Laguerre-Volterra Model and Architecture for MIMO System Identification and Output Prediction

Will X. Y. Li[1], *Member*, *IEEE*, Yao Xin[1], Rosa H. M. Chan[1], *Member*, *IEEE*, Dong Song[2], *Member*, *IEEE*, Theodore W. Berger[2], *Fellow*, *IEEE* and Ray C. C. Cheung[1], *Member*, *IEEE*

***Abstract*—A generalized mathematical model is proposed for behaviors prediction of biological causal systems with multiple inputs and multiple outputs (MIMO). The system properties are represented by a set of model parameters, which can be derived with random input stimuli probing it. The system calculates predicted outputs based on the estimated parameters and its novel inputs. An efficient hardware architecture is established for this mathematical model and its circuitry has been implemented using the field-programmable gate arrays (FPGAs). This architecture is scalable and its functionality has been validated by using experimental data gathered from real-world measurement.**

I. INTRODUCTION

Causal systems which exhibit high-order nonlinearities and dynamic properties are ubiquitous in the physical constitutions of organisms. Such systems may exist as parts/subconfigurations of animal cardiovascular system, endocrine system or nervous system (e.g. the hippocampal DG-CA3-CA1 subsystem). Various computational models have been established to model these complex parts or subconfigurations [1]. Generally, these models have multiple inputs and multiple outputs (MIMO) in structure, as shown in Fig. 1-a. It is an interesting problem how to derive the model outputs at a specific timing from existing information of the system the observer gathered prior to that timing.

Historically, there are two approaches (modeling techniques). One is parametric modeling, which entails *a prior* postulations of the model structure based on the fundamental physical mechanisms of the system that have been understood [2]. The parametric models, although possessing the capability of being directly and physically interpreted, have two major drawbacks. One is the built-in biases in model postulation owning to the existence of undiscovered mechanisms and processes. The other is the super-complicated computational processes that may be incurred due to the large amount of mechanisms underlying the system to be modeled parametrically. For example, in human nervous system, effects of these mechanisms vary greatly with ion channel densities, distributions in dendrites, and many other parameters [3].

In view of the above, the non-parametric models are proposed, which use engineering modeling techniques such as network analysis and statistical methods to investigate the properties of the complex systems [4]. They can take a general model form (Volterra series) and are able to be obtained directly from a broader repertoire of input-output data [1].

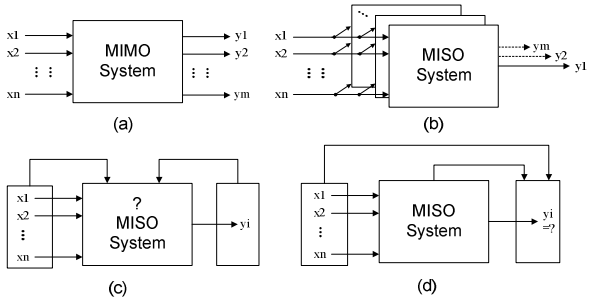

Fig. 1. The non-parametric modeling paradigm for MIMO systems. (a): A multi-input (x1-xn), multi-output (y1-ym) (MIMO) system. (b): Decomposition of MIMO system into multi-input, single-output (MISO) systems. (c): Estimation of system properties. (d): Prediction of system output.

In [5], a generalized Volterra model is introduced and two techniques are proposed thereon: 1) MIMO model decomposition (as shown in Fig. 1-b) and 2) Laguerre expansion of Volterra kernels (the model can also be termed as generalized Laguerre-Volterra model: GLVM) to track the properties of the nonlinear dynamic systems non-parametrically.

Prior to conducting system output prediction, model coefficients (Laguerre coefficients) have to be estimated in the first place using recorded input-output data under broad experimental conditions (as shown in Fig. 1-c) [6]. Subsequently, these coefficients can be utilized, together with the novel inputs, for prediction of the future outputs (as shown in Fig. 1-d). In an earlier publication, we report an efficient hardware framework established for *offline estimation* of the Laguerre coefficients based on field-programmable gate array (FPGA) [7]. In this paper, we introduce the hardware architecture for *online prediction* of model outputs employing high-order kernels and the reconfigurable platform. Although the work introduced in [7] is originally inspired by and dedicated to a neuroinformatics application, both architectures derived (in [7] and this paper) can be adapted and well applied to a broader range of real-world biological causal systems, for prediction of their behaviors.

