n. 10, p. 2381–2393, out. 2012. ISSN 01651684.
- [81] COMON, P. et al. Symmetric tensors and symmetric tensor rank. *SIAM Journal on Matrix Analysis and Applications*, SIAM, v. 30, n. 3, p. 1254–1279, 2008.
- [82] BADEAU, R.; BOYER, R. Fast Multilinear singular value decomposition for structured tensors. *SIAM Journal on Matrix Analysis and Applications*, v. 30, n. 3, p. 1008–1021, set. 2008.
- [83] BRACHAT, J. et al. Symmetric tensor decomposition. *Linear Algebra and its Applications*, v. 433, n. 11-12, p. 1851–1872, dez. 2010. ISSN 00243795.
- [84] COMON, P.; LUCIANI, X.; ALMEIDA, A. L. F. de. Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics*, John Wiley & Sons, v. 23, p. 393–405, ago. 2009. ISSN 1099-128X.
- [85] KOLDA, T. G.; BADER, B. W. Tensor decompositions and applications. *SIAM Review*, SIAM, v. 51, n. 3, p. 455–500, 2009.
- [86] BREWER, J. Kronecker products and matrix calculus in system theory. *IEEE Transactions on Circuits and Systems*, v. 25, n. 9, p. 772–781, set. 1978. ISSN 0098-4094.
- [87] GOULART, J. H.; BURT, P. M. S. *Código em MATLAB para algoritmo eficiente de bilinearização de Carleman*. Última visita em 15 de agosto de 2012. Disponível em: <<http://www.lcs.poli.usp.br/~jgoulart/research.html>>.
- [88] SMALL, R. H. Closed-box loudspeaker systems-part 1: Analysis. *Journal of the Audio Engineering Society*, v. 20, n. 10, p. 798–808, dez. 1972.
- [89] SMALL, R. H. Closed-box loudspeaker systems-part 2: Synthesis. *Journal of the Audio Engineering Society*, v. 21, n. 1, p. 11–18, fev. 1973.
- [90] BRO, R. *The N-way Toolbox*. Última visita em 16 de junho de 2012. Disponível em: <www.mathworks.com/matlabcentral/fileexchange/1088-the-n-way-toolbox>.

- [91] LATHAUWER, L. D.; MOOR, B. D. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, v. 21, n. 4, p. 1253–1278, 2000.

Apêndice A

Revisão de conceitos matemáticos e notação

A.1 Álgebra de tensores

O objetivo desta seção é fornecer uma revisão sucinta de conceitos de álgebra de tensores que são empregados nesta dissertação, introduzindo também a notação adotada.

A.1.1 Notação e conceitos básicos

Tensores são arranjos multidimensionais (no sentido geométrico) de valores numéricos. Tal dimensão geométrica é também denominada ordem do tensor, o que evita ambiguidades com o conceito de dimensão algébrica.

Formalmente, um tensor de ordem p sobre um campo $\mathbb{F}$ é denotado por $\mathcal{H}$ e pertence ao espaço $\mathbb{F}^{n_1 \times n_2 \times \dots \times n_p}$, em que os números inteiros positivos $n_1, \dots, n_p$ formam sua dimensão (no sentido algébrico). Os elementos de $\mathcal{H}$ são denotados por $[\mathcal{H}]_{i_1, \dots, i_p}$, em que $i_j \in \{1, \dots, n_j\}$, com $j \in \{1, \dots, p\}$. No caso particular de vetores e matrizes, i.e., tensores de primeira e segunda ordem, quando é mais conveniente utiliza-se a convenção usual de denotá-los por letras latinas em negrito minúsculas e maiúsculas, respectivamente.

Se um espaço de tensores de ordem p , $\mathbb{F}^{n_1 \times n_2 \times \dots \times n_p}$, é tal que $n_1 = n_2 = \dots = n_p = n$, então ele também pode ser escrito de forma mais compacta como $\mathbb{F}^{n^{(p)}}$. Adicionalmente, os tensores nele contidos são ditos hipercúbicos [29].

Um tensor hipercúbico $\mathcal{H} \in \mathbb{F}^{n^{(p)}}$ é dito simétrico se satisfaz a propriedade

$$\forall i_1, \dots, i_p \in \{1, \dots, n\}, \quad [\mathcal{H}]_{i_1, \dots, i_p} = [\mathcal{H}]_{\pi_1(i), \dots, \pi_p(i)} \quad (A.1)$$

onde $\pi = (\pi_1(i), \dots, \pi_p(i))$ é uma permutação qualquer do multi-índice $i = (i_1, \dots, i_p)$.

A norma de Frobenius de um tensor $\mathcal{H} \in \mathbb{F}^{n_1 \times \dots \times n_p}$ é definida como

$$\|\mathcal{H}\|_F = \sqrt{\sum_{i_1=1}^{n_1} \dots \sum_{i_p=1}^{n_p} ([\mathcal{H}]_{i_1, \dots, i_p})^2} \quad (A.2)$$

sendo, portanto, uma generalização da definição da mesma norma para matrizes.

Para um tensor qualquer $\mathcal{H} \in \mathbb{F}^{n_1 \times \dots \times n_s \times \dots \times n_p}$, definem-se seus vetores do modo s como tensores de primeira ordem do espaço $\mathbb{F}^{n_s}$ cujos elementos são dados por

$$\left[\mathbf{h}_{i_1, \dots, i_{s-1}, i_{s+1}, \dots, i_p}^{(s)} \right]_{i_s} = [\mathcal{H}]_{i_1, \dots, i_{s-1}, i_s, i_{s+1}, \dots, i_p}. \quad (A.3)$$

Observe que existem $n_1 \dots n_{s-1} n_{s+1} \dots n_p$ vetores do modo s de $\mathcal{H}$.

O desdobramento matricial horizontal $\mathbf{H}_s$ de $\mathcal{H}$, também chamado de modo s [82], é uma matriz que pertence ao espaço $\mathbb{F}^{n_s \times n_{s+1} \dots n_p n_1 \dots n_{s-1}}$ e é formada justamente pelos vetores do modo s [29]. Em outras palavras, ela é tal que [91]

$$[\mathbf{H}_s]_{i_s, j_{s+1} + j_{s+2} + \dots + j_p + j_1 + \dots + j_{s-1}} = \mathcal{H}_{i_1, \dots, i_s, \dots, i_p} \quad (A.4)$$

em que

$$j_{s+1} = (i_{s+1} - 1)n_{s+2}n_{s+3} \dots n_p n_1 \dots n_{s-1} \quad (A.5)$$

$$j_{s+2} = (i_{s+2} - 1)n_{s+3}n_{s+4} \dots n_p n_1 \dots n_{s-1} \quad (A.6)$$

$$\vdots$$

$$j_p = (i_p - 1)n_1 \dots n_{s-1} \quad (A.7)$$

e

$$j_1 = (i_1 - 1)n_2 \dots n_{s-1} \quad (A.8)$$

$$\vdots$$

$$j_{s-2} = (i_{s-2} - 1)n_{s-1} \quad (A.9)$$

$$j_{s-1} = i_{s-1}. \quad (A.10)$$

Deve-se notar que, quando $s = 1$, então o último termo j_k que entra no somatório do segundo índice de $[\mathbf{H}_s]$ em (A.4) é $j_p = i_p$. Observe que, como comentado em [29], trata-se de um tensor que contém a mesma informação de $\mathcal{H}$, porém arranjada de outra maneira.

