Continuous-Time to Discrete-Time Conversion via a Novel Parametrized S-to-Z-Plane Mapping

William L. Melvin, *Member, IEEE,* and Douglas R. Frey, *Senior Member, IEEE*

***Abstract*— A parametrized s-to-z-plane map is introduced, where the conventional bilinear map and backward and forward Euler rules appear as special cases. Time and frequency-domain characteristics of the map are discussed, along with a simple technique for applying this map to adaptively reduce truncation error in the continuous-time to discrete-time conversion problem. An example illustrates the method and its potential for yielding more accurate discrete-time models over conventional approaches.**

I. Introduction

In general, no unique equivalence between the continuous-time (CT) and discrete-time (DT) domains exists, and thus a variety of approaches are available for the conversion [1]. A common problem concerns developing discretization rules which are both stable and accurate, especially for the relatively low sampling rates representative of a digital signal processing (DSP) environment. The critical factor regarding accuracy is taken as the difference between responses of the DT and CT models at sampling instances [1]. This is the best measure of performance when system dynamics are of concern, such as in control, robotics and navigation applications, and is superior to magnitude frequency response as the performance metric, since the latter metric represents only part of the performance described by the former [2], [3].

The bilinear map is a popular approach for the CT-to-DT conversion, yielding stable discrete-time models of moderate accuracy [2]. In this paper we develop a novel, first-order mapping strategy which offers improvement over the performance of the bilinear map. Specifically, we propose a time-varying mapping based on a class of first-order s-to-z-plane maps parametrized by a single parameter, for which the bilinear map appears as a special instance, to reduce truncation error ϵ_T. This class of first-order maps is then extended to the general case of "vector mappings," whereby each state variable is uniquely mapped. We restrict attention to first-order mappings since mappings based on higher order polynomials often suffer severe stability problems [4]. Note that numerical simulation environments control ϵ_T by manipulating sampling period T or order of integration [5]. In an analogous fashion, we attempt to reduce ϵ_T, without varying T, by adaptively choosing different members from the aforementioned class of first-order parametrized mappings through the appropriate mapping parameter selection.

Manuscript received October 13, 1995; revised June 7, 1996. This paper was recommended by Associate Editor R. W. Newcomb.

W. L. Melvin is with the USAF Rome Laboratory, Signal Processing Branch, Rome, NY 13441 USA.

D. R. Frey is with the Department of Electrical Engineering and Computer Science, Lehigh University, Bethlehem, PA 18015 USA.

Publisher Item Identifier S 1057-7130(97)07483-1.

II. Parametrized Mapping

We now derive the parametrized mapping. First, consider the linear state-equation

$$\frac{dx(t)}{dt} = f(t) = ax(t) + bu(t), \tag{1}$$

where $x(t)$ is the state-variable and $u(t)$ is the input. Equation (1) is easily related to the transfer function representation [5]. The exact solution to (1) at the sampling instance $t_{n+1} = (n+1)T$ is

$$x_{n+1} = x_n + \int_{nT}^{(n+1)T} f(\tau)\, d\tau. \tag{2}$$

The subscripts n and $n+1$ refer to samples at times nT and $(n+1)T$, respectively. Assuming that $f(\tau)$ is continuous on $(nT, (n+1)T)$, the Mean Value Theorem for Integrals applied to (2) yields

$$x_{n+1} = x_n T f(\xi); \qquad nT \le \xi \le (n+1)T. \tag{3}$$

Relying on the continuity of $f(\tau)$ and further assuming $f(\tau)$ is strictly concave up or down, a real number P between 0 and 2 exists such that

$$f(\xi) = \tfrac{1}{2}[Pf_{n+1} + (2-P)f_n]. \tag{4}$$

If $f(\tau)$ is not strictly concave up or concave down, (4) still holds and P may be any real number, provided $f_{n+1} \neq f_n$. If $f_{n+1} = f_n$, then (4) no longer is generally valid. Substituting (4) into (3) gives

