IEEE Asilomar Conference on Signals, Systems and Computers, Monterey, CA, October ‘92

Short Term Memory Structures for Dynamic Neural Networks

Bert deVries, Jose C. Principe*

David Sarnoff Research Center
CN5300, Princeton, NJ 08543

*Computational NeuroEngineering Lab
University of Florida, Gainesville, FL 32611

Abstract

This paper presents a framework to design and characterize short term memory structures for neural networks. The gamma memory, a recursive linear structure, is presented as a generalization of the tapped delay line or the context memory units. The gamma memory principle can be enhanced to construct non-uniform time warping scales that may be useful in speech recognition.

1 Introduction

In this paper we discuss short term memory design methods for the classification of sequences using neural networks. A dynamic neural network is a neural model that has internal dynamics. Let us review the processing protocol first. We are given a P-dimensional (training) set of input sequences U = {u1(t), u2(t),..., uP(t), t=1,...,T} and a set of L classes C = {c1, c2,...,cL}. The goal is to develop a processor that minimizes the classification error rate. We focus on the case where the processor is implemented by a dynamic neural network. During training of the net a time varying pattern u(t), t=1,...,T, is presented to the net along with a target signal z(t) for output nodes. For classification problems, the target signal is usually specified only at the final time step t=T. At the classification time t=T, the entire pattern u(t) should be available to the net, and it is assumed stored in the short term memory of the neural net. Therefore, the optimal memory depth of the neural net is T time steps. For nets with feedforward delay elements it is easy to design a memory depth of T steps (a tapped delay line with T delays) [Waibel, 1989]. However for nonlinear feedback systems the memory depth is not easily controlled.

Let us review the available methods for recurrent neural net design. In fact, the only formal recurrent neural net design method is based on Lyapunov stability theory [Hopfield, 1982; Johnson, 1982]. This design method has been established in the control theory community and leads to recurrent networks with guaranteed global stability. However, this method does not provide insight into the memory capacity of certain architectures.

An alternative, although hardly a formal method, is to use one of the available restricted recurrent net architectures that have proven to be successful for particular sequential processing tasks. Examples of specific recurrent net architectures for temporal processing are Elman nets, Jordan nets, Mozer nets, local-feedback global-feedforward networks, fully recurrent nets, Giles’ high order recurrent nets, gamma nets, and several others. Each of these networks reflect a different design trade-off, and the sheer number of different recurrent net architectures indicates that the field is very experimental and ad hoc.

This paper introduces an alternative approach for recurrent net design. In this approach the recurrent net consists of a memory-less feedforward topology which communicates with a (adaptive) parameterized linear *memory filter* (Figure 1). The next section introduces the concept of a (neural) memory filter.

2 Neural Memory Filters

Definition (memory filter). *A memory filter is a linear system with an impulse response g(t) obeying the following two conditions.*

1. g(t) is causal, that is g(t)=0 for t<0.

2. g(t) is normalized in the sense that $\sum_{t=0}^{\infty} |g(t)| = 1$.

Theorem. *A memory filter is bounded-input-bounded-output (BIBO) stable.*

Proof. *A filter g(t) is BIBO stable iff* $\sum_{t=-\infty}^{\infty} |g(t)| < \infty$ *[Oppenheim and Schafer, 1975]. This condition is satisfied according to condition 2 in the definition of the memory filter.*