A.1.2 Produto e decomposição de Tucker

Sejam o tensor $\mathcal{H} \in \mathbb{F}^{n_1 \times \dots \times n_s \times \dots \times n_p}$ e a matriz $\mathbf{G} \in \mathbb{F}^{m \times n_s}$. Então, o produto de Tucker no modo s (também chamado simplesmente de produto no modo s) entre $\mathcal{H}$ e $\mathbf{G}$ é definido como

$$\times_s : \mathbb{F}^{n_1 \times \dots \times n_s \times \dots \times n_p} \times \mathbb{F}^{m \times n_s} \rightarrow \mathbb{F}^{n_1 \times \dots \times m \times \dots \times n_p} \quad (A.11)$$

e tal que

$$[\mathcal{H} \times_s \mathbf{G}]_{i_1, \dots, j, \dots, i_p} = \sum_{i_s=1}^{n_s} [\mathcal{H}]_{i_1, \dots, i_s, \dots, i_p} [\mathbf{G}]_{i_s, j} \quad (A.12)$$

em que $j \in \{1, \dots, m\}$ e $i_k \in \{1, \dots, n_k\}$ para $k \in \{1, \dots, p\}$.

Alternativamente, podemos definir o resultado de $\mathcal{V} = \mathcal{H} \times_s \mathbf{G}$ em termos dos modos s de $\mathcal{V}$ e de $\mathcal{H}$, de tal forma que [29]

$$\mathbf{V}_s = \mathbf{G} \mathbf{H}_s \quad (A.13)$$

em que $\mathbf{V}_s$ e $\mathbf{H}_s$ são os modos s de $\mathcal{V}$ e $\mathcal{H}$, respectivamente.

Com base na definição acima, uma decomposição (ou modelo) de Tucker de

$\mathcal{H}$ pode ser escrita como [29]

$$\mathcal{H} = \mathcal{C} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \dots \times_p \mathbf{U}^{(p)} \quad (A.14)$$

em que as matrizes $\mathbf{U}^{(j)} \in \mathbb{F}^{n_j \times R_j}$ são denominadas matrizes de fatores e $\mathcal{C} \in \mathbb{F}^{R_1 \times \dots \times R_p}$ é chamado de tensor núcleo. Uma outra forma de se ver tal decomposição é

$$[\mathcal{H}]_{i_1, \dots, i_p} = \sum_{r_1=1}^{R_1} \dots \sum_{r_p=1}^{R_p} [\mathcal{C}]_{r_1, \dots, r_p} \prod_{j=1}^p [\mathbf{U}^{(j)}]_{i_j, r_j}. \quad (A.15)$$

Uma decomposição de particular importância na forma (A.14) ocorre quando as colunas da matriz $\mathbf{U}^{(j)}$ correspondem aos vetores singulares esquerdos do modo j de $\mathcal{H}$, para $j \in \{1, \dots, p\}$. Neste caso, tem-se uma generalização da decomposição por valores singulares, a qual é chamada de decomposição por valores singulares de alta ordem (HOSVD, do inglês *high-order singular value decomposition*) [29, 82]. Os valores singulares de tal decomposição são definidos como [82]

$$\sigma_i^{(s)} = \|[\mathcal{C}]_{i_s=i}\|_F \quad (A.16)$$

em que s é o modo do valor singular e $[\mathcal{C}]_{i_s=i} \in \mathbb{F}^{n_1, \dots, n_{s-1}, n_{s+1}, \dots, n_p}$ denota o subtensor de $\mathcal{C}$ obtido fixando-se o índice i_s no valor $i \in \{1, \dots, n_s\}$. Por fim, deve-se notar que, quando $\mathcal{H}$ é simétrico, o tensor núcleo da HOSVD, $\mathcal{C}$, também o é. Sendo assim, os valores singulares de todos os modos são iguais, i.e.,

$$\sigma_i^{(1)} = \sigma_i^{(2)} = \dots = \sigma_i^{(p)} = \sigma_i. \quad (A.17)$$

Adicionalmente, para um tensor $\mathcal{H}$ simétrico, as matrizes $\mathbf{U}^{(k)}$ da decomposição (A.14) são todas iguais, i.e., $\mathbf{U}^{(1)} = \mathbf{U}^{(2)} = \dots = \mathbf{U}^{(p)} = \mathbf{U}$.

A.1.3 Produto externo

O produto externo entre dois tensores $\mathcal{H} \in \mathbb{F}^{n_1 \times \dots \times n_p}$ e $\mathcal{G} \in \mathbb{F}^{m_1 \times \dots \times m_q}$ é definido como

$$\circ : \mathbb{F}^{n_1 \times \dots \times n_p} \times \mathbb{F}^{m_1 \times \dots \times m_q} \rightarrow \mathbb{F}^{n_1 \times \dots \times n_p \times m_1 \times \dots \times m_q} \quad (A.18)$$

e é tal que, se

$$\mathcal{J} = \mathcal{H} \circ \mathcal{G} \quad (A.19)$$

então

$$[\mathcal{J}]_{i_1, \dots, i_{p+q}} = [\mathcal{H}]_{i_1, \dots, i_p} [\mathcal{G}]_{i_{p+1}, \dots, i_q}. \quad (A.20)$$

Em se tratando especificamente de tensores de primeira ordem, i.e., vetores $\mathbf{h}$ e $\mathbf{g}$, uma notação familiar que pode ser empregada para o produto externo é $\mathbf{J} = \mathbf{h}\mathbf{g}^T$ (note que o resultado é uma matriz), desde que ambos sejam vetores coluna, como usualmente se admite.

Da definição do produto externo, decorre que, se

$$\mathcal{J} = \mathbf{v}_1 \circ \mathbf{v}_2 \circ \dots \circ \mathbf{v}_p \quad (A.21)$$

em que $\mathbf{v}_i \in \mathbb{F}^{n_i}$, com $i \in \{1, \dots, p\}$, então $\mathcal{J} \in \mathbb{F}^{n_1 \times \dots \times n_p}$ é tal que

$$[\mathcal{J}]_{i_1, \dots, i_p} = \prod_{j=1}^p [\mathbf{v}_j]_{i_j}. \quad (A.22)$$

Analogamente à decomposição de Tucker, pode-se definir uma decomposição de um tensor $\mathcal{H} \in \mathbb{F}^{n_1 \times \dots \times n_p}$ na forma [29]

$$\mathcal{H} = \sum_{r=1}^R \mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \dots \circ \mathbf{a}_r^{(p)} \quad (A.23)$$

em que $\mathbf{a}_r^{(j)}$ é a r -ésima coluna da matriz de fatores $\mathbf{A}^{(j)} \in \mathbb{F}^{n_j \times R}$. Como observado em [29], a forma (A.23), denominada na literatura de decomposição PARAFAC, é um caso particular de (A.14) em que o tensor núcleo é a identidade $\mathcal{I}_R \in \mathbb{F}^{R^{(p)}}$,

i.e.,

$$[\mathcal{I}_R]_{i_1, \dots, i_p} = \begin{cases} 1, & i_1 = i_2 = \dots = i_p \\ 0, & \text{caso contrário} \end{cases} \quad (A.24)$$

e todas as matrizes de fatores $\mathbf{U}^{(j)}$ possuem o mesmo número de colunas, R .