$$x_{n+1} = x_n + Tf_n + \frac{T}{2}P(f_{n+1} - f_n). \tag{5}$$

The mapping that transforms (1) into (5) is the parametrized s-to-z-plane map, defined as

$$s = \frac{2}{T}\left[\frac{1 - z^{-1}}{P + (2-P)z^{-1}}\right] \tag{6}$$

where P is identified as the mapping parameter, and $P = 1$ yields the bilinear mapping, $P = 2$ results in the backward Euler rule, and $P = 0$ provides the forward Euler approximation [5]. Equation (6) maps the s-plane $j\omega$-axis into circular images in the z-plane for various values of $P \neq 0$. The left-half s-plane maps entirely within each respective circular image for values of $P > 0$. The case, $P = 0$, is degenerate, yielding

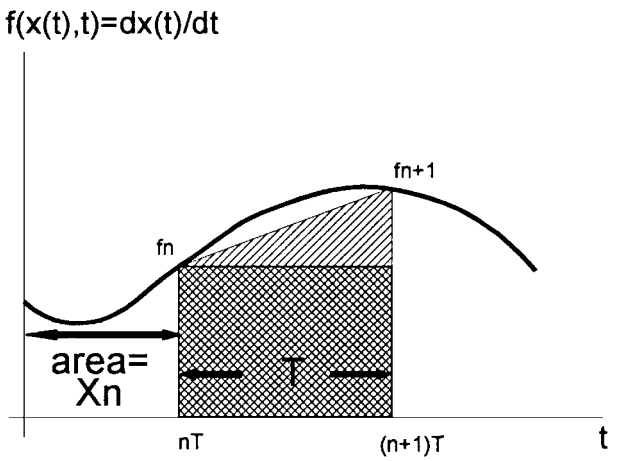

Fig. 1. Geometrical interpretation of the parametrized map in the time-domain.

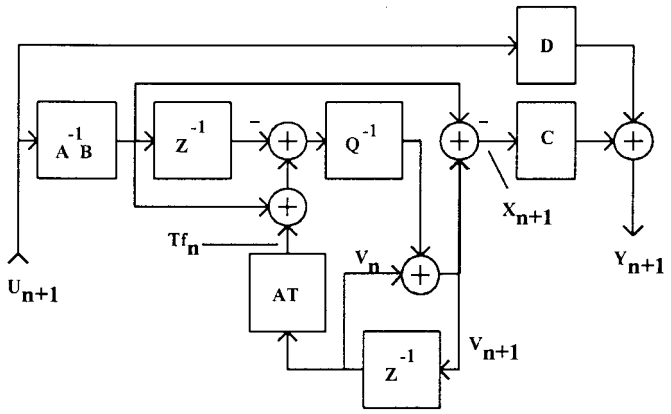

Fig. 2. Block diagram of digital filter realization using parametrized map.

an image which is a straight line. Note that the family of mappings defined by (6) is absolutely stable for all $P \geq 1$.

Equation (5) has an interesting geometrical interpretation in the time-domain. Upon considering the right-hand side of (5), the first term represents the previously computed area under the curve $f(t)$ up to nT. The second term is recognized simply as the area of a rectangle with base T and height f_n. The last term represents the weighted area of a triangle with base T and height, $f_{n+1} - f_n$. Fig. 1 serves to identify these three distinct areas. Observe from Fig. 1 that a unique value of the parameter P may be specified to identically represent the area under the curve, thereby yielding the exact response of (1) at $(n+1)T$ via (5), except when $f_{n+1} = f_n$. The exceptional condition may be thought of as weighting a triangle with zero area. Therefore, an alternative to reducing truncation error other than increasing the sampling rate $1/T$ is available by appropriately manipulating the mapping parameter P. Further examination of (5) or Fig. 1 indicates that best results may be obtained by manipulating the parameter on a time-step basis.