The major contributions of this work and the novelty of our silicon design are reflected in the following aspects.

1) The original generalized Laguerre-Volterra model is adapted for behaviors prediction of *generic* biological MIMO casual systems.
2) An efficient hardware architecture is for the first time

[1]Will X. Y. Li, Yao Xin, Rosa H. M. Chan and Ray C. C. Cheung are with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong xyli@ee.cityu.edu.hk; yaoxin2@student.cityu.edu.hk; rosachan@cityu.edu.hk; rcheung@cityu.edu.hk

[2]Dong Song and Theodore W. Berger are with the Department of Biomedical Engineering, University of Southern California, Los Angeles, CA 90089, USA dsong@usc.edu; berger@bmsr.usc.edu

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

invented based on the adapted model.

3) The second-order Volterra kernel (with cross terms) is for the first time introduced to the hardware. Its silicon module has been successfully integrated to the top-level architectural framework.
4) The proposed design has been validated in experimental settings. It is integratable to the earlier established model parameters estimation architecture [7].

II. Method

For each of the MISO models proposed (shown in Fig. 1-b), a further decomposition can be conducted to partition it into four major components as shown in Fig 2. These are: 1) a feedforward Volterra kernel K which generates the "synaptic potential" u from recorded model inputs, 2) a feedback Volterra kernel H which generates the "after-potential" a from recorded model output, 3) a noise term ε which captures the influences of intrinsic system noise and unobserved model inputs, 4) an output trigger which generates the predicted model output measuring the above quantities. The terms *synaptic potential* and *after-potential* are derived from the study of neural networks. However, they can take on general meanings with regard to generic MIMO system modeling. The detailed algorithmic description of the GLVM can be found in [5].

Owing to the requirement of generic biological causal system behavior identification, the original GLVM is subject to further adaption, which is reflected in the following aspects.

- System inputs and feedback outputs are normalized/denormalized in the first/last stage of calculation. The normalization/denormalization procedure is conducted in consideration of a fixed Laguerre pole being adopted and in prevention of possible data overflow.
- The normalized input and feedback output should be convolved with different orders of Volterra kernels, as shown in (3) and (4) of [6]. In order to reduce the large amount of open parameters arised in direct Volterra modeling [8], an important technique named Laguerre expansion of Volterra kernel (LEV) is adopted. The *i*th order kernel k_i is expanded in terms of Laguerre basis functions b_j [9] so the potential values could be written into Wiener-Bose type expressions which are mathematically convenient. In this design, the Laguerre basis functions are obtainable through inverse Z-transform of transfer function of the Laguerre filter:

$$b_j^{(n)} = Z^{-1}\left\{\frac{\sqrt{1-\alpha_n^2}}{1-\alpha_n z^{-1}}\left(\frac{z^{-1}-\alpha_n}{1-\alpha_n z^{-1}}\right)^{j-1}\right\}, \quad (1)$$

  where j indicates the order of the Laguerre basis function; α_n is the Laguerre pole.
- The pre-threshold membrane potential w can be derived by adding u, a and ε, as:

$$w = u(k,x) + a(h,y) + \varepsilon(\sigma). \quad (2)$$

  In (2), k and h are feedforward and feedback Volterra kernels; x and y are input and feedback output.
  Model output y_{t+1} is predicted by measuring the summation result and with the utilization of a triggering function as shown in (3). In this design, we set the firing probability quantity in the original GLVM [5] directly as the system output, followed by a denormalization procedure, which effectively meet our purpose.

$$y_{t+1} = \frac{1}{2} + \frac{1}{2} erf\left(\frac{w-\theta}{\sqrt{2}\sigma}\right). \quad (3)$$

III. Hardware Architecture

The proposed method can be implemented on different platforms (either software or hardware). There are two major drawbacks with regard to the software implementation, which are inefficiency of model coefficients estimation [7] and incapability of conducting real-time model output prediction. On the other hand, the FPGAs, given their intrinsic hardware-level programmability and massive parallel processing capability, become an ideal choice for fast prototyping of the mathematical model. Fig. 3 shows the general hardware architecture for conducting model outputs prediction.

A. Vector Convolution

The vector convolution circuitry serves a vital component in both parameter estimation architecture [7] and the architecture proposed in this paper. Contrast to [7], the component herein consists an input multiplexion module. Previous studies suggest that not all inputs necessarily contribute to the outputs in a sparsely connected MIMO system, thus *model selection* has to be conducted in the first place for identification of effective channels [10]. The selection procedure can be effectively conducted using the method proposed in [11].

