THE ACCURATE ESTIMATION OF ARTICULATORY SYNTHESISER PARAMETERS THROUGH REDUCING THE DEGREE OF SATURATION IN A NURAL NETWORK HIDDEN LAYER

K M Curtis[*] & H Altun[≅]
RISN Group, University of Nottingham
[*]*Department of Electrical & Electronic Engineering, University of Nottingham, Nottingham, NG7 2RD, UK*
[≅]*Department of Electrical & Electronic Engineering, Nigde University, Nigde, Turkey*

ABSTRACT

A new method is proposed to correctly estimate the parameters of an articulatory speech synthesiser using a MLP neural network. This is achieved through modifying the statistical characteristic of the acoustic input pattern vectors in order to prevent the activation level of the hidden nodes from approaching saturation. The technique results in considerably faster neural learning and a more accurate estimation of the articulatory synthesiser parameters.

1. INTRODUCTION

Articulatory synthesising of speech is reported as being very capable of producing natural speech at very low bit rates [1] and is likely to be the ultimate solution to the production of natural intelligible speech [2]. However, inferring the articulatory parameters from the acoustic parameter of speech is a difficult problem to solve due to the ill-posed, one-to-many relationship between the acoustic and articulatory domain data. The use of a neural network (NN) to carry out the acoustic-to-articulatory mapping has been reported on. However, apart from their computational advantages, NN's have not proved superior to other mapping techniques, such as table-lookup [1]. Improving the mapping requires enhanced neural learning. Researchers have striven towards improving the learning process of the backpropagation (BP) algorithm, which is the most widely used learning algorithm, through modifying the algorithm itself [3] or through optimising the parameters of the algorithm [4-5].

This paper shows how neural learning can be improved through modifying the statistical characteristic of the acoustic input pattern vectors, to achieve accurate acoustic-to-articulatory mapping, based on our earlier results from an analytical and statistical evaluation of neural learning in a MLP NN [6].

2. AN MLP NEURAL NETWORK FOR ACOUSTIC-TO-ARTICULATORY MAPPING

In order to investigate the effect of the statistical characteristic of the acoustic input pattern vector on neural learning in the MLP NN, let a NN with a single hidden layer have a vector space with dimension S, and an input vector in this space have a dimension of K $\mathbf{x_i}^{(s)}=[x_1^{(s)},x_2^{(s)},..,x_K^{(s)}]^T$, a hidden layer output vector with dimension L, $\mathbf{x_h}^{(s)}=[x_1^{(s)},x_2^{(s)},..,x_L^{(s)}]^T$, and an output vector $\mathbf{x_o}^{(s)}=[x_1^{(s)},x_2^{(s)},..,x_M^{(s)}]^T$ with dimension M. The weighted connection between the input-hidden and hidden-output layers are $\mathbf{W_{ih}}$ and $\mathbf{W_{ho}}$. Training of the MLP NN using the BP algorithm requires a training set which consists of corresponding input and target vectors, $\mathbf{x_i}$ and $\mathbf{t_o}$ respectively. Training continues until a predefined error threshold is met between x_o and $\mathbf{t_o}$ as follows

$$\boldsymbol{x_o}^{(s)} = \boldsymbol{t_o}^{(s)}\ \varepsilon \qquad s= 0,1,..,n$$

The BP algorithm presented above has found widespread use in many different applications**.**