III. Digital Filtering Based on a Parametrized Map

While the last section demonstrated that varying P on a continuing basis could reduce truncation error to zero, there is no way to exactly choose P in practice. In this section we describe a very simple approach for estimating P on a time-step basis based on the geometrical interpretation shown in Fig. 1. In Fig. 1, a straight-line approximation $P = 1$ corresponds to the bilinear mapping. A value of P other than unity might be more appropriate in weighting the triangular area shown in the figure to better represent the area under the curve at the given time step, thereby reducing ϵ_T. This is particularly valid for relatively large T. Specifically, in this method the parameter P may be selected after determining whether $f(t)$ is concave up or concave down with respect to a straight-line approximation. If $f(t)$ is increasing $(f_{n+1} > f_n)$ and concave up, or decreasing $(f_{n+1} < f_n)$ and concave down, then P should be selected greater than unity. Alternatively, if $f(t)$ is decreasing and concave up, or increasing and concave down, P should be selected less than unity.

To implement the preceding idea, the bilinear map is initially used ($P = 1$ in (5)) to estimate x_{n+1} and f_{n+1}. A simple test of concavity is to determine whether $f_{n+(1/2)}$ lies above or below the straight-line approximation had by computing

$$f_{SL,n+(1/2)} = \tfrac{1}{2}(f_n + f_{n+1}) \tag{7}$$

and then comparing (7) with the midpoint estimate

$$\begin{aligned} f_{\text{est},n+(1/2)} &= \left(f_n + \frac{T}{2}\dot{f}_n\right) = \left(f_n + \frac{T}{2}\ddot{x}_n\right) \\ &\approx f_n + \frac{T}{2}\left(\frac{x_{n+1} - 2x_n + x_{n-1}}{T^2}\right). \end{aligned} \tag{8}$$

Concavity is "up" if the result in (8) is greater than that given in (7). Otherwise, it is "down." We propose that the parameter P be set to one of two quantized values P_{HI} or P_{LO} based upon the above considerations. After selecting P, the state variable x_{n+1} is then recomputed via (5).

We now recapitulate and derive a digital filter structure based on our method. Let a given filter be represented in a state-space formulation as follows:

$$\overline{f}_{n+1} = A\overline{x}_{n+1} + Bu_{n+1}; \quad y_{n+1} = C\overline{x}_{n+1} + Du_{n+1} \tag{9}$$

where we assume a single input and output but a vector of state variables. Vector quantities are denoted by a bar above the variable and A, B, C, and D are appropriately dimensioned real matrices, vectors, or constants. Next, we choose the generalized parametrized discretization equation for an Nth-order system as

$$\begin{aligned} \overline{x}_{n+1} &= \overline{x}_n + \frac{T}{2}P\overline{f}_{n+1} + \frac{T}{2}(2I - P)\overline{f}_n \\ P &= \text{diag}\{p_1, p_2, \cdots, p_N\} \end{aligned} \tag{10}$$

where I is an N by N identity matrix and P is a time-varying, diagonal parameter matrix allowing the ith state variable to be uniquely mapped by its own parameter p_i along the lines of the above discussion. Combining (9) and (10) yields a digital filter implementation for the system of (9). A more efficient realization is had by first combining (9) and (10), moving terms