The algorithms used for vector convolution is similar to Algorithm 1&2 in [7], with the elimination of the 6th statement in Algorithm 1 of [7] because the convolution product is subject to further processing in a second-order GLVM (see Sec. III-B).

B. Correlation of Convolution Products

The procedure of conducting vector expansion is described by Algorithm 1 shown in this paper (following Algorithm 1&2 in [7]). The expanded vector H' is comprised of three parts: $H' = [H, H_{2s}, H_{2x}]$. H_{2s} accounts for the interactions among different basis functions of individual input by multiplications between each element pair (in all permutation). Given L the number of basis functions, $L \times (L+1)$ multiplication operations are needed. H_{2x} reveals contributions from pairs of different system inputs. $\binom{N}{2} \times L^2$ pair-wise multiplications are required under this premise. H' and the Laguerre coefficients C undergo a further multiplication-and-accumulation (MAC) operation and w can be derived at the root stage of the adder array of the MAC component.

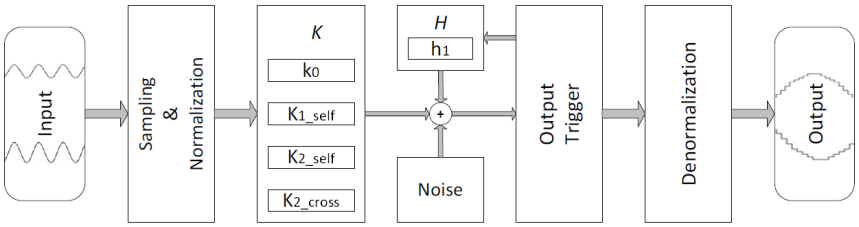

Fig. 2. The basic flow of calculation for the outputs of a generic MIMO causal system utilizing high order Volterra kernels. In the figure, K is the feedforward Volterra kernel, which consists of four sub-kernels, including the zeroth-order Volterra kernel k_0, the first-order self-kernel k_{1_self}, the second-order self-kernel k_{2_self} and the second-order cross-kernel k_{2_cross}; H is the feedback Volterra kernel.

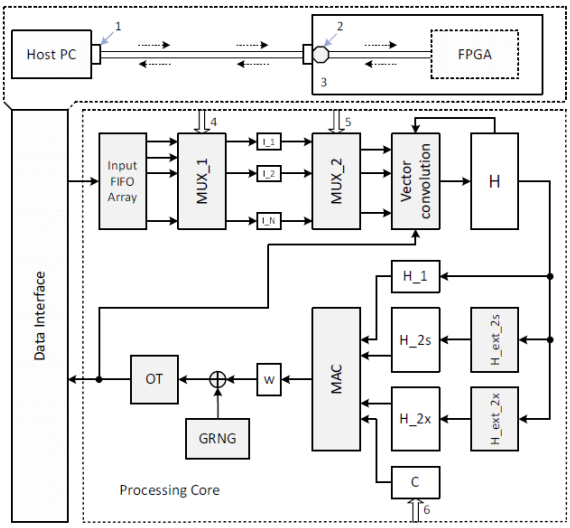

Fig. 3. Overview of the hardware architecture. The grey rectangular boxes indicate key functional units while the white ones indicate the register (arrays). In the figure, MUX_1 is input channel multiplexier; MUX_2 is convolution input multiplexier; H_ext_2s and H_ext_2s are augmented horizontal vector extension units; MAC is the multiplication-and-accumulation unit; GRNG is the Gaussian random number generator; OT is the output trigger; I_1 to I_N are selected effective model inputs; H is the augmented horizontal vector; H_1, H_2s and H_2x form the extended vector; w is the "pre-threshold potential"; C is Laguerre coefficients. In the figure: 1. USB port; 2. voltage converter; 3. development board; 4-5. control path; 6. input from the estimation core [7].

C. Output Generator

The GRNG is built to simulate the intrinsic Gaussian noise (such as the noised contributed by unobserved model inputs) of the stochastic system under modeling. The configuration of the GRNG is based on a uniform random number generator (URNG) which is designed to produce uniformly distributed random numbers $\{\lambda\}$ within the range of [0, 1]. We use $\{\sum_{i=1}^{N} \lambda_i\}$ to form a quasi-Gaussian distribution; The URNG is based on bitwise XOR operations between the lower 32 bits of a 43-bit linear feedback shift register (LFSR) and the lower 32 bits of a 37-bit cellular automata shift register (CASR) (cycle length: $O(2^{80})$), as shown in Fig. 4. This Gaussian noise quantity is further added to the "pre-threshold potential". The summation is passed to the output trigger, which adopts a transfer function to determine the value of model output at each time step.