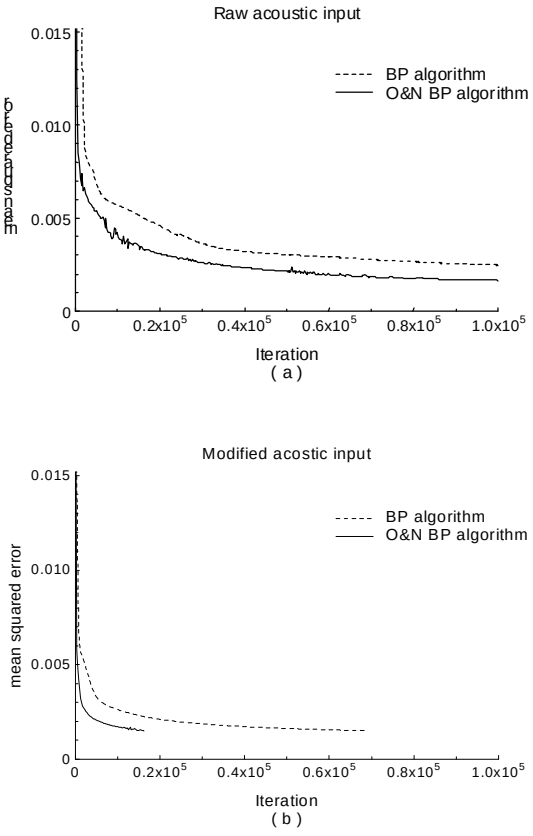

Figure 1. The error curves for standard and O&N BP algorithms.

A training set has been created using a simplified, Kelly-Lochbaum vocal tract (VT) model in order to train the MLP NN,. The assumptions made to simplify the VT implementation is: the VT consists of lossless uniform, concatenated acoustic tubes; the VT consists of rigid walls,

the planar wave propagation is valid. Linear interpolation is applied to the optimised area function of 10 English vowels [7] giving a total of 164 training patterns. Acoustic input pattern vectors $\mathbf{x_i}$ are derived from the transfer function of the VT which has been simulated in the MATCAD software package. The radiation load is approximated by a 1st order IIR filter, setting the reflection coefficient at the boundary of the last section as 0.99 to ensure IIR filter stability.

A NN consisting of a layer of ten inputs, a layer of 18 hidden units and five output units was trained setting the mean squared error threshold to 0.0015. The learning rate and momentum terms were set to 0.01 and 0.3, respectively. Results show that the NN does not converge to the pre-determined error threshold even after 10^5 iterations (Figure 1a). The articulatory target data t_j, in the training set was investigated in the hope of finding a way to improve neural learning. It was found that a considerable amount of the articulatory target data t_j was near the negative saturation region, as seen in Figure 2, which is defined as the activation level of the output nodes x_o, at which derivative of the output node activation x'_o , is reduced by more than 30% of its maximum value of 0.25. This distribution characteristic, of the articulatory target patterns, will be problematic according to the theoretical and empirical results obtained in [8-10]. A very weak weight update signal will be produced by the BP algorithm, due to the small value of the derivative of the output node activation x'_o , when the activation level of the output nodes x_o is forced to approximate the articulatory target values t_j during the process of training.

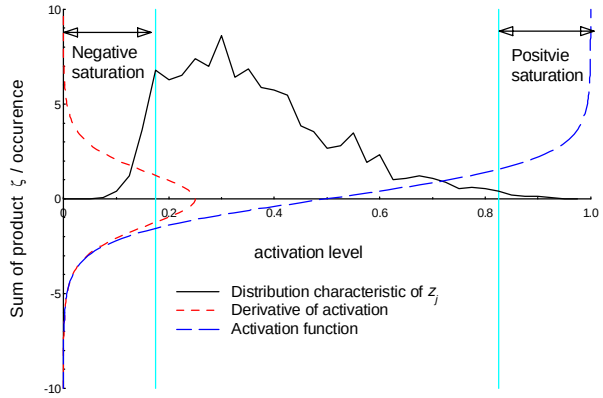

Figure 2. Distribution characteristic of the articulatory data z_j , the activation function and its derivative

3. IMPROVING NEURAL NETWORK MAPPING

A straightforward suggestion to improve neural learning, in acoustic-to-articulatory mapping, is to eliminate the adverse effect of the derivative of the output node activation x'_o . Thus it was decided to use a modified BP algorithm proposed by Ooyen and Nienhuis [11], which eliminates the derivative of the activation x'_o in the weight adjustment formula for the hidden-output layer weights w_{ih} through defining a new error calculation as follows:

$$(1)\quad E = -\sum_{o=1}^{M_z}\left[t_o \ln x_o + (1-t_o)\ln(1-x_o)\right]$$

instead of the standard error calculation in the BP algorithm:

$$(2)\quad E = -\frac{1}{2}\sum_{o=1}^{M_z}(x_o - t_o)^2$$

where o is the index for output layer nodes, and M_z is the number of the nodes in the output layer.

