Appears in IEEE Workshop on Neural Networks for Signal Processing VII, Eds. J. Principe, L. Giles, N. Morgan, E. Wilson. pp. 362–367. IEEE Press, 1997.

The Gamma MLP – Using Multiple Temporal Resolutions for Improved Classification

Steve Lawrence[1], Andrew D. Back[2], Ah Chung Tsoi[3], C. Lee Giles[1]

[1] NEC Research Institute, 4 Independence Way, Princeton, NJ 08540
[2] The Institute of Physical and Chemical Research (RIKEN), Japan
[3] Faculty of Informatics, University of Wollongong, NSW 2522 Australia

Abstract

We have previously introduced the Gamma MLP which is defined as an MLP with the usual synaptic weights replaced by gamma filters and associated gain terms throughout all layers. In this paper we apply the Gamma MLP to a larger scale speech phoneme recognition problem, analyze the operation of the network, and investigate why the Gamma MLP can perform better than alternatives. The Gamma MLP is capable of employing multiple temporal resolutions (the temporal resolution is defined here, as per de Vries and Principe, as the number of parameters of freedom (i.e. the number of tap variables) per unit of time in the gamma memory – this is equal to the gamma memory μ parameter as detailed in the paper). Multiple temporal resolutions may be advantageous for certain problems, e.g. different resolutions may be optimal for extracting different features from the input data. For the problem in this paper, the Gamma MLP is observed to use a large range of temporal resolutions. In comparison, TDNN networks typically use only a single temporal resolution. Further motivation for the Gamma MLP is related to the "curse of dimensionality" and the ability of the Gamma MLP to trade off temporal resolution for memory depth, and therefore increase memory depth without increasing the dimensionality of the network. The IIR MLP is a more general version of the Gamma MLP – however the IIR MLP performs poorly for the problem in this paper. Investigation suggests that the error surface of the Gamma MLP is more suitable for gradient descent training than the error surface of the IIR MLP.

1 Introduction

Machine learning models used for speech recognition are required to account for a high degree of variability in the data (e.g. acoustic variability, *within-speaker* variability, *across-speaker* variability, and phonetic variability). For phoneme recognition, methods of addressing these variabilities include using larger datasets and using models which take into account greater context of the acoustic signal. However, taking into account greater context typically leads to larger models. The amount of training data required for accurate estimation of class distributions can increase significantly when the input dimensionality increases (cf. the "curse of dimensionality" [6])[1]. As the complexity of the desired target function for a given problem increases while the amount of data remains constant, it becomes increasingly problematic to estimate the target function from finite data due to the ill-posed nature of the problem – many of the models which fit the training data closely do not generalize well to unseen data. In order to reduce the difficulty with trying to approximate a function which is too complex for the available data, we often consider looking for a hierarchical solution where initial layers extract features which identify higher level attributes of the data which enhance generalization. These features can be extracted manually, or automatically. The Gamma MLP considers a transformation for the inputs to each node and aims to optimize the transformation for each node individually in order to improve performance. The process can be thought of as automatic feature extraction (if the optimal transformations were known beforehand then those transformations could be used to extract new features from the data).

2 The Gamma Filter

Infinite Impulse Response (IIR) filters have a significant advantage over Finite Impulse Response (FIR) filters in signal processing: the length of the impulse response is uncoupled from the number of filter parameters. The length of the impulse response is related to the memory depth[2] of a system, and hence IIR filters allow a greater memory depth than FIR filters of the same order. However, IIR filters are not widely used in adaptive signal processing [9]. This may be attributed to the fact that a) there may be instability during training and b) the gradient descent training procedures are not guaranteed to locate the global optimum in the possibly non-convex error surface [11].

[1] Additionally, increases in the "complexity" of the desired target function may make gradient descent optimization more difficult – training algorithms may take longer to converge or become "stuck" in local minima or "plateaus" which are increasingly poor compared to the global optimum.