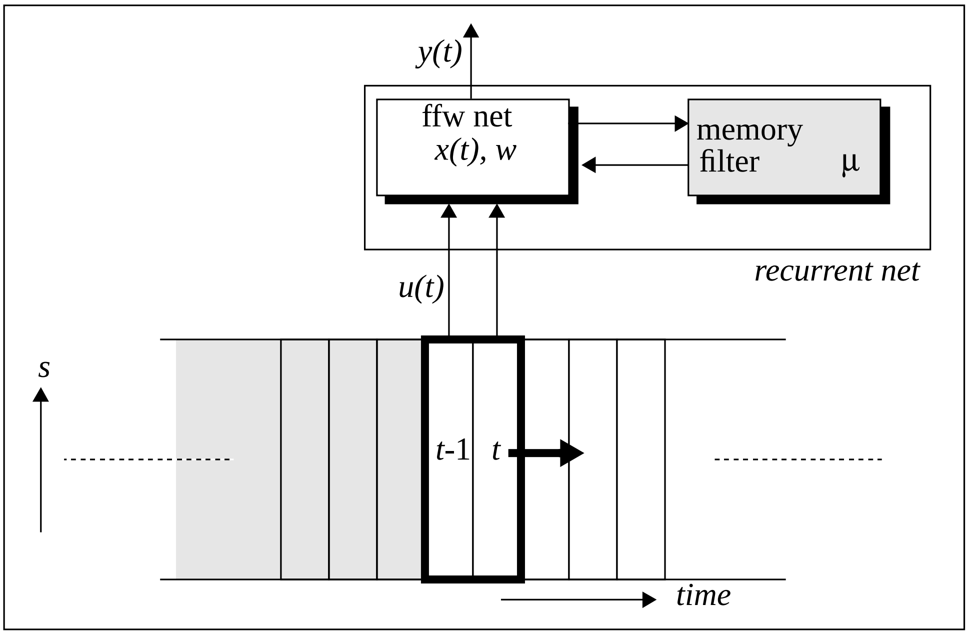

Figure 1 Recurrent net design (feedforward net extended with neural memory filter) for time varying signal processing.

The use of the memory filter is the following. A signal *v(t)* is called a *memory state* (of *the input u(t)*) if $v(t) = \sum_{k \le t} g(t-k)u(k)$ where *g(t)* is the impulse response of a memory filter. The problem in designing memory filters is the choice of g(t) to represent well properties of the input signal, relevant to the classification task. Next, some examples of memory filters are presented.

2•1 the tapped delay line

An important class of memory filters is of course based on the (non-dispersive) delay operator. A *K*-th order tapped delay line, the memory structure for Time Delay Neural Network (TDNN) [Waibel, 1989], can be regarded as a single-input-*K*-output filter with impulse responses $g_k(t) = \delta(t-k)$. The transfer function in the *z*-domain is given by $G_k(z) = z^{-k}$.

The *memory depth D* of a *K*-th order delay line is $D = K$. The *resolution R*, the number of memory state variables per time step delay is *R* = 1.

2•2 the leaky integrator

The impulse response of the leaky integrator is given by

$$g(t) = (1-\mu)\mu^{t-1}u(t-1) \qquad , |\mu| < 1 \quad , \tag{1}$$

where μ, $|\mu| < 1$, is an adaptive parameter and *u(t)* is the unit step function. In the neural net literature, leaky integrators are sometimes referred to as *context nodes* [Elman, 1990] or *memory neurons* [Poddar and Unnikrishnan, 1991].

There are several ways to estimate the “reach” or memory depth of a leaky integrator. Here we will use the mean value of the time variable (the first moment, considering g(t) the probability density function) as defined by

$$D \equiv \sum_{t=0}^{\infty} tg(t) = Z\{tg(t)\}\big|_{z=1} = -z\frac{dG(z)}{dz}\bigg|_{z=1} = \frac{1}{1-\mu} \quad , \tag{2}$$

where $G(z) \equiv \sum_{t=0}^{\infty} g(t) z^{-t}$ is the z-transform of *g(t)*. For the leaky integrator the z-transform evaluates to

$$G(z) = \frac{1-\mu}{z-\mu} . \tag{3}$$

The memory depth of the leaky integrator increases for increasing values of μ (<1). The case $\mu = 0$ reduces the integrator to a unit delay operator z^{-1}, whereas $\mu = 1-\varepsilon$ with ε very small leads to a very deep memory. The cost of increasing memory depth is a reduced temporal resolution. The temporal resolution for the leaky integrator is $R = 1/D = 1-\mu$.

2•3 the gamma memory filter

The gamma memory filter consists of a cascade of *K* leaky integrators with the same parameter μ [deVries and Principe, 1992Figure 2], as depicted in Figure 2. The gamma filter unifies the tapped delay line and the leaky integrator into a single structure, with two parameters, the order and μ. The transfer function to the *k*-th tap is given by

$$G_k(z) \equiv \frac{V_k(z)}{U(z)} = \left(\frac{1-\mu}{z-\mu}\right)^k \tag{4}$$

The mean memory depth for a *K*-th order gamma memory evaluates to

$$D = -z \frac{dG_K(z)}{dz}\bigg|_{z=1} = \frac{K}{1-\mu} . \tag{5}$$

Thus, for the same μ, the gamma memory depth is *K* times larger than for the leaky integrator. Yet, the resolution for the gamma memory is similar to the leaky integrator,

$$R = \frac{K}{\left(\frac{K}{1-\mu}\right)} = 1-\mu \tag{6}$$

Note that all three memory structures (tapped delay line, leaky integrator and gamma filter) obey the relation

$$\textit{order} = \textit{depth} \text{ x } \textit{resolution (K=DR).} \tag{7}$$

Therefore, for μ=1 the gamma memory defaults to the tapped delay line, and for K=1, it defaults to the leaky integrator. The distinct advantage of the gamma structure is that the recurrent parameter μ can be adapted using gradient descent in the same way as filter weights. The Wiener Hopf equations can still be analytically solved in the frequency domain [Principe and deVries, 1993]. Effectively, the memory depth is optimized for the classification task. Gamma nets (an additive model enhanced with a gamma memory) have been applied in diverse temporal processing applications [Principe et al, 1992].