Algorithm 1 Vector space expansion of convolution product (N: number of inputs)

```
H_2s = [];
H_2x = [];
for i = 1:N, % Input #1
    for ps = 1:L,
        for qs = ps:L,
            vs1 = H(:,((i-1)*L+ps));
            vs2 = H(:,((i-1)*L+qs));
            vs = vs1.*vs2;
            H_2s = horzcat(H_2s,vs);
    for j = (i+1):N, % Input #2
        for px = 1:L,
            for qx = 1:L,
                vx1 = H(:,((i-1)*L+px));
                vx2 = H(:,((j-1)*L+qx));
                vx = vx1.*vx2;
                H_2x = horzcat(H_2x,vx);
```

IV. Demonstrative Implementation and Result

The functionality of the proposed silicon architecture has been fully validated, employing the reconfigurable platforms. The application scenario of the demonstrative implementation is biological neural signal processing, for the mammalian nervous system of a brain region is a typical MIMO causal system which well satisfies the prerequisite of model application. In our experiment, the laboratory animals are trained to perform a type of memory task [12] and our silicon architecture is utilized to predict the output of their regional hippocampal brain signals. The model inputs are recorded using the multi-electrode arrays which are implanted into the nervous system. Model coefficients can be estimated at first offline utilizing the platform established in [7]. Then with the architecture proposed herein, model outputs can be predicted by employing the novel inputs and the estimated coefficients.

The validation work consists of two stages. In the first stage, we validate the functionality of the proposed mathematical model. Test result shows that the model has achieved desirable goodness-of-fit considering the fact that all points generated by the Kolmogorov-Smirnov (KS) test lie closely to the 45-degree line in the KS plot (within 95% confidence bounds) [6]. In the second stage, we validate the functionality of the our proposed circuit architecture. We use the FPGA-based hardware platform and the software platform to process a session of neuronal firing data. We compare the results generated by the hardware platform to the previous

TABLE I
THE FPGA RESOURCE UTILIZATIONS (NI: NON-INTERFACED)

	Compact		Paralleled	
	Consumed	Available	Consumed	Available
LUTs	36,819	69,120	55,617	150,720
LUTs (NI)	20,663	69,120	53,646	150,720
BRAMs	4	148	8	416
DSP48s	11	64	56	768
FPGA model	XC5VLX110T		XC6VLX240T	
Board model	Xilinx XUPV5-LX110T		Xilinx ML605	

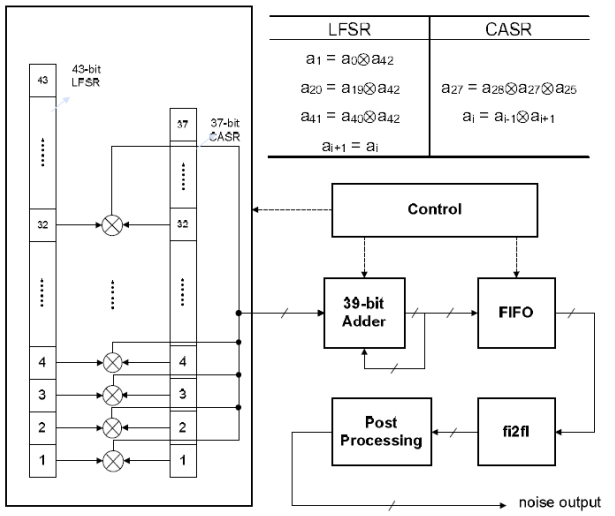

Fig. 4. The structure of the Gaussian white noise generator. (fi2fl is the fixed point to floating point conversion unit; the Post Processing unit is designed to adjust the mean and variance of the generated distribution to a standard normal distribution; the Control unit is in charge of URNG seed loading, FIFO r/w signal generation and component enabling.)

