LSP SPEECH SYNTHESIS USING BACKPROPAGATION NETWORKS

G.C. Cawley and P.D. Noakes

University of Essex, U.K.

ABSTRACT

A multi-layer perceptron (MLP) similar to that used in the NETtalk system is used to form a mapping between sequences of allophones and corresponding frames of LPC synthesizer control parameters. Three parameter sets equivalent to the LPC coefficients, line spectral pair (LSP), PARCOR and log area ratio, are evaluated. In addition to training a standard MLP, networks which have been decomposed according to phonetic class and by allophone, are trained. Decomposition is found to reduce training time and produce greater accuracy on the training set, however the network decomposed by allophone is found to receive to few training patterns too generalise properly on new data.

INTRODUCTION

In continuous speech the boundaries between allophones are not distinct but are considerably blurred, an effect known as coarticulation (O'Shaughnessy [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 carries little of the meaning of an utterance, however we subconsciously expect to hear its effects in natural speech. Our research has been concerned with the use of neural networks to model the effects of coarticulation in synthetic speech.

Linear predictive coding (LPC) (Rabiner and Schafer [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.

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

This paper describes the training of neural networks for speech synthesis through generation of LPC parameters corresponding to a sequence of allophones. 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 (Viswanathan and Makhoul [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) (Sugamura and Itakura [4]) representation is an equivalent parameter set found to have excellent quantization and interpolation properties for use in low bit rate coding. These properties have also been found to be useful in our research in training neural networks for speech synthesis. 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:

$$\begin{aligned} P(z^{-1}) &= A_p(z^{-1}) - z^{-(p+1)} A_p(z) \\ &= 1 + (a_1 - a_p) z^{-1} + \cdots \\ &\quad + (a_p - a_1) z^{-p} - z^{-(p+1)} \\ Q(z^{-1}) &= A_p(z^{-1}) + z^{-(p+1)} A_p(z) \\ &= 1 + (a_1 + a_p) z^{-1} + \cdots \\ &\quad + (a_p + a_1) z^{-p} + z^{-(p+1)} \end{aligned}$$

Where p is the order of polynomial A(z)

The zeros of $P(z)$ and $Q(z)$ lie on the unit circle in the z plane, and 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 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 $Q(z)$ (see Figure 1). The overall spectral sensitivity of LSP parameters is less than that of PARCOR and log area ratio parameters, and also the spectral sensitivity of individual LSP parameters are uniform whereas low order PARCOR parameters exhibit higher sensitivities.

Large artificial neural networks with large training sets inevitably take many hours, even days to train satisfactorily on current workstations. In order to take full advantage of a number of workstations which lie largely unused overnight, the decomposition of the neural network is investigated. Lateral decomposition has been found to reduce the training time and to converge with greater accuracy (Lucas *et al.* [5])). This paper presents results of an experiment in which a network is decomposed into a number of sub-networks each of which produce a different type of speech sound.

NETWORK ARCHITECTURE

An network architecture similar to that used in the NETalk (Sejnowski and Rosenberg [6])) system was employed (see Figure 2). 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 vector of articulatory features such as phonetic class, stress and place of articulation. 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.

TRAINING

All ten sentences from one speaker in the TIMIT database (NTIS [7]) were analysed using tenth order LPC analysis. Eight sentences were used in training and two sentences reserved for testing. Three systems were trained, the first a conventional MLP with 100 hidden layer neurons, the second an array of smaller networks, each with 30 hidden layer neurons, trained to produce allophones belonging to one of seven phonetic classes: affricates, fricatives, plosives, nasals, liquids, vowels and miscellaneous. The third system was comprised of one network per allophone, each network containing 15 hidden layer neurons. Each system was trained using LSP, PARCOR and log area ratio data. The networks were trained using a backpropagation algorithm on a number of Sun Sparcstations using a simulator written in C.