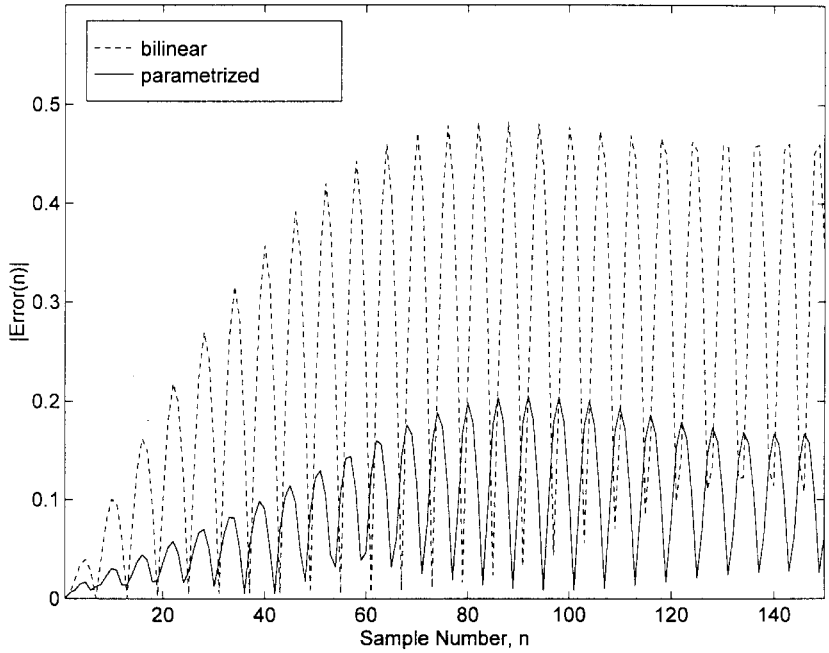

Fig. 3. Comparison of instantaneous error, transient region, parametrized, and bilinear maps.

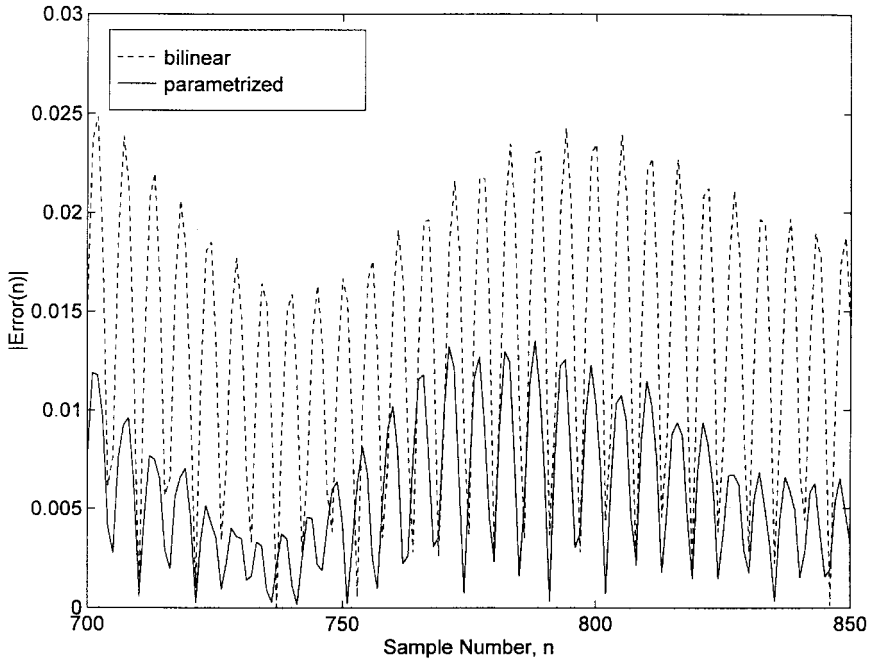

Fig. 4. Comparison of instantaneous error, steady-state region, parametrized and bilinear maps.

with $\overline{x}_{n+1}$ to the left, and adding $(I-(T/2)PA)A^{-1}Bu_{n+1}$ to both sides, yielding

$$\left(I-\frac{T}{2}PA\right)(\overline{x}_{n+1}+A^{-1}Bu_{n+1}) = \left(I+\frac{T}{2}(2I-P)A\right)(\overline{x}_n+A^{-1}Bu_n) + A^{-1}B(u_{n+1}-u_n). \tag{11}$$