Using identical initial conditions a NN was trained employing the O&N modification to the standard BP algorithm. The result in Figure 1a shows that despite a faster initial approach, the NN shows a slower final approach when learning. The NN could not converge to the pre-determined error threshold within the iteration threshold of 10^5. The main reason for this was found to be that despite the elimination of the derivative of the output node activation x'_o, in the optimisation of the hidden-output layer weights w_{ih}, the weight adjustment for the input-hidden layer interconnections w_{ho} in (3) still contains a term which slows down neural learning. This is the derivative of the activation level of the hidden layer node, x'_h

$$(3)\quad \frac{\partial E}{\partial w_{ho}} = \sum_{o=1}^{N_z}(x_o - t_i)w_{ij}\,x'_h\,x_i$$

This term prevents the NN from optimising its weights efficiently due to its very small values at the saturation regions and inevitably prevents the error at the output from propagating back. Thus, the O&N technique still suffers from the adverse effect of the derivative of the activation function of the hidden layer node x'_h .

3.1. A new method to improve acoustic-to-articulatory mapping

To improve neural learning, the adverse effect of the derivative of the hidden node activation x'_h should be alleviated by preventing the hidden nodes from approaching the saturation region. However, no recognised technique has yet been proposed for this purpose. The reason for this is that the activation level of the hidden layer nodes x_h is determined by the network and depends on the training set and the state of the input-hidden layer weights w_{ih} which have dynamic statistical characteristic due to the optimisation process of the BP algorithm. However, results from our statistical analysis [6] show that the adverse effect of the derivative of the hidden layer nodes x'_h can be alleviated. This is done by modifying the statistical characteristic of the input pattern vectors $\mathbf{x_i}$ to an optimum statistical characteristic, as the possibility of having a saturated hidden node is found to be a function of the statistical characteristic of the input pattern vectors x_i. The optimum statistical values of the input pattern vectors are stated as being a mean value , which is around the middle of the activation domain and has a small standard deviation.

The possibility of the hidden nodes being saturated is given as follows:

$$(4)\quad P[\zeta > \psi] = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\psi} \exp\left(- \frac{(x - m\zeta)^2}{2\sigma_\zeta^2} \right) dx$$

$$(5)\quad m\zeta = LKm_{x_i} m_{w_{ih}}$$

$$(6)\quad \sigma_\zeta^2 = L^2K^2\left(\sigma_{x_i}^2 + m_{x_i}^2\right)\left(\sigma_{w_{ih}}^2 + m_{w_{ih}}^2\right) - (m\zeta)^2$$

where m is the mean value and σ is the standard deviation, ζ stands for the sum-of-product to the hidden layer. K and L are the number of the input and the hidden layer nodes, respectively.

Algorithm	Distribution Characteristic m_x	σ_x	Number of Iterations	Converged Error	Generalisation Error
Standard BP	0.4827	0.5164	It_thr	0.0497	0.1284
O&N BP	0.4827	0.5164	It_thr	0.0408	0.1095
Standard BP	0.5028	0.3946	68,500	Er_thr	0.1034
O&N BP	0.5028	0.3946	16,250	Er_thr	0.0969

Table I. The effect of modifying the acoustic input data distribution on neural learning.
It_thr = 10^5 Er_thr = 0.0015

The statistical values of x_i are pre-processed, into the statistical values given in Table I, then the training is repeated using the standard and the O&N BP algorithms. Both NNs converged within the iteration threshold, after 68500 and 16250 iterations respectively, as can be seen in Figure 1-b. The results show that the objective of elimination the adverse affect of the hidden node saturation has been achieved. The distribution characteristic of the hidden layer activation levels, shown in Figure 3, show that the proposed modification of the acoustic input pattern vectors x_i reduces dramatically the degree of saturation in the hidden layer. As the weight update signals are partly determined by the product of the hidden layer activation level x_h and its derivative in (7) and (8), the degree of saturation of the layer is defined as in (9):

$$(7)\quad \Delta w_{h_o} = \eta x'_o \Delta x_o x_h$$

$$(8)\quad \Delta w_{i_h} = \eta x'_h \sum_o w_{h_o} x'_o \Delta x_o x_i$$

$$(9)\quad \text{degree of saturation} = \frac{\sum occurence_h}{\sum (occurence_h * x_h * x'_h)}$$

The reduction in the degree of saturation is calculated as 10.64%, a decrease from 5.64 to 5.504 for the standard BP algorithm, and as 56.84%, a decrease from 15.94 to 6.88 for the O&N BP algorithm.