[2] A greater memory depth implies that the model can retain past information for a longer time.

The use of *gamma* filters as a memory structure at the input of an otherwise standard MLP network was proposed by de Vries and Principe [5]. The gamma filter, a special case of an IIR filter, is designed to retain the uncoupling of memory depth to the number of parameters provided by IIR filters, but to have simple stability conditions. The output of a neuron in a multilayer perceptron is computed using[3] $y_k^l = f\left(\sum_{i=0}^{N_{l-1}} w_{ki}^l y_i^{l-1}\right)$. The addition of short term memory with delays was considered by de Vries and Principe [5]: $y_k^l = f\left(\sum_{i=0}^{N_{l-1}} \sum_{j=0}^{K} g_{kij}^l(t-j) y_i^{l-1}(t-j)\right)$ where $g_{kij}^l(t) = \frac{\mu_{ki}^l}{(j-1)!} t^{j-1}\, e^{-\mu_{ki}^l t}$, $j = 1, 2, \ldots, K$. The depth of the memory is controlled by μ, and K is the order of the filter. For the discrete time case, de Vries and Principe [5] obtain the following recurrence relation:

$$z_j(t) = \begin{cases} x(t), & j = 0 \\ (1-\mu) z_j(t-1) + \mu z_{j-1}(t-1), & j = 1, 2, \ldots, K \end{cases} \quad (1)$$

where $x(t)$ is the filter input and $z_j(t)$ are the filter outputs. For $\mu < 1$ the gamma filter may be considered as a low pass filter. For $\mu = 1$, the memory is a tapped delay line corresponding to the memory structure in an FIR MLP (An MLP where the weights are replaced by FIR filters and optional gain terms [2]) or a TDNN. For $\mu < 1$ the gamma memory structure implements a tapped dispersive delay line where the degree of dispersion is controlled by μ.

de Vries and Principe [9] define the temporal resolution, R, of a gamma memory structure as the number of parameters of freedom (i.e. the number of tap variables) per unit of time in the filter memory: $R = K/D = \mu$ where D is the memory depth of the structure (the temporal mean value of the impulse response of the last tap) [10]: $D = K/\mu$. When $\mu = 1$, the memory depth is equal to the order of the memory, K. The memory depth increases when $\mu < 1$, and the temporal resolution decreases, i.e. the gamma memory can trade resolution for memory depth. Therefore the gamma memory can be used to create models which can take into account greater context with fewer parameters (without resorting to the use of a single low temporal resolution) in comparison to TDNN or FIR MLP models.

[3] where y_k^l is the output of neuron k in layer l, N_l is the number of neurons in layer l, w_{ki}^l is the weight connecting neuron k in layer l to neuron i in layer $l-1$, $y_0^l = 1$ (bias), and f is commonly a sigmoid function.

3 The Gamma MLP

3.1 Motivation

The *focused gamma network* which uses the gamma memory as a preprocessing layer for a standard MLP has been proposed by de Vries and Principe [5]. This network allows for the use of only one temporal resolution per input. However, it may be desirable to use multiple temporal resolutions (e.g. different resolutions may be optimal for extracting different features or for classifying different phonemes). The Gamma MLP is similar to a standard MLP except every synapse contains a gamma memory structure and a gain factor. The temporal resolution of the memory in each synapse is adjusted separately. Therefore, in contrast with the focused gamma network, the Gamma MLP is able to use multiple temporal resolutions. Additionally, the Gamma MLP can contain gamma memory structures in all layers of the network.

Other motivation for the Gamma MLP can be seen with comparison to TDNN, FIR MLP and IIR MLP (An MLP where the weights are replaced by IIR filters and optional gain terms [1]) models. In comparison to the TDNN and FIR MLP models, the Gamma MLP may provide improved performance because it allows temporal resolution to be traded for memory depth, i.e. for a system of given dimensionality, the Gamma MLP can employ filters with a greater memory depth. Additionally, in comparison with the IIR MLP, the Gamma MLP may be significantly easier to train, which is discussed further in section 5.