Next, define the intermediate vector $\overline{v}_n$ and time-varying matrix Q as

$$\overline{v}_n=\overline{x}_n+A^{-1}Bu_n;\quad Q=I-\frac{T}{2}PA. \tag{12}$$

Substituting (12) into (11) leads to a convenient form for the parametrized digital filter as

$$\begin{aligned}\overline{v}_{n+1}&=\overline{v}_n+Q^{-1}[TA\overline{v}_n+A^{-1}B(u_{n+1}-u_n)];\\ \overline{x}_{n+1}&=\overline{v}_{n+1}-A^{-1}Bu_{n+1}\quad \overline{f}_{n+1}=A\overline{x}_{n+1}+Bu_{n+1}\\ y_{n+1}&=C\overline{x}_{n+1}+Du_{n+1}.\end{aligned} \tag{13}$$

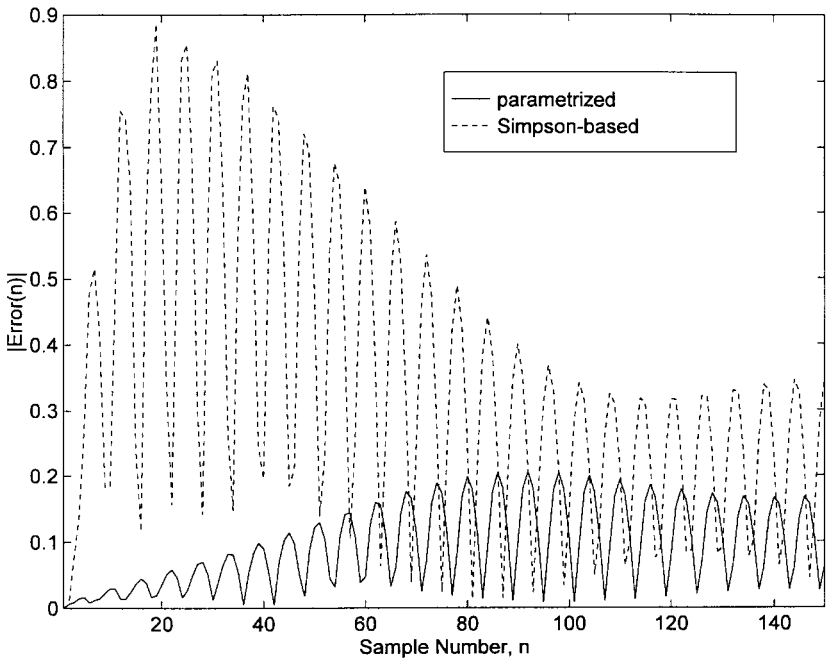

Fig. 5. Comparison of instantaneous error, transient region, parametrized, and Simpson-based maps.

Fig. 2 shows a block diagram of the parametrized filter structure based on (13). Observe that two passes are made on the parametrized filter realization for each time step. Initially, P is set to I (bilinear case) to perform the concavity test using new and stored values of x and f. Then, each diagonal element p_i is set to one of two quantized values $P_{\rm HI}$ or $P_{\rm LO}$ based on the outcome of the concavity test, and used to compute the final updated output on the second pass.

IV. Example

As an example, consider a second-order underdamped system, equivalent to a notch filter, with a natural frequency of 5.032 kHz, described by the following state-equation and output equation:

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix}=\begin{bmatrix}-1000 & -1E6\\ 1000 & 0\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix}+\begin{bmatrix}1000\\ 0\end{bmatrix}u(t)$$
$$y(t)=[-0.5\ 0]\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix}+0.5. \tag{14}$$

The forcing function is $\boldsymbol{u}(nT)=\boldsymbol{u_n}=\cos(\omega_1\cdot n\cdot T)$, with $\omega_1=2\pi\cdot 5.5$ kHz, $f_s=1/T=60$ kHz, and $X(0)=[2\ -0.001]^T$. For comparison, three approaches are used to discretize (14)—the parametrized filter realization of (13) with $\boldsymbol{P_{\rm HI}}=1.1$ and $\boldsymbol{P_{\rm LO}}=0.9$, the DT models resulting from the bilinear map $(P=I)$ [2], and the stabilized differentiator based on Simpson's integration rule [3]. For clarity, the procedure for applying the parametrized map to compute the output at a given sampling instance is as follows.