Um resultado importante existente na literatura é que, quando $\mathcal{H} \in \mathbb{C}^{n^{(p)}}$ é um tensor simétrico, então ele admite uma decomposição PARAFAC tal que todas as matrizes de fatores são idênticas [29], ou seja,

$$\mathcal{H} = \sum_{r=1}^R \mathbf{a}_r \circ \mathbf{a}_r \circ \dots \circ \mathbf{a}_r \quad (A.25)$$

em que $\mathbf{a}_r$ é a r -ésima coluna da matriz de fatores $\mathbf{A}$ e cada termo do somatório envolve $p - 1$ produtos externos (logo, o vetor $\mathbf{a}_r$ é repetido p vezes em sua expressão).

A.1.4 Produto de Kronecker

Um operador de particular importância na análise de equações de estado não-lineares é o produto de Kronecker, o qual é definido para tensores de primeira e segunda ordens, i.e., vetores e matrizes, respectivamente. Em linhas gerais, trata-se de um operador que resulta em um vetor ou uma matriz cujos componentes são todos os produtos possíveis entre elementos de seus operandos, como ficará claro na descrição a seguir.

Sejam as matrizes

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,1} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{bmatrix} \in \mathbb{F}^{m \times n} \quad (A.26)$$

e

$$\mathbf{B} = \begin{bmatrix} b_{1,1} & b_{1,1} & \dots & b_{1,q} \\ b_{2,1} & b_{2,2} & \dots & b_{2,q} \\ \vdots & \vdots & \vdots & \vdots \\ b_{p,1} & b_{p,2} & \dots & b_{p,q} \end{bmatrix} \in \mathbb{F}^{p \times q}. \quad (A.27)$$

Então, o produto de Kronecker entre $\mathbf{A}$ e $\mathbf{B}$ é definido como [32, 86]

$$\otimes : \mathbb{F}^{m \times n} \times \mathbb{F}^{p \times q} \rightarrow \mathbb{F}^{mp \times nq} \quad (A.28)$$

e tal que

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{1,1}\mathbf{B} & a_{1,1}\mathbf{B} & \dots & a_{1,n}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \dots & a_{2,n}\mathbf{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m,1}\mathbf{B} & a_{m,2}\mathbf{B} & \dots & a_{m,n}\mathbf{B} \end{bmatrix}. \quad (A.29)$$

Observe que, considerando que vetores coluna podem ser representados como matrizes compostas por uma única coluna, então o produto de Kronecker é definido por extensão para vetores

$$\mathbf{v} = [v_1 \ v_2 \ \dots \ v_m]^T \in \mathbb{F}^m \quad (A.30)$$

$$\mathbf{w} = [w_1 \ w_2 \ \dots \ w_n]^T \in \mathbb{F}^n \quad (A.31)$$

$$(A.32)$$

resultando, neste caso, em outro vetor dado por

$$\mathbf{v} \otimes \mathbf{w} = \begin{bmatrix} v_1 \mathbf{w} \\ \vdots \\ v_m \mathbf{w} \end{bmatrix} \quad (A.33)$$

ou seja,

$$\mathbf{v} \otimes \mathbf{w} = \begin{bmatrix} v_1 w_1 \\ \vdots \\ v_1 w_n \\ v_2 w_1 \\ \vdots \\ v_2 w_n \\ \vdots \\ v_m w_1 \\ \vdots \\ v_m w_n \end{bmatrix}. \quad (A.34)$$

O produto de Kronecker entre dois vetores linha é definido analogamente, de modo a resultar também em um vetor linha.

Com base na definição acima, pode-se mostrar que as seguintes propriedades do produto de Kronecker são válidas [32, 86]:

- (i) $(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C})$
- (ii) $(\mathbf{AB}) \otimes (\mathbf{CD}) = (\mathbf{A} \otimes \mathbf{C})(\mathbf{B} \otimes \mathbf{D})$
- (iii) $(\mathbf{A} + \mathbf{B}) \otimes (\mathbf{C} + \mathbf{D}) = (\mathbf{A} \otimes \mathbf{C}) + (\mathbf{A} \otimes \mathbf{D}) + (\mathbf{B} \otimes \mathbf{C}) + (\mathbf{B} \otimes \mathbf{D})$.

A segunda propriedade da lista acima é conhecida como propriedade do “produto misto” [86].

A.1.5 Operador vec

Seja a matriz

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{bmatrix} \in \mathbb{F}^{m \times n}. \quad (A.35)$$

O operador vec é definido como

$$\text{vec} : \mathbb{F}^{m \times n} \rightarrow \mathbb{F}^{mn} \quad (A.36)$$

e é tal que

$$\text{vec}(\mathbf{A}) = \begin{bmatrix} a_{1,1} & \dots & a_{m,1} & a_{1,2} & \dots & a_{m,2} & \dots & a_{1,n} & \dots & a_{m,n} \end{bmatrix}^T. \quad (A.37)$$

Em outras palavras, o resultado de sua aplicação sobre uma matriz é um vetor formado pelo “empilhamento” de suas colunas.

A.2 Expansão de Taylor de funções multivariadas

Nesta seção, faz-se uma breve revisão sobre a expansão em série de Taylor de funções de múltiplas variáveis. Inicialmente, a expansão de funções que resultam em um escalar é descrita. Em seguida, aborda-se o caso em que as funções consideradas são campos vetoriais.

A.2.1 Expansão de Taylor de funções reais multivariadas

Seja $f : \mathbb{R} \rightarrow \mathbb{R}$ uma função real de uma única variável infinitamente diferenciável em uma vizinhança de $\xi \in \mathbb{R}$. Relembremos que a expansão em série de Taylor de f em torno ξ é dada por

$$s(x) = f(\xi) + \left. \frac{df}{dx} \right|_{x=\xi} (x - \xi) + \frac{1}{2!} \left. \frac{d^2 f}{dx^2} \right|_{x=\xi} (x - \xi)^2 + \frac{1}{3!} \left. \frac{d^3 f}{dx^3} \right|_{x=\xi} (x - \xi)^3 + \dots \quad (A.38)$$

$$= \sum_{p=0}^{\infty} \frac{1}{p!} \left. \frac{d^p f}{dx^p} \right|_{x=\xi} (x - \xi)^p \quad (A.39)$$

em que

$$\left. \frac{d^p f}{dx^p} \right|_{x=\xi} \quad (A.40)$$

denota o valor da p -ésima derivada de f em relação a x no ponto ξ , admite-se que $0! = 1$ e que o resultado de (A.40) para $p = 0$ é $f(\xi)$. Observe que a série é composta por um somatório de potências ponderadas pelos valores de

$$\frac{1}{p!} \frac{d^p f}{dx^p} \Big|_{x=\xi} \quad (A.41)$$

os quais são denominados coeficientes da expansão de Taylor. Adicionalmente, o $(p + 1)$ -ésimo termo da série é chamado de termo de ordem p .

Como mencionado no Capítulo 3, f é dita analítica se e somente se, $\forall \xi \in \mathbb{R}$, existe uma vizinhança de ξ na qual $f(x) = s(x)$, isto é, na qual a função é dada com exatidão por sua expansão em série de Taylor em torno de ξ .