3.2 Definition

Definition 1 A Gamma MLP with L layers excluding the input layer $(0, 1, ..., L)$, gamma filters of order K, and $N_0, N_1, ..., N_L$ neurons per layer, is defined as:

$$y_k^l(t) = f\left(x_k^l(t)\right) \tag{2}$$

$$x_k^l(t) = \sum_{i=0}^{N_{l-1}} c_{ki}^l(t) \sum_{j=0}^{K} w_{kij}^l(t) z_{kij}^l(t) \tag{3}$$

$$z_{kij}^l(t) = \begin{cases} (1-\mu_{ki}^l(t)) z_{kij}^l(t-1) + \mu_{ki}^l(t) z_{ki(j-1)}^l(t-1), & 1 \le j \le K \\ y_i^{l-1}(t), & j = 0 \end{cases} \tag{4}$$

where $y_k^l(t)$ is the output of neuron k in layer l at time t, c_{ki}^l = synaptic gain, $\mathrm{f}(\alpha) = \tanh(\alpha) = (e^{\alpha/2} - e^{-\alpha/2})/(e^{\alpha/2} + e^{-\alpha/2})$, $k = 1, 2, ..., N_l$ (neuron index), $l = 0, 1, ..., L$ (layer), and $z_{kij}^l|_{i=0} = 1$, $w_{kij}^l|_{i=0, j\neq 0} = 0$, $c_{kij}^l|_{i=0} = 1$ (bias).

□

A Gamma MLP is defined as a multilayer perceptron where every synapse contains a gamma filter and a gain term (introduced in [7]), as shown in the definition above. The Gamma MLP is therefore a special case of the IIR MLP [1]. The motivation behind the inclusion of the gain term is discussed in section 5. A separate μ parameter is used for each filter. Gradient descent update equations for the Gamma MLP are given in [7]. In practice, it is often desirable to restrict the Gamma MLP structure by using Gamma filter only in the first layer and/or not using the synaptic gain terms (c_{ki}^l), as is also the case for FIR and IIR MLP networks.

4 Phoneme Recognition

4.1 Task Details

Our data consists of the "sa" sentences spoken by male members of demographic region 3 in the TIMIT database. There are 79 speakers. The problem is therefore speaker independent phoneme prediction. The speakers in the training and test sets do not overlap.

The raw speech data was preprocessed into a sequence of frames using PLP. The analysis window (frame) was 20 ms. Each succeeding frame overlapped with the preceding frame by 10 ms. 9 PLP coefficients plus the signal power were extracted and used as features describing each frame of data. The difference between the current and previous frames was added to the input vectors, as is commonly done [4]. Periods of silence before and after the sentences were reduced to two frames in order to limit any skew of the results caused by a disproportionate percentage of silence frames.

The models had 40 outputs corresponding to the 40 phonemes[4]. The FIR and gamma filter orders were 4 (5 taps), and the TDNN model had an input window of 5 steps in time. The training set contained 10,000 frames, the test set and validation sets contained 5,000 frames, and the networks had 40 hidden nodes. The networks were trained for 200,000 updates. We used standard backpropagation with stochastic update. The $\tanh$ activation function was used. A "search then converge" learning rate schedule was used with an initial learning rate of 0.1 for the μ parameters and 0.2 for all other parameters.

[4]The TIMIT allophones were converted to the standard 40 phoneme set [8].

4.2 Results

Results are presented for frame level phoneme recognition, i.e. for each frame the recognizer predicts the current phoneme. A few observations regarding the expected results: guessing would result in 97.5% error, coarticulation[5] makes the task difficult, and the possibility of zero error would not be expected due to inevitable difficulty and errors in the phoneme labelling.