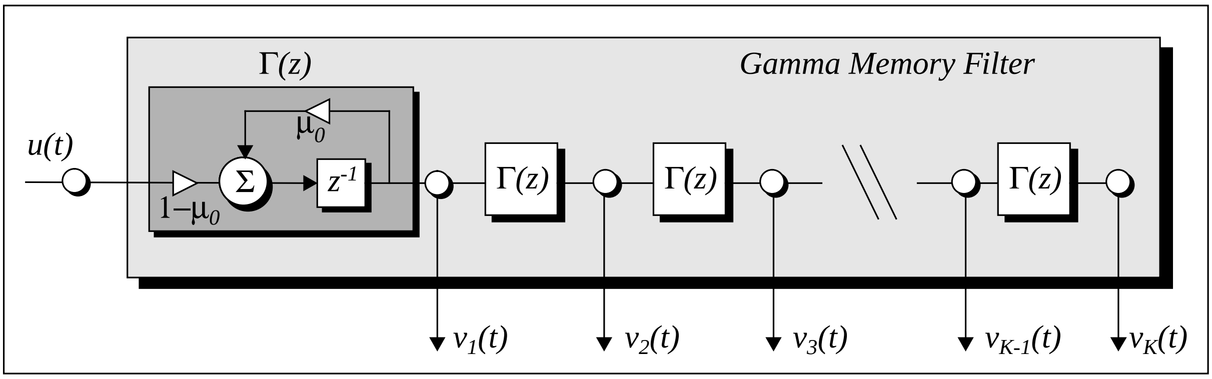

Figure 2 The Gamma memory filter.

2•4 The Time-Alignment Filter

The gamma filter could be compared to a homogeneous rubber band that spans the delay domain. By stretching the band (increase μ) the depth of the memory increases, but as the band remains homogeneous distances between the taps increase proportionally for all taps. Thus, resolution of the gamma filter is the same at each tap. For some applications that involve nonuniform warping of the time axis, an adaptive *tap-dependent* resolution is needed. In particular for speech signals, variations in speech rate create a necessity for more freedom in resolution.

The *time-alignment (TA) filter* presented next provides this kind of memory structure (Figure 3). The backbone of a TA filter is a gamma memory structure, parametrized by μ_0, which controls the depth and resolution of the delay line. *Tap-dependent modulations* of depth and resolution are implemented by variation of the parameters μ_i, i = 1,...,*K*. The connection pattern between $\tilde{u}(t)$ and *v(t)* defines the boundaries of the tap-dependent modulations. Note that for $\{\mu_0, \mu_1, ..., \mu_K\} = \{0\}$, the TA filter reduces to a tapped delay line and for $\{\mu_1, \mu_2, ..., \mu_K\} = \{0\}$ the TA filter becomes a gamma filter.

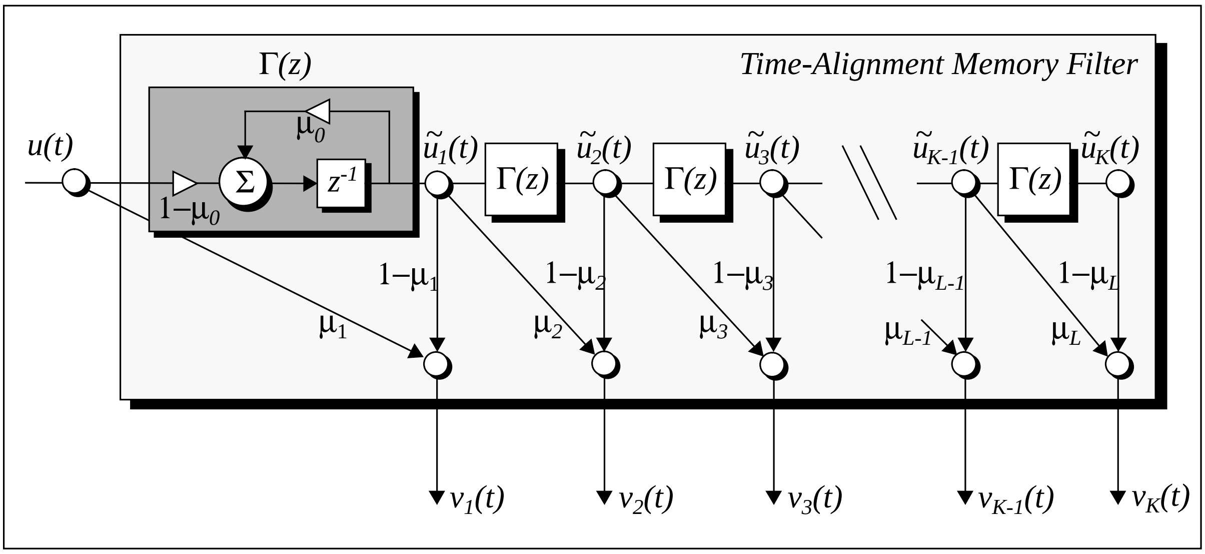

Figure 3 The time-alignment memory filter.