- (**Step 1**) : Compute $[x_{1,n+1},\ x_{2,n+1}]^T$ and $f_{1,n+1}$, $f_{2,n+1}]^T$ from (13) with $p_1=p_2=1$ (bilinear case);
- (**Step 2**): Compute the vectors $\overline{f}_{SL,n+(1/2)}$ and $\overline{F}_{\rm est,n+(1/2)}$ from (7) and (8) using results from the previous step;
- (**Step 3**): Select the values of p_1 and p_2 to be either $\boldsymbol{P_{\rm HI}}=1.1$ or $\boldsymbol{P_{\rm LO}}=0.9$, as discussed in the preceding section;
- (**Step 4**): Recompute $[x_{1,n+1},x_{2,n+1}]^T$ using (13) and the values of p_1 and p_2 computed in Step 3 and then compute y_{n+1}.

Fig. 3 shows the instantaneous error (relative to the analytical solution) for both the bilinear and the parametrized approach over the first 150 sampling instances. Fig. 4 displays the instantaneous error in the steady-state region. One may observe from Figs. 3 and 4 that the parametrized approach results in a much more accurate discrete-time model than the bilinear mapping for this test case. Figs. 5 and 6 compare the instantaneous error resulting from use of the differentiator based on Simpson's rule to the parametrized discretization. The error is significantly larger for the Simpson-based differentiator than both the bilinear and parametrized maps for this test case. It is pointed out that Simpson's rule used for differentiation is always unstable [4], and therefore must be stabilized by inverting the anomalous poles as discussed in [3], an approach which favors preserving magnitude frequency response. We attribute the larger error in the Simpson-based filter to inaccurate phase response, despite good magnitude frequency response.

Different values for $P_{\rm HI}$ and $P_{\rm LO}$ were tried during simulation. The values of $P_{\rm HI}=1.1$ and $P_{\rm LO}=0.9$ were not generally best for this example, but were selected as reasonable choices to demonstrate the possibility of performance improvements using this very simple algorithm. In fact, for this example, significantly better results were obtained using $P_{\rm HI}=1.15$ and $P_{\rm LO}=0.85$, as shown by the instantaneous error in Fig. 7 over 1000 samples. In general, we expect the optimum values of $P_{\rm HI}$ and $P_{\rm LO}$ to depend on the sampling rate, where values close to one are best for high sampling