The frequencies of the forty classes varies significantly, and it was found that all models had a tendency to "ignore" the rarer phonemes [3] due to biases inherent in the neural network architecture and training algorithm. We therefore employed a scaling technique whereby weight updates are scaled on a class by class basis. The amount of scaling is varied using a control parameter, c_s, from none ($c_s = 0$) to scaling according to the prior probabilities of the classes ($c_s = 1$). Yaeger et al. have recently introduced a very similar technique which they call "frequency balancing" [12].

Reporting results in terms of the percentage of correct classifications can be misleading when the frequency of the individual classes varies significantly (e.g. a relatively low error rate may be achieved by a network which ignores low frequency classes). For this reason, results are reported here in terms of the MSSE which is defined as: MSSE $= \frac{1}{N_c} \sum_{i=1}^{N_c} (1 - S_i)^2$ where N_c = the number of classes and S_i = the sensitivity of class i. The sensitivity of a class is defined as the proportion of events labelled as that class which are correctly detected. This criterion was chosen because each class is given equal importance and the square causes lower individual sensitivities to be penalized more (e.g. for a two class problem, class sensitivities of 100% and 0% produce a higher MSSE than sensitivities of 50% and 50%).

Figure 1 shows the results for the Gamma MLP, FIR MLP, and TDNN networks. The degree of scaling, c_s, was varied from 0 to 1. Five trials were performed in each case. The FIR MLP and Gamma MLP networks contained filters in both layers. The Gamma MLP contained synaptic gains, however the FIR MLP was found to perform significantly better without the synaptic gains for this problem. Scaling with $c_s = 0.75$ resulted in the best performance for each of the networks and, therefore, scaling with $c_s = 0.75$ was used for the later results.

Results for the IIR MLP are not shown because it was not possible to obtain significant convergence. Theoretically, the IIR MLP model is the most powerful model used here (in the sense that it can represent a greater variety of computational structures than the other networks with the same number of hidden nodes). In particular, the Gamma MLP is a special case of the IIR MLP. Although the IIR MLP is prone

[5] Coarticulation refers to changes in the way a speech segment is articulated depending on previous (backward coarticulation) and following segments (forward coarticulation).

to stability problems, the stability of the model can and was controlled in the simulations performed here (by reflecting poles that move outside the unit circle back inside). The most obvious hypothesis for the difficulty in training the model is related to the error surface and the nature of gradient descent. It is expected that the error surface of the IIR MLP presents greater difficulty to gradient descent optimization. This is discussed further in the next section.

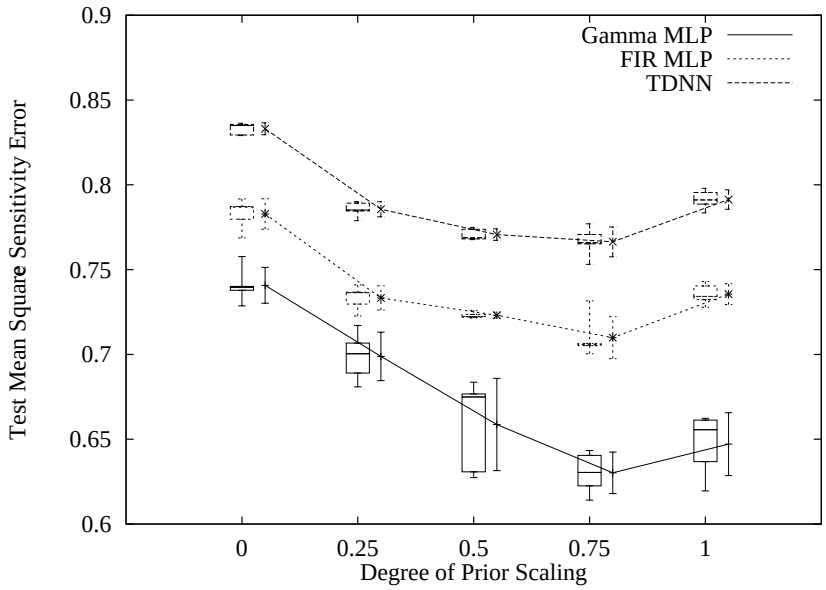

Figure 1. Test MSSE results as the degree of scaling is modified. The best error corresponds to a scaling degree of 0.75 for each network type. At each point, box-whiskers plots are shown on the left and the mean plus and minus one standard deviation is shown on the right. Five trials were performed in each case.