RESULTS

Decomposition of neural networks has been found to greatly reduce training time, each network requires fewer hidden layer neurons and has a smaller training set. Training can also be performed in parallel if a number of workstations are available. The results obtained are displayed in Figure 3, which shows RMS error against cycles trained for a standard MLP and networks decomposed according to phonetic class and by allophone. Decomposition can clearly be seen to improve performance on the training set, the greater the decomposition the simpler the learning problem for each network becomes, the better the results obtained. The more important results obtained on the test set suggests that some decomposition is beneficial, but that a network can clearly be decomposed too far. Our interpretation of this result is that the greater the decomposition, the fewer training examples are presented to each network, so eventually the network sees too few examples to generalise those it has seen correctly. The results were obtained using networks trained using LSP data, similar results are obtained using PARCOR and log area ratio data.

Figure 4 shows a graph of spectral distortion against cycles trained for networks trained using LSP, PARCOR and log area ration data.

It can be seen that best performance on both training and test sets is obtained using LSP data, followed by log area ratio and lastly PARCOR data. The results shown were obtained using networks decomposed by allophone, similar results are obtained using standard MLPs and networks decomposed according to phonetic class. The results confirm those obtained during initial experiments indicating the superior qualities of LSP parameters for use in training neural speech synthesizers (Cawley and Noakes [8]).

CONCLUSIONS

We have shown LSP coding to be superior to the PARCOR and log area ratio parameter sets for use in training neural networks for speech synthesis. Decomposition of neural networks is found to be a useful approach for reducing training time and exploiting opportunities for parallelism. The decomposed network converges more quickly and more accurately than a conventional MLP, however a highly decomposed network will require a substantially larger training set in order to maintain performance on new utterances.

ACKNOWLEDGEMENT

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

References

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

[2] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*, chapter 8. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, USA, 1978.

[3] R. Viswanathan and J. Makhoul. Quantization properties of transmission parameters in linear predictive systems. *I.E.E.E. Transactions on Accoustics, Speech, and Signal Processing*, 23(3):309–321, June 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] S. Lucas, Z. Zhao, G. Cawley, and P. Noakes. Pattern recognition with decomposed multilayer perceptron. *I.E.E. Electronic Letters*, 29(5):442–443, March 1993.

[6] T. J. Sejnowski and C. R. Rosenberg. Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145–68, 1987.

[7] National Technical Information Service (NTIS), Computer Systems Laboratory, Gaithesburg, MD, USA. *DARPA acoustic-phonetic continuous speech corpus (TIMIT).*

[8] G. C. Cawley and P. D. Noakes. Allophone synthesis using a neural network. In *Proceedings of the World Conference on Neural Networks*, volume 2, pages 122–125, Portland, Oregon, USA, 1993.