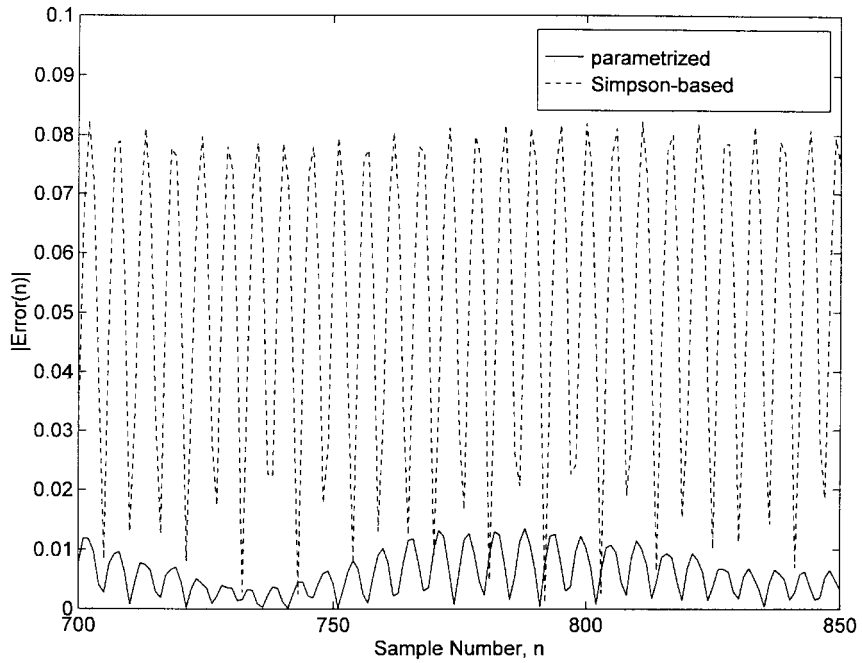

Fig. 6. Comparison of instantaneous error, steady-state region, parametrized, and Simpson-based maps.

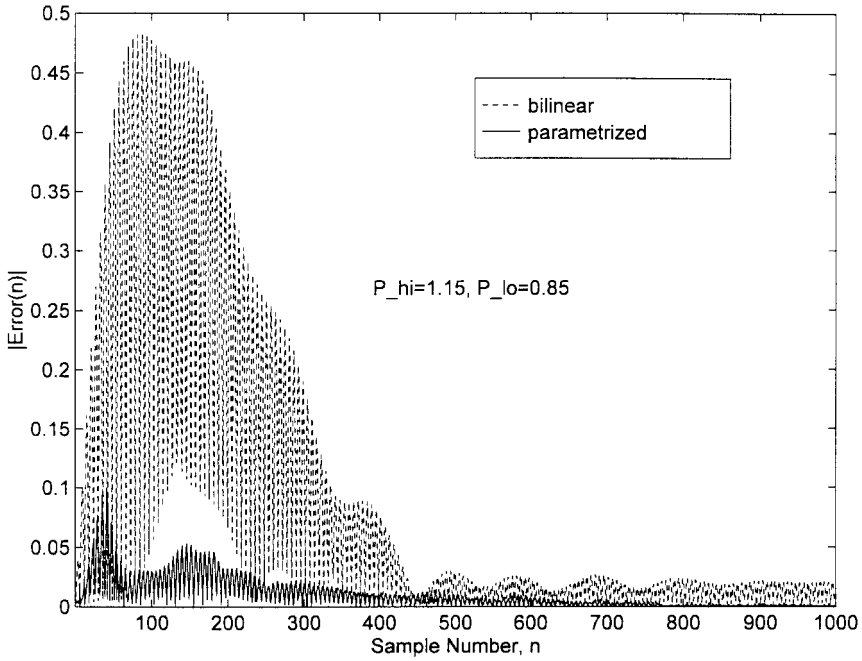

Fig. 7. Instantaneous error, parametrized, and bilinear case, P_{HI} and P_{LO} changed to 1.15 and 0.85.

rates and values of 1.2 and 0.8 provide better results for relatively lower sampling rates. We also point out that for our specific underdamped example, the complex conjugate poles of the time-varying DT system have a magnitude of 1.005 when $p_1 = p_2 = 0.9$. For all other combinations of p_1 and p_2, the poles of the time-varying system map within the unit circle. Thus, the time-varying application of the parametrized mapping allows for reduction in truncation error despite momentary pole placement outside the unit circle, which does not lead to instability for this example. Furthermore, in the case yielding the best results ($P_{\mathrm{HI}} = 1.15$, $P_{\mathrm{LO}} = 0.85$) the complex conjugate poles move even further from the unit circle for $p_1 = p_2 = 0.85$, with a temporary magnitude of 1.012. In addition, a normally distributed random signal of unit variance and zero mean, having excursions as large as ± 3 V, was added to the input $\boldsymbol{u_n}$ to further test the stability of the parametrized filter structure. The filter remained stable under such conditions. These results, coupled with the fact that it is a relatively simple task to monitor situations which give rise to instability and make appropriate adjustments to the selection of $p_1, p_2, \cdots, p_N$, suggest that the proposed method can be made robust in practice.