The transfer function at each tap in the TA filter is given by

$$G_k(z) \equiv \frac{V_k(z)}{U(z)} = \Gamma^{k-1}(z)[(1-\mu_k)\Gamma(z)+\mu_k] \qquad \text{for k = 1,...,K} \tag{8}$$

The memory depth at tap k is

$$D_k = -z\frac{dG_k(z)}{dz}\bigg|_{z=1} = \frac{k-\mu_k}{1-\mu_0} \tag{9}$$

and the resolution at tap k is given by

$$R_k \equiv \frac{1}{D_k - D_{k-1}} = \frac{1-\mu_0}{1-\mu_k+\mu_{k-1}} = \frac{1-\mu_0}{1-\Delta_k\mu_k} \tag{10}$$

where $\Delta_k\mu_k \equiv \mu_k - \mu_{k-1}$ .For instance, the parametrization $\{\mu_0, \mu_1, \mu_2, \mu_3, \mu_4\} = \{0, 0, 1, 1, 0\}$ leads to a tapped delay line with unequal tap delays. This feature is very interesting when dealing with speech where sometimes signal values change slowly (vowels) and sometimes rapidly (consonants). Our goal is to adapt μ_i using gradient descent on the classification mean square error (i.e. difference between desired class and actual neural net output)

3 Discussion

Memory structures are of paramount importance for the classification of time varying signals with artificial neural networks. Several memory structures are presently being utilized, but a coherent theory for the design of memory structures was lacking. We have proposed the gamma memory as an unifying model for linear memories. It includes as special cases the most significant linear memories presently being used in the neural network literature. The framework also leads to the definition of memory parameters, and shows the clear trade-off between memory depth and resolution. The time scale in a gamma memory is controlled by μ, i.e. the time is uniformly warped. This memory parameter μ, which controls memory depth, and the filter weights can be analytically computed solving an extension of the Wiener-Hopf equations. Therefore, we can say that the gamma memory uniformly warps time optimally for the classification task. We showed that the gamma memory principle can be extended to create more versatile memory structures with nonuniform time warping scales, that yield promise in dealing with speech recognition problems. Another interesting enhancement of the gamma memory concept for system identification is the design of tap transfer functions that have complex poles, and will therefore yield frequency sensitive memories (Silva et al, 1992).

4 Acknowledgments

This work was partially supported by NSF grant # ECS-9208789, to Jose Principe.

5 References

deVries B. and Principe, J., “The gamma model - A new neural model for temporal processing”, *Neural Networks*, vol 5, pp565-576, 1992.

Elman J.L, Finding structure in time, *Cognitive Science 14*, 179-211, 1990.

Hopfield J.J., Neural networks and physical systems with emergent collective computational abilities, *Proceedings of the national academy of sciences 79*, USA, 2554-2558, 1982.

Johnson C.R., The common parameter estimation basis of adaptive filtering, identification, and control, *IEEE trans. on ASSP 30-4*, 587-595, 1982.

Morgan D.P. and Scofield C.L., *Neural Networks and Speech Processing*, Kluwer Academic Publishers, Boston, MA, 1991.

Oppenheim A.V. and Schafer R.W., *Digital signal processing*, Prentice-Hall, Inc., Englewood, NJ, 1975.

Poddar P. and Unnikrishnan K.P., Non-linear prediction of speech signals using memory neuron networks, *Proceedings of the 1991 IEEE workshop on neural networks for signal processing*, 395-404, 1991.

Principe J., deVries B., Kuo J., Oliveira P., “Modeling applications with the focused gamma network”, in *Advances of Neural Information Processing Systems*, NIPS 4, pp 143-150, Morgan Kaufmann

Principe J., deVries, Oliveira P., “The gamma filter - A new class of adaptive IIR filters with restricted feedback”, *IEEE Trans. Signal Processing*, in Press.

Silva T., Oliveira P., Principe J., deVries B., “Generalized feedforward filters with complex poles”, Proc. of Second IEEE Conf. Neural Networks for Signal Processing, 503-510, 1992.

Waibel A., Hanazawa T., Hinton G., Shikano K., Land K.,”Phoneme recognition using time-delay neural networks”, *IEEE Trans. ASSP*, vol 37, #3, 328-339, 1989.