Quando a expansão em série de Taylor é calculada incluindo-se somente os termos de ordem até P , diz-se que trata-se de uma expansão truncada na ordem P , a qual fornece uma aproximação polinomial da função original.

Consideremos agora uma função real $g(\mathbf{x})$ infinitamente diferenciável em relação aos componentes $x_1, \dots, x_n$ de $\mathbf{x} \in \mathbb{R}^n$ em uma vizinhança do ponto

$$\xi = [\xi_1 \ \dots \ \xi_n]^T. \quad (A.42)$$

Analogamente a (A.39), a expansão em série de Taylor de g é dada por

$$v(\mathbf{x}) = \sum_{p_1=0}^{\infty} \dots \sum_{p_n=0}^{\infty} \frac{1}{p_1! \dots p_n!} \frac{\partial^{p_1+\dots+p_n} g}{\partial^{p_1} x_1 \dots \partial^{p_n} x_n} \Big|_{\mathbf{x}=\xi} \prod_{j=1}^p (x_j - \xi_j)^{p_j} \quad (A.43)$$

em que

$$\frac{\partial^0 g}{\partial^0 x_1 \dots \partial^0 x_n} \Big|_{\mathbf{x}=\xi} = g(\xi). \quad (A.44)$$

Neste caso, cada valor dado por

$$\frac{1}{p_1! \dots p_n!} \frac{\partial^{p_1+\dots+p_n} g}{\partial^{p_1} x_1 \dots \partial^{p_n} x_n} \Big|_{\mathbf{x}=\xi} \quad (A.45)$$

é um coeficiente de ordem $p_1 + \dots + p_n$ da série de Taylor de g .

Assim como no caso de uma única variável, g é dita analítica se e somente se, $\forall \xi \in \mathbb{R}^n$, existe uma vizinhança de ξ em que $v(\mathbf{x}) = g(\mathbf{x})$.

Também à semelhança do caso anterior, a expansão de Taylor de uma função multivariável truncada na ordem P é obtida mantendo-se apenas os termos do somatório cujas ordens são menores ou iguais a P .

Vejamos um exemplo concreto da expansão (A.43). Seja a função real $g(\mathbf{x}) \in \mathbb{R}^2 \rightarrow \mathbb{R}$ dada por

$$g(\mathbf{x}) = e^{2x_1 - x_2} \quad (\text{A.46})$$

sendo $\mathbf{x} = [x_1 \ x_2]^T$. Vamos determinar os termos da expansão em série de Taylor de g truncada na ordem $P = 2$ em torno de $\xi = \mathbf{0} \in \mathbb{R}^2$. Para tanto, calculam-se as derivadas

$$\frac{\partial g}{\partial x_1} = 2e^{2x_1 - x_2} \quad (\text{A.47})$$

$$\frac{\partial g}{\partial x_2} = -e^{2x_1 - x_2} \quad (\text{A.48})$$

$$\frac{\partial^2 g}{\partial x_1^2} = 4e^{2x_1 - x_2} \quad (\text{A.49})$$

$$\frac{\partial^2 g}{\partial x_2^2} = e^{2x_1 - x_2} \quad (\text{A.50})$$

$$\frac{\partial^2 g}{\partial x_1 \partial x_2} = \frac{\partial^2 g}{\partial x_2 \partial x_1} = -2e^{2x_1 - x_2}. \quad (\text{A.51})$$

Calculando-se os valores dados pelas expressões acima para $x_1 = \xi_1 = 0$ e $x_2 = \xi_2 = 0$ e substituindo-se os coeficientes resultantes na fórmula (A.43), obtém-se

$$v(\mathbf{x}) = 2x_1 - x_2 + 2x_1^2 + \frac{1}{2}x_2^2 - 2x_1x_2. \quad (\text{A.52})$$

A.2.2 Expansão de Taylor de campos vetoriais

Seja $\mathbf{f}$ um campo vetorial definido sobre o $\mathbb{R}^n$. Por definição, $\mathbf{f}$ é tal que cada um de seus componentes é dado por uma função real de múltiplas variáveis. Denotemos tais componentes por $f_i \in \mathbb{R}^n \rightarrow \mathbb{R}$, para $i \in \{1, \dots, n\}$.

Como visto na subseção anterior, se cada uma das funções $f_i(\mathbf{x})$ é infinitamente diferenciável em relação aos componentes de $\mathbf{x}$ em torno de uma vizinhança de $\xi \in \mathbb{R}^n$, então é possível determinar a expansão em série de Taylor v_i na forma (A.43) de cada um dos componentes f_i de $\mathbf{f}$.

Vejamos agora como se pode utilizar o conceito de produto de Kronecker para se escrever de forma compacta um campo vetorial formado pelas expansões em série de Taylor de cada um dos componentes de $\mathbf{f}$ em torno de $\xi = [\xi_1 \ \dots \ \xi_n]^T$. Primeiro, vamos definir

$$\bar{\mathbf{x}} = [\bar{x}_1 \ \bar{x}_2 \ \dots \ \bar{x}_n]^T \quad (A.53)$$

$$= [x_1 - \xi_1 \ x_2 - \xi_2 \ \dots \ x_n - \xi_n]^T \quad (A.54)$$

$$= \mathbf{x} - \xi \quad (A.55)$$

e

$$\bar{\mathbf{x}}_0 = 1 \quad (A.56)$$

$$\bar{\mathbf{x}}_1 = \bar{\mathbf{x}} \quad (A.57)$$

$$\bar{\mathbf{x}}_i = \bar{\mathbf{x}} \otimes \bar{\mathbf{x}}_{i-1}, \quad \text{para } i > 1. \quad (A.58)$$

Com isso, podemos escrever um campo vetorial $\mathbf{v}(\mathbf{x})$ que contém todas as expansões dos elementos de $\mathbf{f}(\mathbf{x})$ em torno de ξ como

$$\mathbf{v} = \sum_{i=0}^{\infty} \mathbf{F}_i \bar{\mathbf{x}}_i \quad (A.59)$$

em que cada matriz $\mathbf{F}_i \in \mathbb{R}^{n \times n^i}$ é tal que sua k -ésima linha contém os coeficientes de ordem i da expansão de Taylor de $f_k(\mathbf{x})$, com $k \in \{1, \dots, n\}$.