V. Conclusions

This paper discusses a novel parametrized map and corresponding digital parametrized filter structure, where the conventional bilinear transform and backward and forward Euler maps appear as special cases. This map has some interesting properties, including a simple geometrical interpretation and the capability to drive truncation error to much lower levels than conventional maps when the mapping parameter is allowed to vary with time. An elementary example demonstrates the potential performance improvement offered by the parametrized map, as compared with the bilinear transform and a higher-order Simpson-based differentiator, when applied in time-varying fashion with a simple parameter estimation technique. To conclude, we note that improved estimation methods allowing mapping parameters to vary over a continuum will result in potentially dramatic reduction in truncation error. Second, the parametrized mapping approach discussed allows unique selection of mapping parameters for each state-variable. This is a significant development since mutual integral/differential relationships may be accounted for, thus avoiding the expected performance degradation with growing system order that occurs with fixed mapping techniques such as the bilinear mapping [1]. Finally, the structure of the parametrized map may be useful for adaptive IIR filtering [6].

References

[1] Z. Kowalczuk, "On discretization of continuous-time state-space models: A stable-normal approach," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 1460–1477, Dec. 1991.
[2] A. Oppenheim and R. Schaeffer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[3] M. A. Al-Alaoui, "Novel IIR differentiator from the Simpson integration rule," *IEEE Trans. Circuits Syst. I*, vol. 41, pp. 186–187, Feb. 1994.
[4] A. M. Schneider, J. D. Kaneshige, and F. D. Groutage, "Higher order s-to-z mapping functions and their application in digitizing continuous-time filters," *Proc. IEEE*, vol. 79, pp. 1660–1674, Nov. 1991.
[5] L. Chua and P. Lin, *Computer Aided Analysis of Electronic Circuits: Algorithms and Computational Techniques*. Englewood Cliffs, NJ: Prentice-Hall, 1975.
[6] U. Forssén, "Adaptive bilinear digital filters," *IEEE Trans. Circuits Syst. II*, vol. 40, pp. 729–735, Nov. 1993.

William L. Melvin (S'89–M'94) was born in Danbury, CT, on February 27, 1967. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from Lehigh University, Bethlehem, PA, in 1989, 1992, and 1994, respectively. He also received a commission as an officer in the United States Air Force on June 3, 1989, postponing his service commitment until completion of the Ph.D. degree.

He is currently assigned to the United States Air Force Rome Laboratory, Signal Processing Branch, Rome, NY, where he develops algorithms and conducts experiments for improved detection of weak targets by airborne radar surveillance platforms. His areas of research interest include digital signal processing, nonlinear system theory and numerical methods.

Douglas R. Frey (S'74–M'76–SM'88) received the B.S., M.S., and Ph.D. degrees from Lehigh University, Bethlehem, PA, in 1973, 1974, and 1977, respectively.

He was with Bell Laboratories, Holmdel, NJ, in 1977 before accepting a teaching position at Lehigh University as an Assistant Professor. In 1983, he was promoted to Associate Professor of Electrical Engineering and continues to be a faculty member with the EECS Department. His main areas of interest are in circuit design and nonlinear circuits and systems theory. He has developed and taught a number of courses in these areas. His research includes work on Log filters, analog adaptive filters, nonlinear digital fixed and adaptive filters, and chaotic encoding. He is also an engineering consultant, in which capacity he has been involved in discrete, and semi- and full-custom integrated circuit design, including consumer electronics. He is currently a consultant for Analog Devices, Inc. He has seven patents and has presented and published a number of papers in his areas of interest.

Dr. Frey is a member of the Audio Engineering Society and a registered Professional Engineer in Pennsylvania.