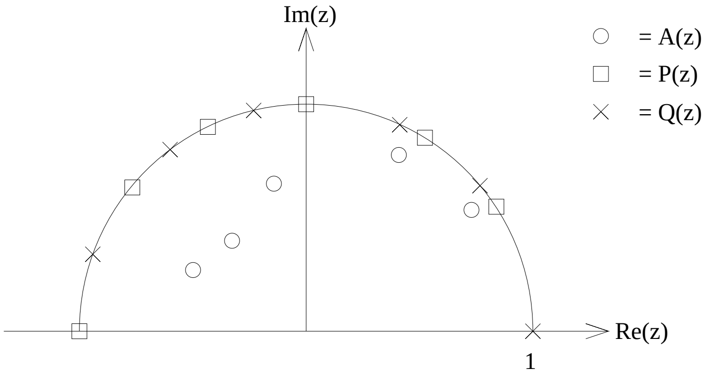

Figure 1: Relation between roots of $A(z)$, $P(z)$ and $Q(z)$

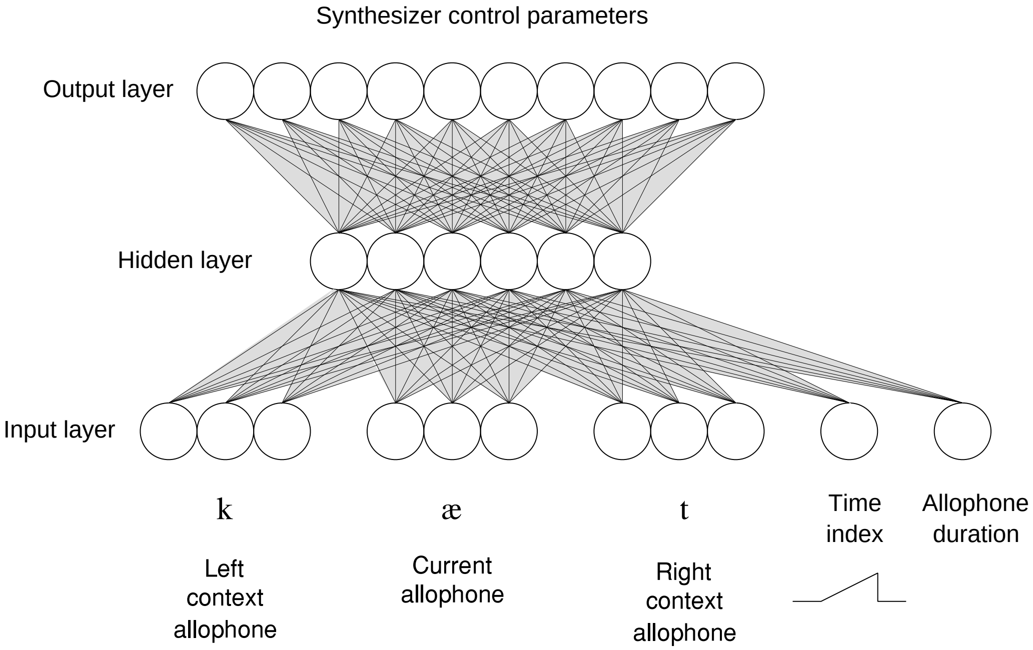

Figure 2: Schematic drawing of network architecture

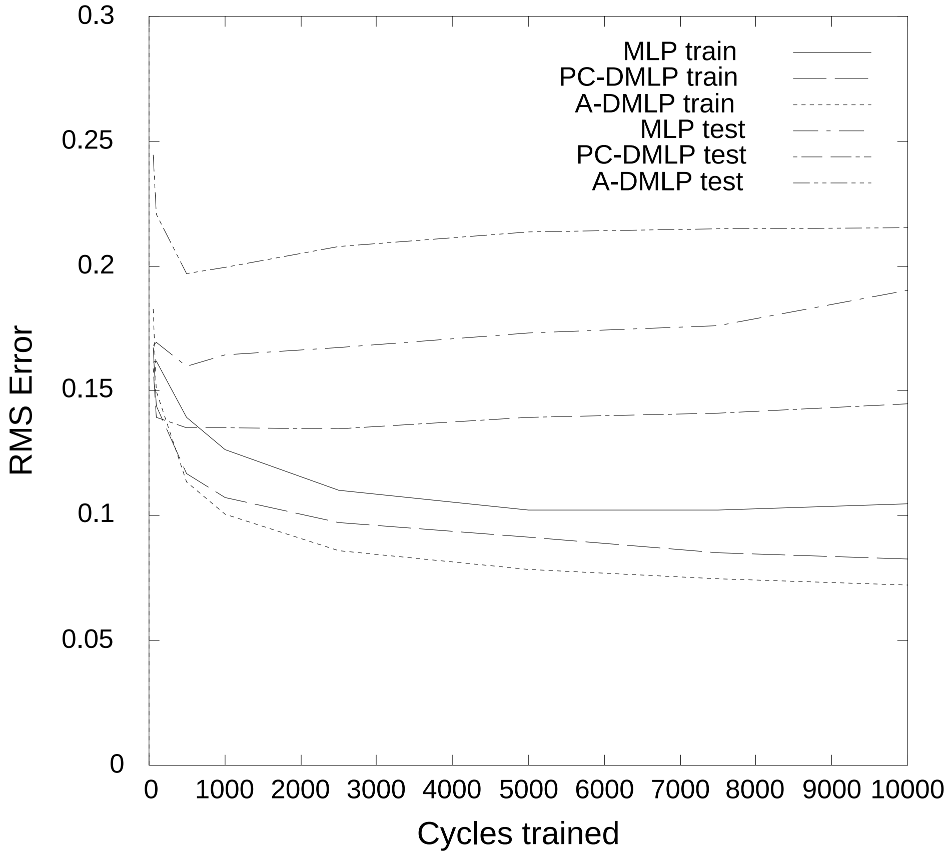

Figure 3: Graph of RMS error against cycles trained for standard multi-layer perceptron (MLP), MLP decomposed according to phonetic class (PC-DMLP), and MLP decomposed by allophone (A-DMLP), trained using LSP data. Results obtained on both training and test sets are plotted