Consideremos um exemplo da forma acima em que $n = 2$ e as séries de $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$ em torno de $\xi = [\xi_1 \ \xi_2]^T$ são truncadas na ordem $P = 3$. Das definições dadas, temos que

$$\mathbf{x}_0 = 1 \quad (A.60)$$

$$\mathbf{x}_1 = [\bar{x}_1 \ \bar{x}_2]^T \quad (A.61)$$

$$\mathbf{x}_2 = [\bar{x}_1^2 \ \bar{x}_1 \bar{x}_2 \ \bar{x}_2 \bar{x}_1 \ \bar{x}_2^2]^T \quad (A.62)$$

$$\mathbf{x}_3 = [\bar{x}_1^3 \ \bar{x}_1^2 \bar{x}_2 \ \bar{x}_1 \bar{x}_2^2 \ \bar{x}_2 \bar{x}_1^2 \ \bar{x}_2 \bar{x}_1 \bar{x}_2 \ \bar{x}_2^2 \bar{x}_1 \ \bar{x}_2^3]^T. \quad (A.63)$$

Já as matrizes $\mathbf{F}_i$ são dadas por

$$\mathbf{F}_0 = [f_1(\xi_1) \quad f_2(\xi_2)]^T \quad (A.64)$$

$$\mathbf{F}_1 = \left[\begin{array}{cc} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{array} \right] \bigg|_{\mathbf{x}=\xi} \quad (A.65)$$

$$\mathbf{F}_2 = \frac{1}{2!} \left[\begin{array}{cccc} \frac{\partial^2 f_1}{\partial x_1^2} & \frac{\partial^2 f_1}{\partial x_1 \partial x_2} & \frac{\partial^2 f_1}{\partial x_2 \partial x_1} & \frac{\partial^2 f_1}{\partial x_2^2} \\ \frac{\partial^2 f_2}{\partial x_1^2} & \frac{\partial^2 f_2}{\partial x_1 \partial x_2} & \frac{\partial^2 f_2}{\partial x_2 \partial x_1} & \frac{\partial^2 f_2}{\partial x_2^2} \end{array} \right] \bigg|_{\mathbf{x}=\xi} \quad (A.66)$$

$$\mathbf{F}_3 = \frac{1}{3!} \left[\begin{array}{ccccccccc} \frac{\partial^3 f_1}{\partial x_1^3} & \frac{\partial^3 f_1}{\partial x_1^2 \partial x_2} & \frac{\partial^3 f_1}{\partial x_1 \partial x_2 \partial x_1} & \frac{\partial^3 f_1}{\partial x_1 \partial x_2^2} & \frac{\partial^3 f_1}{\partial x_2 \partial x_1^2} & \frac{\partial^3 f_1}{\partial x_2 \partial x_1 \partial x_2} & \frac{\partial^3 f_1}{\partial x_2^2 \partial x_1} & \frac{\partial^3 f_1}{\partial x_2^3} \\ \frac{\partial^3 f_2}{\partial x_1^3} & \frac{\partial^3 f_2}{\partial x_1^2 \partial x_2} & \frac{\partial^3 f_2}{\partial x_1 \partial x_2 \partial x_1} & \frac{\partial^3 f_2}{\partial x_1 \partial x_2^2} & \frac{\partial^3 f_2}{\partial x_2 \partial x_1^2} & \frac{\partial^3 f_2}{\partial x_2 \partial x_1 \partial x_2} & \frac{\partial^3 f_2}{\partial x_2^2 \partial x_1} & \frac{\partial^3 f_2}{\partial x_2^3} \end{array} \right] \bigg|_{\mathbf{x}=\xi}. \quad (A.67)$$

Assim, temos que o campo vetorial

$$\mathbf{v}_3(\mathbf{x}) = \mathbf{F}_0 + \mathbf{F}_1 \bar{\mathbf{x}}_1 + \mathbf{F}_2 \bar{\mathbf{x}}_2 + \mathbf{F}_3 \bar{\mathbf{x}}_3 \quad (A.68)$$

contém como componentes as expansões em série de Taylor de $f_1(\mathbf{x})$ e de $f_2(\mathbf{x})$ em torno de ξ truncadas na ordem $P = 3$.

É fácil vermos que, em decorrência do que foi discutido anteriormente, se todos os componentes de $\mathbf{f}$ são funções analíticas, então qualquer $\xi \in \mathbb{R}^n$ possui uma vizinhança em que $\mathbf{f}$ é dada exatamente pelo campo vetorial $\mathbf{v}$ contendo as expansões (não-truncadas) de seus componentes em torno de ξ .

Note que, devido ao uso do produto de Carleman e da maneira como as matrizes são definidas, todos os respectivos termos do somatório (A.43) com ordem menor ou igual a três estão incluídos na expansão de cada um dos componentes de $\mathbf{f}$. É precisamente esta ideia que é empregada no Capítulo 5 para se representar as expansões em série de Taylor das equações diferenciais que compõem um sistema na forma linear analítica (5.48)–(5.49), tanto para a exposição do método da exponencial crescente (ver Seção 5.2) quanto para a do método de Carleman (ver Seção 5.3).

Apêndice B

Publicação

Neste apêndice, apresenta-se o artigo que foi publicado durante a realização deste trabalho. A referência completa de tal publicação é descrita abaixo. Nas páginas seguintes, disponibiliza-se o artigo.

GOULART, J. H. M.; BURT, P. M. S. Efficient kernel computation for Volterra filter structure evaluation. *IEEE Signal Processing Letters*, v. 19, n. 3, p. 135–138, mar. 2012. ISSN 1070-9908.

Efficient Kernel Computation for Volterra Filter Structure Evaluation

José Henrique de Morais Goulart and Phillip Mark Seymour Burt

Abstract—Despite their generality, conventional Volterra filters are inadequate for some applications, due to the huge number of parameters that may be needed for accurate modelling. When a state-space model of the target system is known, this can be assessed by computing its kernels, which also provides valuable information for choosing an adequate alternate Volterra filter structure, if necessary, and is useful for validating parameter estimation procedures. In this letter, we derive expressions for the kernels by using the Carleman bilinearization method, for which an efficient algorithm is given. Simulation results are presented, which confirm the usefulness of the proposed approach.

Index Terms—Nonlinear systems, Volterra filters.

I. INTRODUCTION

VOLTERRA filters (VF) are discrete-time models used in several digital signal processing applications to represent nonlinear systems [1]–[3]. They provide an input/output relation that is characterized by a set of points of hypersurfaces called Volterra kernels.

For systems that admit a Volterra representation and have dynamic nonlinearities, the series underlying the VF is composed by an infinite number of kernels. In [4] it was shown that, if the system has fading memory, then its associated operator can be approximated within arbitrary precision by a Volterra model with a finite number of kernels. This number depends on the desired precision, on system characteristics and on the set of bounded inputs considered (see [4, Ths. 1, 3]).

The major drawback of a conventional VF (CVF) is the exponential growth of its number of parameters with the order and the memory of the model, which causes severe problems particularly when the modelled system is strongly nonlinear. This has motivated several researchers to study alternate Volterra filter (AVF) structures that offer a good compromise between precision and structural complexity [2], [3], especially when real time use (and adaptation) is sought.

In some applications, an approximate physical model (PM) of the target system may be available, in the form of state-space equations [5]. However, it may be the case that only typical parameter values are known, and not the exact ones. Furthermore, the model may represent a class of devices whose parameters vary, as occurs, e.g., in a production line. In this situation, the

use of a VF is often more convenient, due to the simplicity in the formulation of the parameter estimation or adaptation problem [1], since the model is linear in the parameters. Still another advantage of a VF is its guaranteed stability that stems from the absence of feedback, which is not true for PMs.

However, the use of an AVF structure as those in [2], [3] may be necessary in order to avoid excessive computational complexity. In principle, this can be assessed with the calculation of the kernels from the state equations, by which one can estimate how many parameters a CVF should have to meet a certain precision level. If needed, an adequate AVF structure can be chosen with the information given by those kernels, which are also useful as a reference for further evaluation of the AVF and even for validating parameter estimation procedures, possibilities that seem unexplored in the literature (in general, the kernels used in simulations are either artificial, and thus unrealistic, or obtained from measured signals, and thus affected by noise and estimation uncertainties).

