WEIGHT ZERO ENHANCEMENT IN SPEECH SYNTHESIS USING NEURAL NETWORKS

G. C. Cawley, M. I. Heywood and P. D. Noakes

Neural and VLSI Systems Laboratory, Department of Electronic Systems Engineering, University of Essex, Wivenhoe Park, Colchester, Essex C04 3SQ, United Kingdom.

Abstract

This paper applies a general algorithm for dynamically pruning weights from feed forward networks to the problem of allophone speech synthesis. Work using higher order and sigma-pi networks has indicated that many of the weights in the trained network are redundant. The weight zero enhanced (WZE) algorithm has been developed which removes weights during the learning process with the aim of improving generalisation, whilst having as small an effect on the learning process, and adding as low a computational overhead as possible. In simple conventional speech synthesis systems, pre-recorded allophones are simply concatenated to form the required utterance. Unfortunately the boundaries between allophones are considerably slurred (an effect known as coarticulation) and such a simplistic approach leads to very unnatural sounding speech. This paper describes part of an on going project into the use of neural networks for allophone synthesis with better modelling of coarticulation.

1 INTRODUCTION

In continuous speech the boundaries between allophones are not distinct but are considerably blurred, an effect known as coarticulation [1], caused by the inertia of articulators such as the lips and tongue. Coarticulation can also be caused by articulators positioning themselves in anticipation of subsequent allophones during production of the current allophone. Coarticulation is redundant in that it carries little of the semantic content of an utterance, however we subconsciously expect to hear its effects in natural speech.

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n}}$$

This paper describes the training of neural networks for speech synthesis through generation of LPC parameters corresponding to a sequence of allophones. Linear predictive coding (LPC) [2] attempts to find the coefficients a_k of an all pole filter, with transfer function $H(z)$, such that its spectral properties are optimally similar to that of a segment of sampled speech. Given a suitable excitation signal, speech can be reconstructed from these coefficients, which must be updated roughly every 10ms to allow for the time varying nature of speech. For voiced speech the excitation signal can be approximated by a train of impulses, and for unvoiced speech by random noise.

1.1 Line Spectral Pair (LSP) representation

Unfortunately LPC coefficients are not themselves suitable for training neural networks as they are highly sensitive to error. Small changes in the predictor coefficients can lead to large changes in the spectral properties of the synthesis filter, at worst leading to instability. The LPC coefficients must be transformed into an equivalent parameter set with more suitable properties. The PARCOR [2] and log area ratio [3] parameter sets are widely used in low bit rate coding of speech and have

also been used in training neural speech synthesizers. Line spectral pair (LSP) [4] representation is an equivalent parameter set found to have excellent quantization and interpolation properties for use in low bit rate coding of speech. These properties have also been found to be useful in our research in training neural networks for speech synthesis [5]. Line spectral pair coding records the frequency of the zeros of two polynomials $P(z)$ and $Q(z)$ which are related to the predictor polynomial $A(z)$ by the following equations:

$$\begin{aligned} P(z^{-1}) &= A_p(z^{-1}) - z^{-(p+1)}A_p(z) \\ Q(z^{-1}) &= A_p(z^{-1}) + z^{-(p+1)}A_p(z) \end{aligned}$$

The zeros of $P(z)$ and $Q(z)$ lie on the unit circle in the z plane, this reduction in the search space allows efficient root finding methods to be employed (the roots of $A(z)$ can also form a useful parameter set, however the generalised root-finding process is computationally expensive). For the synthesis filter to be stable, the zeros of $P(z)$ alternate around the unit circle with the zeros of $Q(z)$. The overall spectral sensitivity of LSP parameters is less than that of PARCOR or log area ratio parameters, and also the spectral sensitivity of individual LSP parameters are uniform whereas low order PARCOR parameters exhibit higher sensitivities.

1.2 Weight zero enhancement (WZE)

Heywood and Noakes [6]introduces an enhancement to the standard back-propagation (BP) learning rule, which removes redundant weights during the learning process by a zero-weight mechanism without incurring significant deviations from the standard back-propagation learning algorithm. To achieve this, two extra learning parameters are introduced. One is a stability threshold which is compared to the stability of each neuron in the network during the weight update cycle. When an individual neuron's stability betters that of the threshold, then the weights feeding this neuron can be considered for weight zeroing. The second user-specified parameter defines a maximum weight magnitude such that any weight below this magnitude satisfying the stability threshold is set to zero.

