To appear in G. Tesauro, D.S. Touretzky and T.K. Leen, eds
Neural Information Processing Systems 7, MIT Press Cambridge MA. 1995.

Non-linear Prediction of Acoustic Vectors Using Hierarchical Mixtures of Experts

S.R.Waterhouse **A.J.Robinson**
Cambridge University Engineering Department,
Trumpington St., Cambridge, CB2 1PZ, England.
Tel: [+44] 223 332800, Fax: [+44] 223 332662,
Email: srw1001, ajr @eng.cam.ac.uk
URL: http://svr-www.eng.cam.ac.uk/~srw1001

Abstract

In this paper we consider speech coding as a problem of speech modelling. In particular, prediction of parameterised speech over short time segments is performed using the Hierarchical Mixture of Experts (HME) (Jordan & Jacobs 1994). The HME gives two advantages over traditional non-linear function approximators such as the Multi-Layer Perceptron (MLP); a statistical understanding of the operation of the predictor and provision of information about the performance of the predictor in the form of likelihood information and local error bars. These two issues are examined on both toy and real world problems of regression and time series prediction. In the speech coding context, we extend the principle of combining local predictions via the HME to a Vector Quantization scheme in which fixed local codebooks are combined on-line for each observation.

1 INTRODUCTION

We are concerned in this paper with the application of multiple models, specifically the Hierarchical Mixtures of Experts, to time series prediction, specifically the problem of predicting acoustic vectors for use in speech coding. There have been a number of applications of multiple models in time series prediction. A classic example is the *Threshold Autoregressive model* (TAR) which was used by Tong &

Lim (1980) to predict sunspot activity. More recently, Lewis, Kay and Stevens (*in* Weigend & Gershenfeld (1994)) describe the use of Multivariate and Regression Splines (MARS) to the prediction of future values of currency exchange rates. Finally, in speech prediction, Cuperman & Gersho (1985) describe the Switched Inter-frame Vector Prediction (SIVP) method which switches between separate linear predictors trained on different statistical classes of speech. The form of time series prediction we shall consider in this paper is the *single step prediction* $\hat{\boldsymbol{y}}^{(t)}$ of a future quantity $\boldsymbol{y}^{(t)}$, by considering the previous p samples. This may be viewed as a regression problem over input-output pairs $\{\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}\}_{t=0}^{N}$ where $\boldsymbol{x}^{(t)}$ is the *lag vector* $(\boldsymbol{y}^{(t-1)}, \boldsymbol{y}^{(t-2)}, ..., \boldsymbol{y}^{(t-p)})$. We may perform this regression using standard linear models such as the Auto-Regressive (AR) model or via nonlinear models such as connectionist feed-forward or recurrent networks. The HME overcomes a number of problems associated with traditional connectionist models via its architecture and statistical framework. Recently, Jordan & Jacobs (1994) and Waterhouse & Robinson (1994) have shown that via the EM algorithm and a 2nd order optimization scheme known as Iteratively Reweighted Least Squares (IRLS), the HME is faster than standard Multilayer Perceptrons (MLP) by at least an order of magnitude on regression and classification tasks respectively. Jordan & Jacobs also describe various methods to visualise the learnt structure of the HME via 'deviance trees' and histograms of posterior probabilities. In this paper we provide further examples of the structural relationship of the trained HME and the input-output space in the form of *expert activation plots.* In addition we describe how the HME can be extended to give local error bars or measures of confidence in regression and time series prediction problems. Finally, we describe the extension of the HME to acoustic vector prediction, and a VQ coding scheme which utilises likelihood information from the HME.

2 HIERARCHICAL MIXTURES OF EXPERTS

The HME architecture (Figure 1) is based on the principle of 'divide and conquer' in which a large, hard to solve problem is broken up into many, smaller, easier to solve problems. It consists of a series of 'expert networks' which are trained on different parts of the input space. The outputs of the experts are combined by a 'gating network' which is trained to stochastically select the expert which is performing best at solving a particular part of the problem. The operation of the HME is as follows: the gating networks receive the input vectors $\boldsymbol{x}^{(t)}$ and produce as outputs probabilities $P(m_j|x^{(t)}, \eta_j)$ for each local branch m_j of assigning the current input to the different branches, where η_j are the gating network parameters. The expert networks sit at the leaves of the tree and each output a vector $\hat{\boldsymbol{y}}_j^{(t)}$ given input vector $\boldsymbol{x}^{(t)}$ and parameters θ_j. These outputs are combined in a weighted sum by $P(m_j|x^{(t)}, \eta_j)$ to give the overall output vector for this region. This procedure continues recursively upwards to the root node. In time series prediction, each expert j is a linear single layer network with the form:

$$\hat{\boldsymbol{y}}_j^{(t)} = \theta_j \boldsymbol{x}^{(t)}$$

where θ_j is matrix and $\boldsymbol{x}^{(t)}$ is the lag vector discussed earlier, which is identical in form to an AR model.