Although several methods have been proposed for calculating kernels related to state equations [6], [7], some are not appropriate for practical use. In this letter, the problems involved in the use of the approaches of [6] will be briefly described. After that, we will develop expressions for the kernels of a CVF, given a PM, by using the Carleman bilinearization method [7], [8]. Also, we propose an efficient implementation for this method, which has shown itself to be effective for the problem at hand. Finally, important aspects to consider when using the computed kernels for evaluating a VF structure are discussed and simulation results are presented.

II. PROBLEM STATEMENT

The input/output relation of a conventional Volterra filter with triangular kernels [3] is given by

$$y(n) = \sum_{p=1}^P \sum_{k_1=0}^{N_p-1} \dots \sum_{k_p=k_{p-1}}^{N_p-1} v_p(k_1, \dots, k_p) \times u(n - k_1) \dots u(n - k_p) \quad (1)$$

where $v_p(k_1, \dots, k_p)$ is the p th order kernel, with memory extension (in samples) N_p , and P is the maximum nonlinearity degree of the CVF. To model a physical system by (1), we consider the structure shown in Fig. 1, which operates as follows. First, the excitation sequence $u(n)$ is converted to the input signal, $u(t)$, which passes through the reconstruction filter $G_1(j\Omega)$, resulting in $u_c(t)$. Following that, $u_c(t)$ is fed to the continuous-time system, denoted by the $V[\cdot]$ operator, which produces the response $y_c(t)$. This signal, in turn, excites the anti-aliasing

Manuscript received November 02, 2011; revised December 12, 2011; accepted December 23, 2011. Date of publication December 30, 2011; date of current version January 19, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kenneth E. Barner.

The authors are with the Laboratory of Communication and Signals, Department of Telecommunication and Control Engineering of EPUSP, University of São Paulo, São Paulo – SP, Brazil, CEP 05508-900 (e-mail: jgoulart@lcs.poli.usp.br; phillip@lcs.poli.usp.br).

Digital Object Identifier 10.1109/LSP.2011.2182342

Fig. 1. Structure considered to obtain a reference conventional Volterra filter model of the physical system V .

filter $G_2(j\Omega)$, resulting in the output signal $y(t)$, which is sampled to produce $y(n)$. For simplicity, we will consider that V is time invariant.

From the expressions given in [7, pp. 62–63], we conclude that the continuous-time system formed by the series connection of G_1 , V and G_2 can be represented by the Volterra series

$$y(t) = \sum_{p=1}^{\infty} \int_0^{\infty} \cdots \int_{\tau_{p-1}}^{\infty} \tilde{h}_p(\tau_1, \dots, \tau_p) \times u(t - \tau_1) \dots u(t - \tau_p) d\tau_p \dots d\tau_1 \quad (2)$$

with kernels $\tilde{h}_p(\tau_1, \dots, \tau_p)$ such that their generalized frequency response functions (GFRFs)

$$\tilde{H}_p(j\Omega_1, \dots, j\Omega_p) = \int_0^{\infty} \cdots \int_0^{\infty} \tilde{h}_p(\tau_1, \dots, \tau_p) \times e^{-j(\Omega_1\tau_1 + \dots + \Omega_p\tau_p)} d\tau_1 \dots d\tau_p \quad (3)$$

are given by

$$\tilde{H}_p(j\Omega_1, \dots, j\Omega_p) = \left(\prod_{q=1}^p G_1(j\Omega_q) \right) H_p(j\Omega_1, \dots, j\Omega_p) \times G_2(j\Omega_1 + \dots + j\Omega_p) \quad (4)$$

where $H_p(j\Omega_1, \dots, j\Omega_p)$ is the p th order GFRF of V , which is assumed to admit a Volterra representation. Basically, this requires the existence of a series that, under certain conditions, converges and provides an exact representation of V [8].

The complete structure of Fig. 1 is a discrete-time system, V_d , which is to be modelled by a VF. With the frequency-domain characterization of (4), we can, by using the approach of [9], impose that the (discrete-time) GFRFs associated with the CVF be such that

$$V_p(e^{j\omega_1}, \dots, e^{j\omega_p}) = \tilde{H}_p\left(j\frac{\omega_1}{\pi}\Omega_N, \dots, j\frac{\omega_p}{\pi}\Omega_N\right) \quad (5)$$

for $-\pi \leq \omega_i \leq \pi$, $i = 1, \dots, p$, where $\Omega_s = 2\Omega_N$ is the sampling frequency. Assuming that G_2 eliminates any aliasing effects and that Ω_s is sufficiently large to represent the modelled system characteristics, then (5) will provide a discrete-time model analogous to V on the band of interest.

From the above discussion we see that, in order to determine the GFRFs of the CVF and, consequently, its kernels, it is necessary to know G_1 , G_2 and the GFRFs of V . It should be noted that this frequency-domain derivation is advantageous over the direct sampling of the time-domain kernels, because it prevents the aliasing of components with frequency greater than Ω_N [9].

It is also important to point out that environmental factors, as, e.g., a room acoustic response, are not included in G_2 , which is simply an anti-aliasing filter but is sufficient for obtaining a reference VF model of the nonlinear system alone. For applications where such factors are present, it can be shown from (5), [7, p. 62, eq. 23] and a discrete-time version of it that a linear system G_E between V and G_2 can be perfectly taken into account by a discrete-time linear system after V_d , which remains

unaffected.¹ Thus, in the following, we will assume that G_1 and G_2 are fixed and known, focusing ourselves on the determination of the GFRFs $H_p(j\Omega_1, \dots, j\Omega_p)$ from the state equations of the PM of V .

III. KERNELS COMPUTATION IN THE FREQUENCY-DOMAIN

Let us consider that an approximate PM of V is known, consisting in state equations in the linear-analytic form [7], [8]

$$\mathbf{x}'(t) = \mathbf{f}(\mathbf{x}(t)) + \mathbf{g}(\mathbf{x}(t))u_c(t) \quad y_c(t) = \lambda(\mathbf{x}(t)) \quad (6)$$

where $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{x}'$ is its derivative w.r.t. time and $\mathbf{f}$, $\mathbf{g}$ and λ are analytic functions of $\mathbf{x}$. This form is motivated by the fact that it represents a wide class of real systems [8] and admits a convergent Volterra representation for a set of bounded inputs, provided there exists a solution to the unforced differential equation [7]. It should be stressed out that the analyticity of the functions in (6) is essential, basically because the Volterra series arises from their Taylor series approximation around the initial state. We will also suppose that $\lambda : \mathbb{R}^m \rightarrow \mathbb{R}$ is a linear function $\lambda(\mathbf{x}) = \mathbf{c}^T \mathbf{x}$, that the initial state is $\mathbf{x}(0) = \mathbf{0}$ and the solution for $u(t) \equiv 0$ is $\mathbf{x}(t) \equiv \mathbf{0}$, simplifications that, in the present case, do not imply in loss of generality [7].