By removing weights, sparsely connected networks are generated during the learning process. Training using the Weight Zero Enhanced (WZE) algorithm continues until a given RMS error is reached. At this point a final flush of redundant weights is performed, such that all the weights with a magnitude above the user specified maximum weight magnitude are extracted, so defining the sparsely connected network. Training then continues using the standard BP procedure until the final maximum output layer error is satisfied, preferably using a higher learning rate and momentum. In order to minimise disruption to neuron stability, caused by setting weights to zero in large networks, a limit to the number weights which can be set to zero at any one time is employed [6]. This paper continues the development of the WZE algorithm by applying the algorithm to a large problem involving the forming of a continuous mapping rather than the pattern classification problems to which the algorithm has previously been applied.

1.3 Network decomposition

Large artificial neural networks with large training sets inevitably take many hours, even days to train satisfactorily on current workstations. In order to take advantage of a number of workstations which lie largely unused overnight, the network is decomposed into a number of sub-networks which may be trained in parallel. Each subnet is responsible for generating a different type of speech sound. The use of one network per allophone has been found to lead to poor generalisation as each network receives too few training patterns [5], and so one network is used for each of seven phonetic classes.

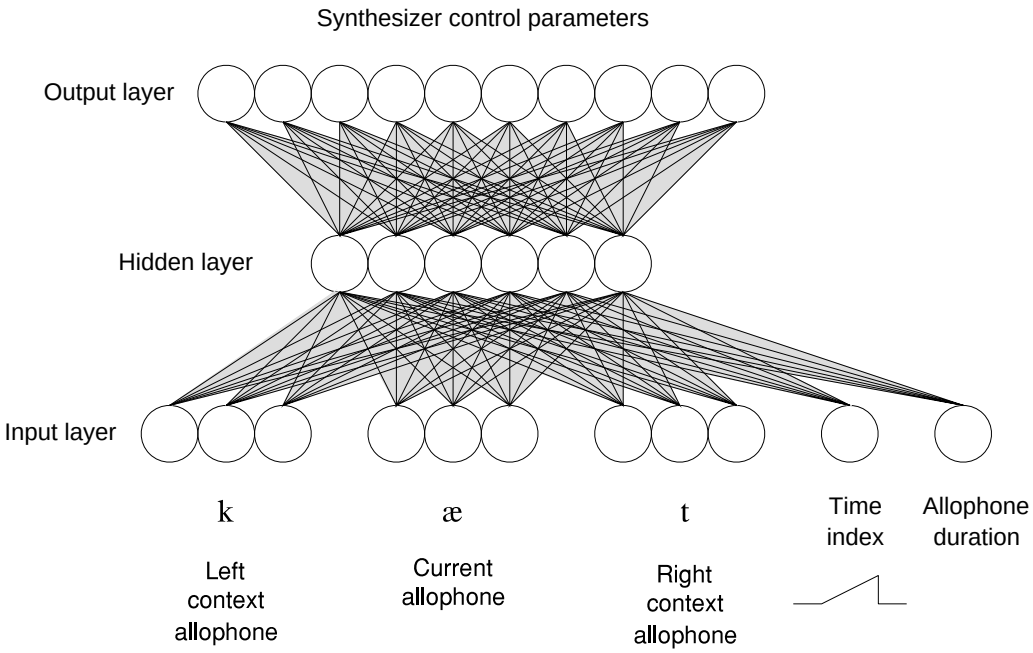

Figure 1: Schematic drawing of network architecture.

2 NETWORK ARCHITECTURE

A network architecture similar to that used in the NETalk [7] system is employed (see Figure 1). The input layer forms a sliding window over the input stream of tokens representing allophones. The input layer consists of three groups of neurons which represent the current allophone and the previous and subsequent allophones to provide partial context. Each allophone is represented by a arbitrary binary code. In addition one input neuron is used to indicate the duration of the current allophone and an index neuron is used to indicate how much of the current allophone has already been generated. In order to synthesize speech parameters for a complete allophone, the input layer is set to the appropriate pattern for the central and context allophones and the required duration. A ramp input is then applied to the index neuron. As the index increases, the outputs of the network step out the parameters required to synthesize the allophone.

3 TRAINING

All ten sentences from one speaker in the TIMIT database [8] were analysed using tenth order LPC analysis. Eight sentences were used in training and two sentences reserved for testing. All inputs and outputs were normalised to values between ± 0.5. Each sub-network was trained to produce allophones belonging to one of seven phonetic classes: affricates, fricatives, plosives, nasals, liquids and glides, vowels, and miscellaneous. The networks were trained using the weight zero enhanced back-propagation algorithm on a number of Sun Sparcstations using a simulator written in C.