software platform. The normalized mean square error (NMSE) is defined as: $NMSE = \sum_{t=1}^{T}(y(t) - \tilde{y}(t))^2 / \sum_{t=1}^{T} \tilde{y}(t)^2$, where y and $\tilde{y}$ present the two data sets under comparison while t is the timing of sampling. We observe that the NMSE has been successfully controlled at the 10^{-13} scale, which is an even stricter condition being met with regard to data comparison. Thereupon we conclude that the hardware system is functionally equivalent to the biological neural network. In this demonstrative application, the proposed architecture is implemented using two different FPGA models (XC5VLX110T and XC6VLX240T). The information of hardware resource utilization is shown in Table I. For the compact architecture, only 2 FPGA processing units are used in the vector convolution module; for the paralleled architecture, the number of PEs equals to $N_{in} + 1$ (N_{in} is the number of effective inputs). For this demonstrative implementation, the compact architecture can suffice with the processing core operating with a 16MHz clock rate and producing a 98.16k samples/sec data throughput. For more data intensive DSP applications, the fully paralleled architecture can be adopted and the data throughput can reach up to 800k samples/sec.

V. CONCLUSIONS

An efficient reconfigurable architecture is established for behavior prediction of MIMO biological causal systems. The contribution of this work is reflected in several aspects. First, it is the first application of Volterra kernel based non-parametric approach to the computation of *generic* biological causal systems. Second, high-order kernels are for the first time adopted by the hardware. Third, the architecture is effective, efficient and capable of conducting real-time prediction of the GLVM outputs, using the estimated Laguerre coefficients. This design can be integrated with our previous model estimation architecture to form a complete full-scale signal analysis system and has good utilization potentiality in future engineering practice.

VI. ACKNOWLEDGEMENT

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 123312), and the Croucher Startup Allowance.

REFERENCES

[1] V. Z. Marmarelis, "Nonlinear dynamic modeling of physiological systems," *Hoboken: Wiley-IEEE Press*, 2004.
[2] P. Eykhoff, "System identification: Parameter and state estimation," *Wiley, New York*, 1974.
[3] M. Migliore and G. M. Shepherd, "Emerging rules for the distributions of active dendritic conductances," *Nature Reviews Neuroscience*, vol. 3, pp. 362–370, 2002.
[4] S. A. Fahmy and A. R. Mohan, "Architecture for real-time nonparametric probability density function estimation," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 5, pp. 910–920, 2013.
[5] D. Song, R. H. M. Chan, V. Z. Marmarelis, R. E. Hampson, S. A. Deadwyler, and T. W. Berger, "Nonlinear dynamic modeling of spike train transformations for hippocampal-cortical prostheses," *IEEE Transactions on Biomedical Engineering*, vol. 54, pp. 1053–1066, Jun 2007.
[6] D. Song, R. H. M. Chan, V. Z. Marmarelis, R. E. Hampson, S. A. Deadwyler, and T. W. Berger, "Nonlinear modeling of neural population dynamics for hippocampal prostheses," *Neural Networks*, vol. 22, pp. 1340–1351, 2009.
[7] W. X. Y. Li, R. H. M. Chan, W. Zhang, R. C. C. Cheung, D. Song, and T. W. Berger, "High-performance and scalable system architecture for the realtime estimation of generalized Laguerre-Volterra MIMO model from neural population spiking activity," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 1, pp. 489–501, 2011.
[8] H. Koeppl, "A local nonlinear model for the approximation and identification of a class of systems," *IEEE Transactions on Circuits and Systems II*, vol. 56, no. 4, 2009.
[9] V. Z. Marmarelis, "Identification of nonlinear biological systems using Laguerre expansions of kernels," *Annals of Biomedical Engineering*, vol. 21, pp. 573–589, 1993.
[10] D. Song, R. H. M. Chan, V. Z. Marmarelis, R. E. Hampson, S. A. Deadwyler, and T. W. Berger, "Sparse generalized Laguerre-Volterra model of neural population dynamics," *Proceedings of the 31st Annual International Conference of the IEEE EMBS*, pp. 4555–4558, 2009.
[11] D. Song, R. H. M. Chan, V. Z. Marmarelis, R. E. Hampson, S. A. Deadwyler, and T. W. Berger, "Statistical selection of multiple-input multiple-output nonlinear dynamic models of spike train transformation," *Proceedings of the 29th Annual International Conference of the IEEE EMBS*, pp. 4727–4730, 2007.
[12] R. E. Hampson, J. D. Simeral, and S. A. Deadwyler, "Distribution of spatial and nonspatial information in dorsal hippocampus," *Nature*, vol. 402, no. 6762, pp. 610–614, 1999.