In [6], two approaches are proposed for systems in this form. The first one is through closed formulas involving the solution of the unforced differential equation (the “flow of $\mathbf{f}$ ”), which is obviously difficult to obtain for general systems. Also, using those formulas with numerical integration instead is not a good alternative, because they involve a time-reversed flow, making the corresponding differential equations unstable (assuming the system is stable). The other method consists in approximating the kernels by their Taylor series and requires the use of a symbolic computation package, since the expressions involve repeated differentiations. In practice, many terms may be necessary for a good approximation, which implies a high computational cost and makes this approach unfeasible.

A technique that has shown itself adequate in practice, for allowing an accurate and efficient computation, is known as Carleman bilinearization (or Carleman linearization [7], [8]). Additionally, it conveniently permits the direct calculation of the GFRFs, without the necessity of computing time-domain kernels. In the rest of this section, we will show how this can be done and propose an efficient algorithmic implementation.

A. Carleman Bilinearization

The central idea of Carleman bilinearization consists in representing the dynamics of the nonlinear system by a differential equation in the bilinear form, for whose kernels and GFRFs there are closed formulas. The brief review given below is based on [7]. For clarity, the time argument will be omitted.

Suppose that the first P kernels are desired. The starting point is to calculate the Taylor series expansions of $\mathbf{f}$ and $\mathbf{g}$, up to the P th order, around the initial state. This gives

$$\mathbf{x}' = \sum_{i=1}^P \mathbf{F}_i \mathbf{x}^{(i)} + \sum_{i=0}^{P-1} \mathbf{G}_i \mathbf{x}^{(i)} u \quad (7)$$

¹If, however, it is desired that the reference discrete-time model include environmental factors, then, of course, G_E can be absorbed in G_2 .

with the vector $\mathbf{x}^{(i)}$ containing every i th order product of the elements of $\mathbf{x}$, according to the formula

$$\mathbf{x}^{(i)} = \mathbf{x} \otimes \mathbf{x}^{(i-1)}$$

where $\otimes$ is the Kronecker product, $\mathbf{x}^{(1)} = \mathbf{x}$ and $\mathbf{F}_i, \mathbf{G}_i \in \mathbb{R}^{m \times m^i}$. Then, by using the rule

$$(\mathbf{x} \otimes \mathbf{x}^{(i-1)})' = \mathbf{x}' \otimes \mathbf{x}^{(i-1)} + \mathbf{x} \otimes (\mathbf{x}^{(i-1)})' \quad (8)$$

the bilinear differential equation

$$(\mathbf{x}^\otimes)' = \mathbf{F}\mathbf{x}^\otimes + \mathbf{G}\mathbf{x}^\otimes u + \mathbf{b}u \quad (9)$$

is constructed from (7), where

$$\mathbf{x}^\otimes \triangleq [(\mathbf{x}^{(1)})^T \dots (\mathbf{x}^{(P)})^T]^T. \quad (10)$$

The first m rows of $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{b}$ are obtained directly from (7) and the others calculated using (8) [7, pp. 108–110].

It is evident that, for $m, i > 1$, the vector $\mathbf{x}^{(i)}$ necessarily has redundant components. For practical purposes, it is more efficient to compute each row of $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{b}$ of (9) separately, disregarding the redundancies to obtain a state vector with exactly one component for each product $x_{i_1} \dots x_{i_k}$, with $1 \leq i_j \leq m$ and $k \leq P$. This will be explored by the algorithm described in Section III-B.

An important property of the Carleman method is that it gives exact kernels, provided that the expansions are calculated up to the desired nonlinearity degree. This is due to the form assumed by the matrices $\mathbf{F}$ and $\mathbf{G}$ [7].

B. Efficient Implementation

The inputs are the expansions of the components of $\mathbf{f}$ and $\mathbf{g}$, and thus completely determine the first m rows of $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{b}$. Note that, if a symbolic computation package is available, they can be obtained computationally. The strategy proposed is capable of efficiently computing nonredundant (and, consequently, smaller) versions of $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{b}$ by exploring two aspects of the Carleman method. First, the elements of $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{b}$ which belong to a same row correspond, actually, to the coefficients of the (truncated) Taylor expansion that gives the evolution of the associated component of $\mathbf{x}^\otimes$. Second, since each element $x_{i_1} \dots x_{i_k}$ of $\mathbf{x}^\otimes$ with $1 < k \leq P$ can be expressed as a product of preceding ones, we can compute its associated row recursively by using preceding row vectors analogously with the rule

$$(\mathbf{x}_{i_1} \dots \mathbf{x}_{i_k})' = (\mathbf{x}_{i_{q+1}} \dots \mathbf{x}_{i_k}) \times (\mathbf{x}_{i_1} \dots \mathbf{x}_{i_q})' + (\mathbf{x}_{i_1} \dots \mathbf{x}_{i_q}) \times (\mathbf{x}_{i_{q+1}} \dots \mathbf{x}_{i_k})' \quad (11)$$

where $q = \lfloor k/2 \rfloor$ and $1 \leq i_j \leq m$. In other words, the computation of each row of (9) corresponds to a symbolic derivation of the expansion it represents in terms of two previously derived expansions, which are represented by preceding rows. Note that the (symbolic) multiplications of (11) can be implemented efficiently by shifting the elements of the vectors according to the ordering of (the nonredundant) $\mathbf{x}^\otimes$, and the sum is performed as a usual vector sum.

TABLE I
TOTAL NUMBER OF OPERATIONS: RATIO OF THE DIRECT TO THE PROPOSED APPROACH FOR SOME VALUES OF m AND P

	$m = 2$	$m = 3$	$m = 4$	$m = 5$
$P = 2$	1.14	1.50	1.78	2.00
$P = 3$	2.69	5.30	8.22	11.20
$P = 4$	6.87	21.56	45.36	76.60
$P = 5$	18.47	97.13	286.02	611.83

A MATLAB version of this implementation was made available in [10]. It is significantly more efficient than the direct computation described in [7, pp. 108–110]. However, due to space limitations, we will not derive this gain, which is a complicated function of m and P [10] and is due to the fact that the redundancy eliminated with our algorithm grows exponentially with these variables. For instance, for $m = 3$ and $P = 4$, the nonredundant $\mathbf{F}$ and $\mathbf{G}$ contain less than 9% of the elements of their redundant counterparts. To exemplify the improvement, in Table I the ratio of the number of operations required by the direct approach to the number required by the proposed one is given for some values of m and P .

C. Calculating the Kernels With Use of the GFRFs

With $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{b}$ and defining $\mathbf{c}_V \triangleq [\mathbf{c}^T \ 0 \ \dots \ 0]^T \in \mathbb{R}^{M_P}$, being M_P the dimension of $\mathbf{b}$, the GFRFs of the system (9) can be calculated by [7]

$$H_p(j\Omega_1, \dots, j\Omega_p) = \mathbf{c}_V^T \mathbf{W}_p \mathbf{G} \dots \mathbf{G} \mathbf{W}_2 \mathbf{G} \mathbf{W}_1 \mathbf{b} \quad (12)$$

with $p = 1, \dots, P$, where

$$\mathbf{W}_k = (j(\Omega_1 + \dots + \Omega_k) \mathbf{I}_{M_P} - \mathbf{F})^{-1}, \quad k = 1, \dots, p. \quad (13)$$

In principle, with the GFRFs H_p and the filters G_1 and G_2 , the kernels of the CVF can be determined through the inverse transform of (5). In practice, assuming that the p th kernel contains little energy after some instant T_p , an approximation of the anti-transforms of (5) that will suffice is [9]

$$v_p(n_1, \dots, n_p) = \text{IFFT} \left\{ \hat{V}_p(k_1, \dots, k_p) \right\} \quad (14)$$

with $0 \leq k_1, \dots, k_p < N_p$, where

$$\hat{V}_p(k_1, \dots, k_p) = \hat{H}_p \left(j\Omega_s \frac{2k_1 - N_p}{2N_p}, \dots, j\Omega_s \frac{2k_p - N_p}{2N_p} \right) \quad (15)$$

and $N_p = \lceil T_p \Omega_s / 2\pi \rceil$. As observed in [9], it is important to note that this corresponds to sampling a circular convolution of the kernels with the sinc function in the time-domain, and thus there will be a noncausal portion of the kernels appearing at samples next to N_p , which may impair the results. One way of attenuating this effect is to multiply the right hand side of (15) by a multidimensional sinusoidal window.