The generalisation ability of the NNs have also been tested using 18 unseen patterns and the results are given in Table I. NNs trained with the modified acoustic input pattern vectors are faster to converge and show improved generalisation.

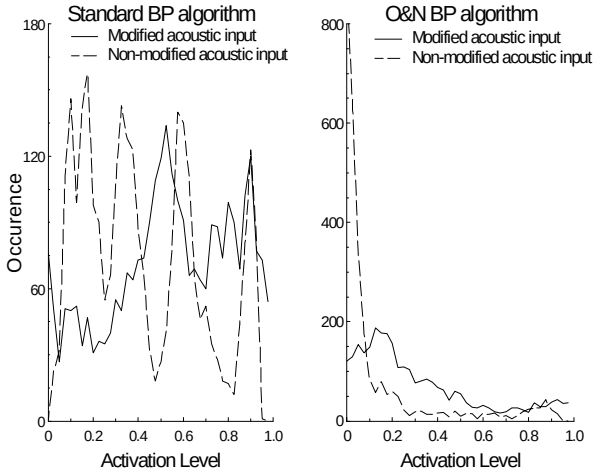

Figure 3. The distribution of the hidden layer activation levels after training.

CONCLUSIONS

It has been shown that neural learning can be improved upon, in acoustic-to-articulatory mapping, in the estimation of articulatory synthesiser parameters when the statistical characteristic of the acoustic input pattern vectors are statistically adjusted according to the optimum statistical values stated in [6]. This statistical modification to the acoustic input pattern vectors x_i prevents the hidden nodes from approaching the saturation regions and hence improved neural learning results. The results presented in this paper show that an improvement in both the speed of learning and in the accuracy of the estimation, is achieved.

REFERENCES

[1] Schroeter J. and Sondhi M.M. 1994. Technique for Estimating Vocal Tract Shape from Speech Signals, *IEEE Trans. on Speech and Audio Processing*, Vol. 2, pp. 133-149.

[2] D.H. Klatt. 1987. Review of Text-to-Speech Conversation for English. *JASA*, Vol. 82, pp.737-793.

[3] S. Kodiyalam and R. Gurumoorthy. 1996. Neural Networks with Modified Backpropagation Learning Applied to Structural Optimisation, *AIAA Journal*, 34, pp. 408-412.

[4] H.B. Kim et. al.1996. Fast learning-Method For Backpropagation Neural-Network By Evolutionary Adaptation of Learning Rates, *Neurocomputing*, 11 (1) , pp. 101-106.

[5] J.S.N. Jean and J. Wang. 1994. Weight smoothing to improve network generalisation, *IEEE Transactions on Neural Networks*, 5 (5), pp.752-763.

[6] Altun, H. and Curtis, K.M.. 1998. Exploiting the statistical characteristic of the speech signal for improved neural learning in a MLP neural network, *IEEE Neural Networks for Signal Processing VIII*, Edited by Tony Constantinides et al, ISBN 0-7803-5060-X, pp 547-556.

[7] Rahim, M. 1991. Neural Networks in Articulatory Speech Synthesis, PhD thesis, University of Liverpool.

[8] Altun H and Curtis K. M. 1997. Improving neural network learning through output vector scaling, *13th Int. Conf. on Digital Signal Processing*, DSP'97, Vol. 2, pp. 723-726.
[9] Altun H. and Curtis K. M. 1999. Acoustic-to-articulatory neural mapping under different statistical characteristics of articulatory pattern vectors, ICPhS'99.
[10] Altun H. and Curtis K. M. Evaluation of neural learning in a MLP NN under different statistical characteristics of target pattern vectors, (Submitted to Neural Processing Letters)
[11] Van ooyen A and Nienhuis B. 1992. Improving the Convergence of the Back-Propagation Algorithm', *Neural Networks*, Vol. 5 , pp. 463-471.