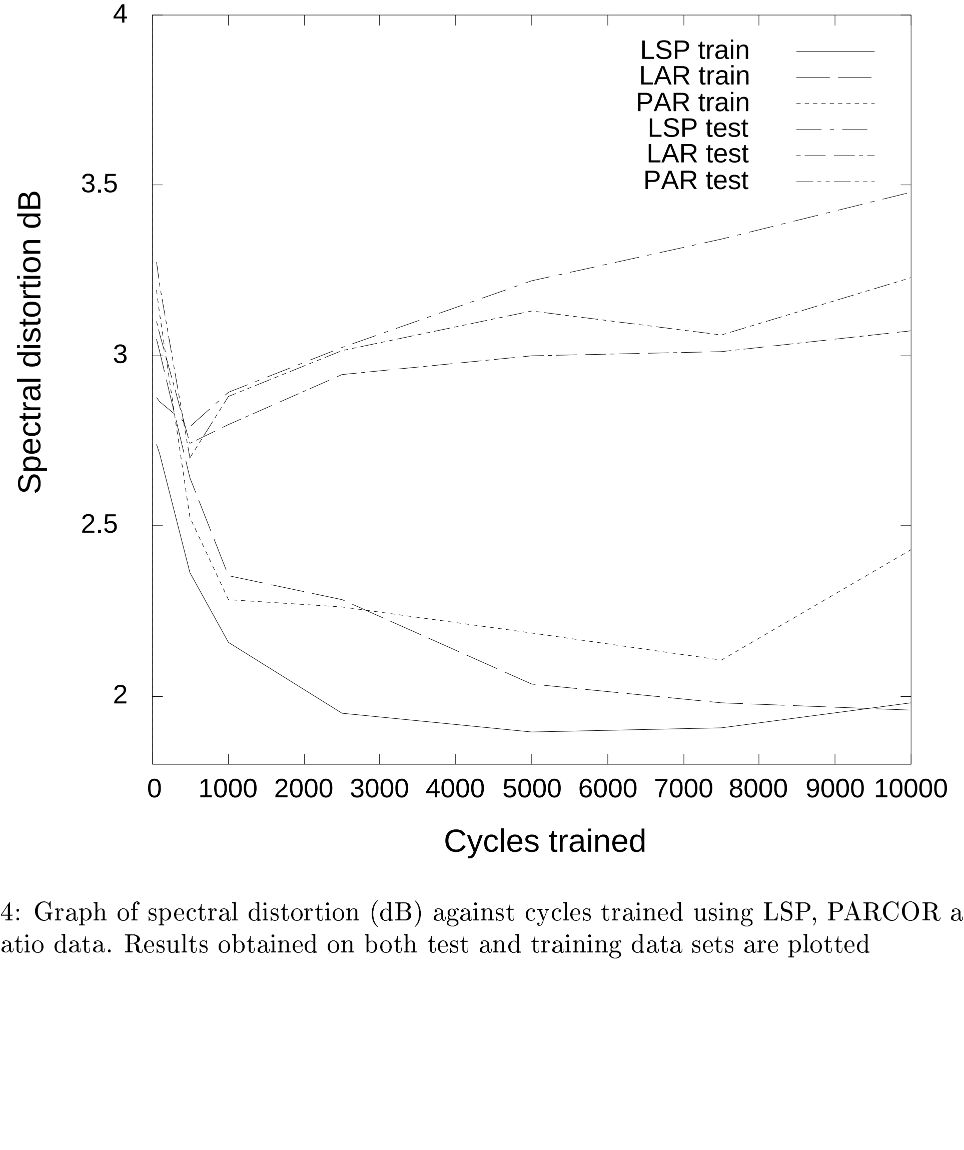

Figure 4: Graph of spectral distortion (dB) against cycles trained using LSP, PARCOR and Log Area Ratio data. Results obtained on both test and training data sets are plotted