IV. VOLTERRA FILTER STRUCTURE EVALUATION

The kernels associated with the PM provide many valuable information for choosing an appropriate VF structure, as explained next. First, by using a set of inputs with variances in-

creasing from about zero up to the desired maximum $\sigma_{\max}^2$, we can determine which combination of filter order P and memories N_p , $p = 1, \dots, P$, leads to a sufficiently accurate CVF. This is done by comparing the responses of some CVFs with different P and N_p to those inputs with the respective responses produced by the system of Fig. 1, which are computed by means of numerical integration of the PM (and filtering). To quantify this comparison, a useful measure is the fit index

$$\text{FIT}(\hat{\mathbf{y}}) = 100 \times \left(1 - \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|}{\|\mathbf{y} - \mu_{\mathbf{y}}\|} \right) \quad (16)$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}$ are vectors with samples of the outputs of the PM and the CVF, respectively, and $\mu_{\mathbf{y}}$ is the sample mean of $\mathbf{y}$. Obviously, one should choose the CVF with the lowest resulting number of coefficients that meets the desired precision, because it is the least computationally costly to use.

Second, if this complexity is inevitably high, some characteristics of the computed kernels may be explored for choosing an adequate AVF structure. For instance, if the kernels are smooth, then a sparse-interpolated scheme, as proposed in [2], can be used. Also, if the energy of the kernels is concentrated around their main diagonals, a diagonal representation with limited width can be used, as suggested in [3].

V. SIMULATION RESULTS

To exemplify the discussed evaluation, the kernels of first to fourth orders associated with the loudspeaker PM of [5] were computed. Since it involves a nonanalytic function $L(x)$, it was first adjusted by replacing this curve by an approximation in the form $L_a(x) = l_0 + l_1 \text{tg}^{-1}(\omega_l x + \phi_l)$. The sampling rate chosen was $F_s = 5$ kHz. For the memory, it was verified that $N = 100$ is a good upper bound to N_p , $p = 1, \dots, 4$, hence the kernels were calculated initially with this memory. Then, two CVFs were defined: H_{V_1} , with the four kernels and $N_1 = \dots = N_4 = 100$, and H_{V_2} , with only the three first kernels and $N_1 = 100$, $N_2 = N_3 = 90$. The inputs used were 50 different realizations of a Gaussian noise sequence with 500 samples having zero mean and with energy only up to approximately 625 Hz, since the PM represents a bass loudspeaker. Their variances were varied, and the fit indexes of the P th order approximations given by H_{V_1} and H_{V_2} w.r.t. the output of the PM were computed for different values of P . The ensemble average of the results is shown in Fig. 2 as a function of the variance of the input. Immediately, we see that the linear approximations (which are identical) degrade rapidly, while the higher order ones give fairly good results. Also, we note that H_{V_2} performs almost as good as H_{V_1} for $P = 2$ or $P = 3$. Thus, if, e.g., a 98% fit is desired and $\sigma_{\max}^2 \leq 7$ (as defined in Section IV), we can define the reference CVF with $P = 2$, $N_1 = 100$ and $N_2 = 90$, resulting in 4195 coefficients. However, if $\sigma_{\max}^2 = 9$, then a sufficient choice is $P = 3$, $N_1 = 100$ and $N_2 = N_3 = 90$, resulting in 129 775 coefficients, which is an increase by a factor of about 31.

Fig. 2. Fit index of the approximations given by H_{V_1} with orders $P = 1$ (*), $P = 2$ (▲), $P = 3$ (■), $P = 4$ (◆) and by H_{V_2} with orders $P = 1$ (●), $P = 2$ (■), $P = 3$ (■).

VI. CONCLUSION

In this letter, we approached the problem of evaluating Volterra filter structures for nonlinear system modelling in DSP applications. The proposed strategy consists in computing the resulting discrete-time system kernels through expressions that involve the filters of the A/D and D/A converters and the physical system kernels, calculated with the Carleman bilinearization. To this end, it is necessary that a state-space model is known *a priori*. As was shown by simulation results, this provides valuable information for choosing an adequate VF structure, i.e., one that exhibits a satisfactory accuracy and a sufficiently low computational complexity. In addition, the computed kernels may be used as a reference for validation of parameter estimation procedures and, if an alternate Volterra filter structure is necessary, for its quantitative evaluation.

REFERENCES

- [1] L. Azpicueta-Ruiz, M. Zeller, A. Figueiras-Vidal, J. Arenas-Garcia, and W. Kellermann, "Adaptive combination of Volterra kernels and its application to nonlinear acoustic echo cancellation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 1, pp. 97–110, Jan. 2011.
- [2] E. Batista, O. Tobias, and R. Seara, "A sparse-interpolated scheme for implementing adaptive Volterra filters," *IEEE Trans. Signal Process.*, vol. 58, no. 4, pp. 2022–2035, Apr. 2010.
- [3] M. Zeller and W. Kellermann, "Fast and robust adaptation of DFT-domain Volterra filters in diagonal coordinates using iterated coefficient updates," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1589–1604, Mar. 2010.
- [4] S. Boyd and L. Chua, "Fading memory and the problem of approximating nonlinear operators with Volterra series," *IEEE Trans. Circuits Syst.*, vol. CAS-32, no. 11, pp. 1150–1161, Nov. 1985.
- [5] D. Franken, K. Meerkotter, and J. Wassmuth, "Observer-based feedback linearization of dynamic loudspeakers with Ac amplifiers," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 233–242, Mar. 2005.
- [6] A. Isidori, *Nonlinear Control Systems*, 3rd ed. New York: Springer, 1995.
- [7] W. J. Rugh, *Nonlinear System Theory: The Volterra/Wiener Approach*. Baltimore, MD: Johns Hopkins Univ. Press, 1981.
- [8] R. Brockett, "Volterra series and geometric control theory," *Automatica*, vol. 12, pp. 167–176, Mar. 1976.
- [9] H. Koepl and D. Schwingshackl, "Comparison of discrete-time approximations for continuous-time nonlinear systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Montreal, QC, Canada, May 2004, pp. 881–884.
- [10] J. H. Goulart and P. M. S. Burt, Code for the Efficient Bilinearization Algorithm and Complexity Evaluation 2011 [Online]. Available: <http://www.lcs.poli.usp.br/~jgoulart/research.html>