4 RESULTS

The WZE algorithm is applied until the RMS error measure has reduced to 0.1 for all but the Affricate and Miscellaneous cases, where RMS errors of 0.15 and 0.05 are applied respectively. Following this training continues using the standard BP algorithm until either an RMS error of 0.05 (0.01 in the case of the Miscellaneous case) or 500 iterations of the training set are completed. At this point this test data is presented. The main aim of these experiments is to determine the

extent, if any, of weight redundancy and resulting generalisation improvements. Simulation results (Figures 2 and 3) indicate that weight redundancy is extensive throughout all the networks. In the case of the more important phonetic classes of vowel, nasal and plosive, where the data set is larger, useful improvements over the generalisation abilities previously attained using BP alone. Figure 2 summarises RMS error on test data (i.e. generalisation ability) as a function of zero weight threshold, and Figure 3 number of weights employed as a function of zero weight threshold. A zero weight threshold of zero corresponds to the standard BP algorithm. The figures indicate that higher zero-weight thresholds result in either to many weights being set to zero or the location of an optimal solution (nasal data class). As the same network size is used for totally different data complexities (23, 30, 10), optimum network complexities will occur at different weight counts. Consequently, when using the WZE algorithm it is important to perform simulations over a range of zero-weight thresholds in order to locate the threshold returning best performance. A zero-weight threshold of 0.1 typically returns a generalisation performance near that of the standard BP algorithm, after this the zero-weight threshold returning the best generalisation performance is a function of the data complexity. Hence, those data classes with a complex data structure reach an optimum generalisation ability for a lower zero-weight threshold (plosive data class) than the more simple data structures, which have a lower interconnect density, hence make use of the higher zero-weight thresholds (nasal data class).

5 CONCLUSIONS

Given the limited training data available, the authors believe that useful performance increases result from application of the WZE algorithm. These data sets of a more complete nature (nasal, vowel, plosive) allow the WZE algorithm to return better generalisation abilities than available from standard BP alone. Furthermore, substantial reductions in weight requirements are possible without disturbing the network performance (a function of the WZE algorithms dynamic pruning technique which integrates pruning with the BP process). However, further work addressing the effects of:

- Changing the network architecture to a four layer structure.
- Incorporation of product terms.
- Application of an improved data set.

6 ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of the United Kingdom Science and Engineering Research Council (SERC).

References

[1] D. O'Shaughnessy. *Speech Communication — Human and Machine.* Addison Wesley, 1987.

[2] L. R. Rabiner and R. W. Schafer. *Digital processing of speech signals.* Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1978.

[3] R. Viswanathan and J. Makhoul. Quantization properties of transmission parameters in linear predictive systems. *IEEE Trans. on acoustics, speech and signal processing*, ASSP-23(3):309–321, 1975.

[4] N. Sugamura and F. Itakura. Speech analysis and synthesis methods developed at ECL in NTT — from LPC to LSP. In *Speech Communication*, volume 5, pages 199–215, 1986.

[5] G. C. Cawley and P. D. Noakes. LSP speech synthesis using backpropagation networks. In *in press, ANN-93*, Brighton, May 1993.

[6] M. I. Heywood and P. D. Noakes. Simple addition to back-propagation learning for dynamic weight pruning, sparse network extraction and faster learning. In *in press, IEEE ICNN*, San Fransisco, 1993.

[7] T. J. Sejnowski and C. R. Rosenberg. Parallel networks that learn to pronounce english text. *Complex Systems*, 1:145–168, 1987.

[8] DARPA acoustic-phonetic continuous speech corpus (TIMIT).