5 Discussion

The Gamma MLP may perform better than the standard TDNN and the FIR MLP for speech recognition because the gamma filtering operation allows processing the input data using multiple temporal resolutions. The Gamma MLP can therefore account for more past history of the signal for a system of a given order (without resorting to the use of a single low temporal resolution). Figure 2 shows the distribution of the gamma μ parameters in a typical trained Gamma MLP. It can be seen that

a range of μ parameters, and therefore a range of temporal resolutions, is employed by the network.

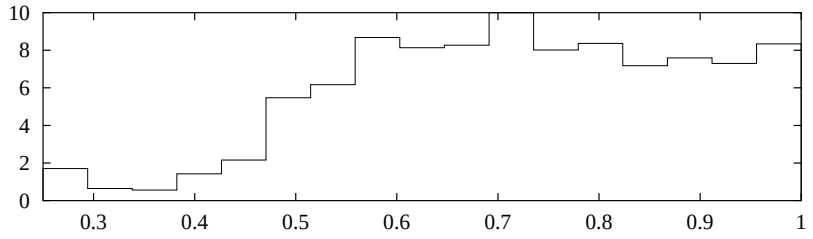

Figure 2. The final distribution of the gamma μ parameters for a sample Gamma MLP.

The Gamma MLP often performs better when using the synaptic gain terms. This improvement may be considered non-intuitive to many – the synaptic gains add degrees of freedom, but no additional representational power. However, the error surface will be different in each case, and results indicate that the surface for the synaptic gains case can often be more amenable to gradient descent.

For the problem considered here, the Gamma MLP performs significantly better than the IIR MLP although the Gamma MLP is a special case of the IIR MLP. It is reasonable to believe that the IIR MLP could perform as well as, or possibly better than, the Gamma MLP, but in practice it is difficult to make it do so for the problem considered here. Figure 3 shows sample plots of the error surface for Gamma and IIR MLP networks. In order to reduce computational expense and use networks with fewer parameters to aid visualization, a simpler task has been chosen. The task is Mackey-Glass prediction using networks that contain only five hidden nodes (the order of the filters was 4, the initial learning rate was 0.1, the training, test, and validation sets contained 500 points, and 100,000 stochastic updates were performed in each case). Even with such small networks, the error surface has many dimensions making visualization difficult. Each plot in the figures is with respect to two randomly chosen dimensions. In each case, the center of the plot corresponds to the values of the parameters after training and the range of each parameter on the plot is 8. The NMSE was evaluated at 225 points equally spaced in a grid. For the IIR MLP, a greater percentage of "flat spots" and complex surfaces can be observed. On average, the error surface for the IIR MLP appears to be less suitable for gradient descent optimization, reinforcing the conclusion that the poorer performance of the IIR MLP is due to optimization being more difficult. Hence, in using the Gamma MLP instead of the IIR MLP, we are trading off computational capacity for easier training. The test NMSE results for 20 simulations each using these networks show that the best performing IIR MLP was only slightly worse than the best performing Gamma MLP. However, the Gamma MLP was significantly

better on average (NMSE of 0.0341 versus 0.185 for the IIR MLP).