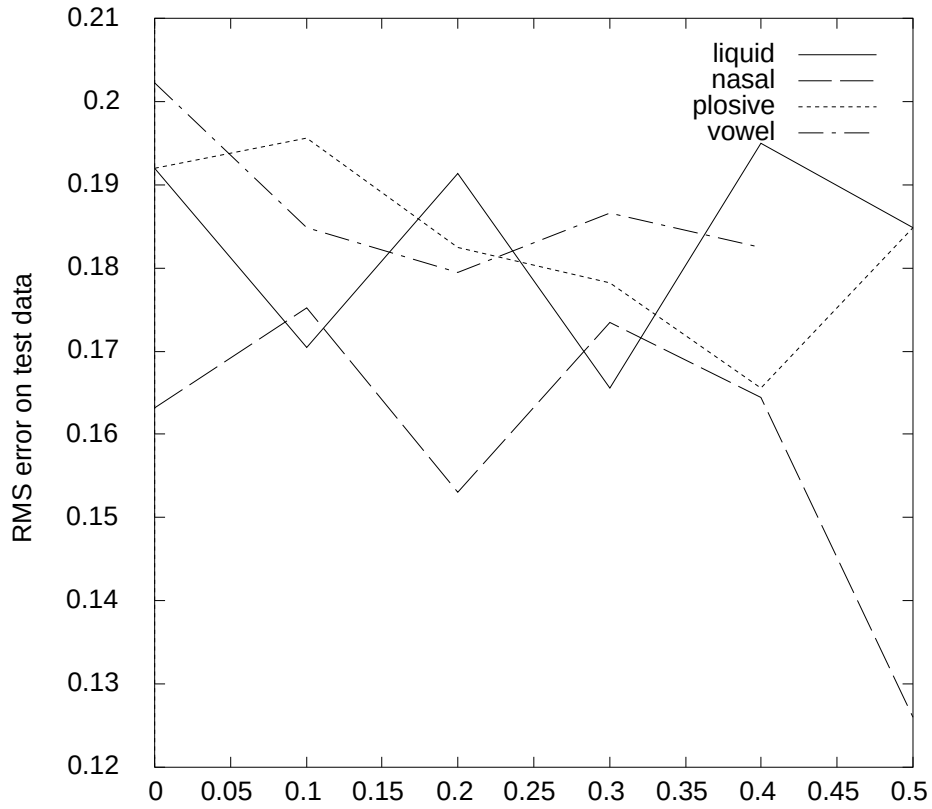

Figure 2: Graph RMS error on test data as a function of zero-weight threshold.

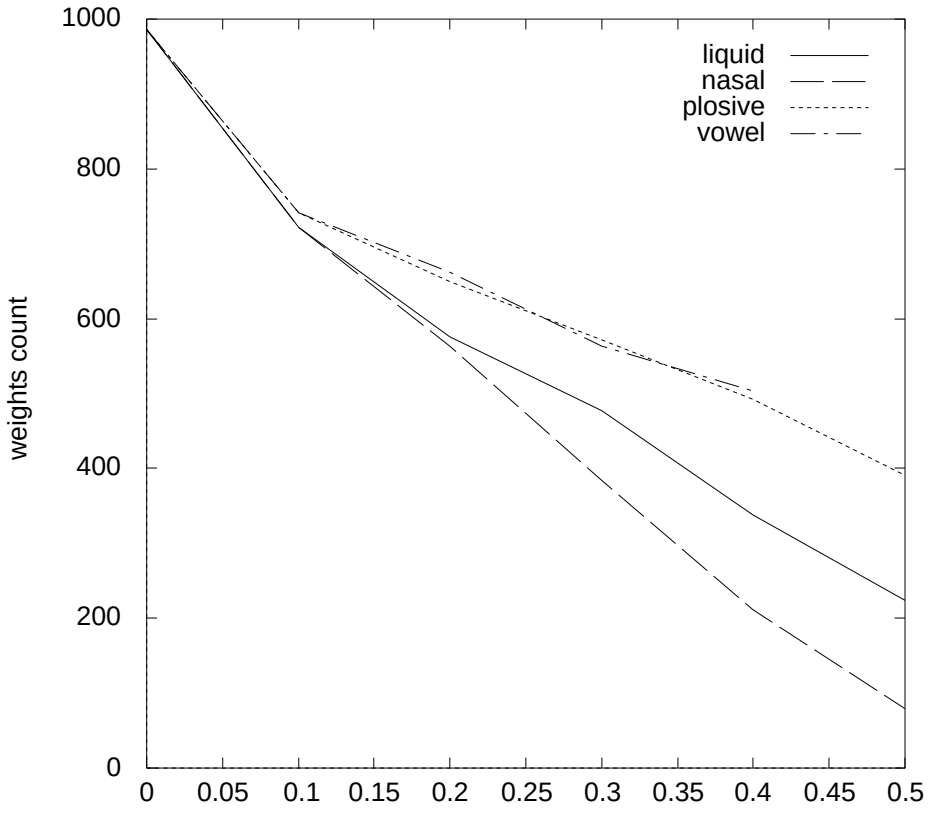

Figure 3: Graph of weight count as a function of zero-weight threshold.