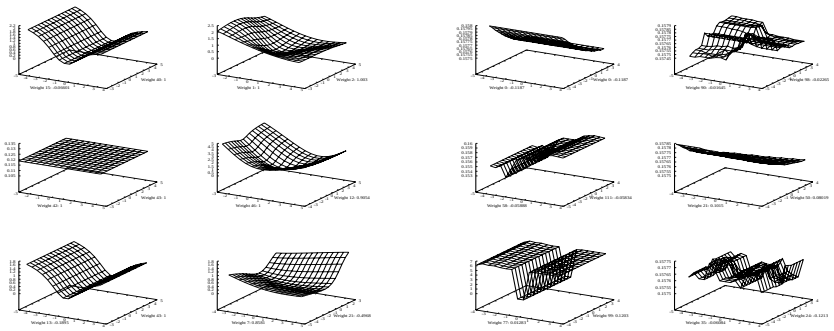

Figure 3. Error surface plots for a sample Gamma MLP (left two columns) and a sample IIR MLP (right two columns). Each plot is with respect to two randomly chosen dimensions. In each case, the center of the plot corresponds to the values of the parameters after training. The z-axis scale varies from plot to plot in order to show the qualitative aspects of the surface (the plots only cover variation in two dimensions and are only plotted around one point in weight space, therefore quantitative conclusions should be drawn from the final NMSE results). From many of these plots we have observed that there is a greater percentage of "flat spots" and complex surfaces for the IIR MLP.

6 Conclusions

We have applied the Gamma MLP to a speech phoneme recognition problem, analyzed the operation of the network, and investigated why the Gamma MLP can perform better than alternatives. The Gamma MLP is capable of employing multiple temporal resolutions, which may be advantageous for certain problems, e.g. different resolutions may be optimal for extracting different features from the input data. For the problem in this paper, the Gamma MLP is observed to use a large range of temporal resolutions. In comparison, TDNN networks typically use only a single temporal resolution. The Gamma MLP is able to trade off temporal resolution for memory depth, and therefore increase memory depth without increasing the dimensionality of the network (or using a single low temporal resolution). The IIR MLP is a more general version of the Gamma MLP – however the IIR MLP performed poorly for the problem in this paper. Investigation suggested that the error surface of the Gamma MLP is more suitable for gradient descent training than the error surface of the IIR MLP.

References

[1] A.D. Back. *New Techniques for Nonlinear System Identification: A Rapprochement Between Neural Networks and Linear Systems*. PhD thesis, Department of Electrical Engineering, University of Queensland, 1992.

[2] A.D. Back and A.C. Tsoi. FIR and IIR synapses, a new neural network architecture for time series modeling. *Neural Computation*, 3(3):375–385, 1991.

[3] Etienne Barnard, R.A. Cole, and L. Hou. Location and classification of plosive constants using expert knowledge and neural-net classifiers. *Journal of the Acoustical Society of America*, 84 Supp 1:S60, 1988.

[4] H.A. Bourlard and N. Morgan. *Connnectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, Boston, MA, 1994.

[5] B. de Vries and J.C. Principe. The Gamma model – a new neural network for temporal processing. *Neural Networks*, 5(4):565–576, 1992.

[6] J.H. Friedman. Introduction to computational learning and statistical prediction. Tutorial Presented at Neural Information Processing Systems, Denver, CO, 1995.

[7] Steve Lawrence, A.C. Tsoi, and A.D. Back. The Gamma MLP for speech phoneme recognition. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 785–791. MIT Press, 1996.

[8] Kai-Fu Lee and Hsiao-Wuen Hon. Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1641–1648, 1989.

[9] J.C. Principe, B. de Vries, and P. Oliveira. The gamma filter – a new class of adaptive IIR filters with restricted feedback. *IEEE Transactions on Signal Processing*, 41:649–656, 1993.

[10] J.C. Principe, J.M. Kuo, and S. Celebi. An analysis of the Gamma memory in dynamic neural networks. *IEEE Transactions on Neural Networks*, 5(2):331–337, 1994.

[11] J.J. Shynk. Adaptive IIR filtering. *IEEE ASSP Magazine*, pages 4–21, 1989.

[12] L. Yaeger, R. Lyon, and B. Webb. Effective training of a neural network character classifier for word recognition. In M.C. Mozer, M.I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, Cambridge, MA, 1997. MIT Press.