Allophone Synthesis Using A Neural Network

G. C. Cawley and P. D. Noakes
Department of Electronic Systems Engineering,
University of Essex
Wivenhoe Park, Colchester C04 3SQ, UK
email ludo@uk.ac.essex.ese

Abstract

Most people reading this paper will be aware of the NETalk system of Sejnowski and Rosenberg [1], in which a multi-layer perceptron was trained to select the correct allophone for combinations of letters occurring in plain English text. Once suitable allophones have been selected, the problem remains of how should the sounds corresponding to a sequence of allophones be produced? The most straight forward approach is to store pre-recorded examples of each allophone and then simply concatenate them to form the required utterance. Unfortunately the boundaries between adjacent allophones in continuous speech are not distinct, an effect known as coarticulation, and such a simplistic approach leads to very unnatural sounding speech. This paper presents some initial findings of experiments to evaluate different parametric forms of speech based on linear predictive coding (LPC) for training neural networks. These experiments were performed as part of a project to improve the subjective quality of speech synthesizers, through the use of neural networks for allophone synthesis.

Introduction

The realisation of an allophone is context-sensitive due to the inertia of articulators such as the lips, jaw and tongue. Articulators can only move at a finite speed in recovering from the position assumed during the previous allophone, causing a gradual transition from one allophone to the next. Coarticulation can also be caused by low level neural processes within the brain, where articulators are also able to position themselves in anticipation of the subsequent allophones. These movements are redundant in that they convey little of the semantic content of the utterance, however we sub-consciously expect to hear effects of these movements in natural speech.

The most simple speech synthesis systems do not attempt to model coarticulation at all, but simply concatenate pre-recorded allophones. A more sophisticated approach concatenates diphones, each consisting of the adjacent halves of two allophones. Diphones capture the immediate effects of coarticulation and abut during the relatively steady state conditions during the central part of each allophone. However this is at the expense of increased storage, as about 1200 diphones are required for the allowable permutations of around 60 allophones. If a parametric description of speech is used, such as formant data, which records the frequency and amplitudes of the spectral peaks known as formants, templates may be used to interpolate the value of each parameter between target values set for each allophone. The rate at which each parameter changes is determined according to the rank of each allophone, which reflects the degree to which it affects others. This allows more natural sounding speech to be produced, but at the expense of increased complexity, and requires manual analysis of human speech to determine targets and rankings for each allophone.

Our research has been concerned with investigating the use of neural networks for allophone synthesis based on formant data [2, 3]. Unfortunately formant analysis of continuous speech is a complex and computationally expensive procedure, making it difficult to obtain the large

amounts of training data needed. This paper presents results of initial experiments to the evaluate different coding techniques based on linear predictive coding, which are less complex and less computationally expensive.

Linear Predictive Coding (LPC) [4] is a technique used to find the coefficients a_k of an all pole filter, with transfer function $H(z)$, such that its spectral properties are similar to that of a segment of sampled speech. Given a suitable excitation signal, speech can be reconstructed from these coefficients, which are updated every 10ms to allow for the time-varying nature of speech. For voiced speech the excitation signal can be approximated by a periodic train of impulses, and for unvoiced speech by random noise.

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

In this paper, neural networks trained using three coding schemes based on linear predictive coding are compared, PARCOR [4, 5], log area ratio [4] and line spectral pair (LSP) [5]. Table 1 summarises some of the relative merits of each of method.

Table 1: A comparison of the properties of PARCOR, log area ratio and LSP coding schemes

Property	PARCOR	Log Area Ratio	LSP
Inter-parameter spectral sensitivity	Lower order coefficients more sensitive	Lower order coefficients more sensitive	Uniform
Individual parameter sensitivity	Non-Uniform	Uniform	Uniform
Overall spectral sensitivity	Good	Good	Very Good
Interpolation properties	Poor	Poor	Good

Network architecture

An architecture similar to that employed in the NETalk system [1] was used, the input layer forming a sliding window over the input stream of allophones (see Figure 1). The input layer consists of three groups of neurons corresponding to the current and right and left context allophones. Each allophone is represented by a vector of phonetic features such as the broad phonetic class 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. All ten sentences from one speaker in the TIMIT database [6] were then analysed using tenth order LPC analysis to generate PARCOR, log area ratio and LSP training data. The network was trained using the backpropagation algorithm simulator written in C running on a Sun Sparcstation.

Results

The results obtained are displayed in Figures 2 and 3 which show graphs of RMS error and spectral distortion against cycles trained for each coding scheme. The results given here were obtained using a hidden layer of 50 neurons, however similar results are obtained using different hidden layer sizes. The log area ratio is a transformation of the PARCOR parameter set designed to flatten the spectral sensitivity of individual parameters, and was expected to produce marginally better

results for this reason. This proved to be the case, the improvement was especially noticeable during voiced sounds where PARCOR coefficients tend to approach ± 1 where the coefficients spectral sensitivity is at its greatest. LSP coding was expected to out-perform the other coding schemes. Firstly the overall spectral sensitivity of LSP parameters is slightly lower, secondly LSP coefficients exhibit better interpolation properties and lastly because each LSP coefficient has roughly the same spectral sensitivity. The spectral sensitivity of lower order PARCOR and log area ratio coefficients are higher, and so some method is required to concentrate training on reducing the error in the low order coefficients. The networks trained using LSP data seemed to train faster than those trained on PARCOR and log area ratio data, and the sentences learned with less spectral distortion. Speech generated using the network generated using LSP data was also judged to be subjectively better.

Conclusions

We have shown that the use of LSP parameters in training neural networks for speech synthesis results in faster training and higher objective and subjective speech quality than is obtained using PARCOR or log area ratio parameters. Work is currently underway to produce a complete neural network allophone speech synthesizer using line spectral pair representation.

References

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

[2] G. C. Cawley and A. D. P. Green. The application of neural networks to cognitive phonetic modelling. In *Proc. 2nd IEE Int. Conf. on Artificial Neural Networks*, pages 280–284, 1991.

[3] G. C. Cawley and P. D. Noakes. Diphone synthesis using a neural network. In *Proc. 1992 Int. Conf. on Artificial Neural Networks (ICANN-92)*, volume 1, pages 795–798, 1992.

[4] L. R. Rabiner and R. W. Schafer. *Digital processing of speech signals*, chapter 8. Prentice-Hall, 1978.

[5] 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.

[6] National technical information service (NTIS), Computer systems laboratory, Gaithesburg, MD, USA, 20899. *DARPA acoustic-phonetic continuous speech corpus (TIMIT).*

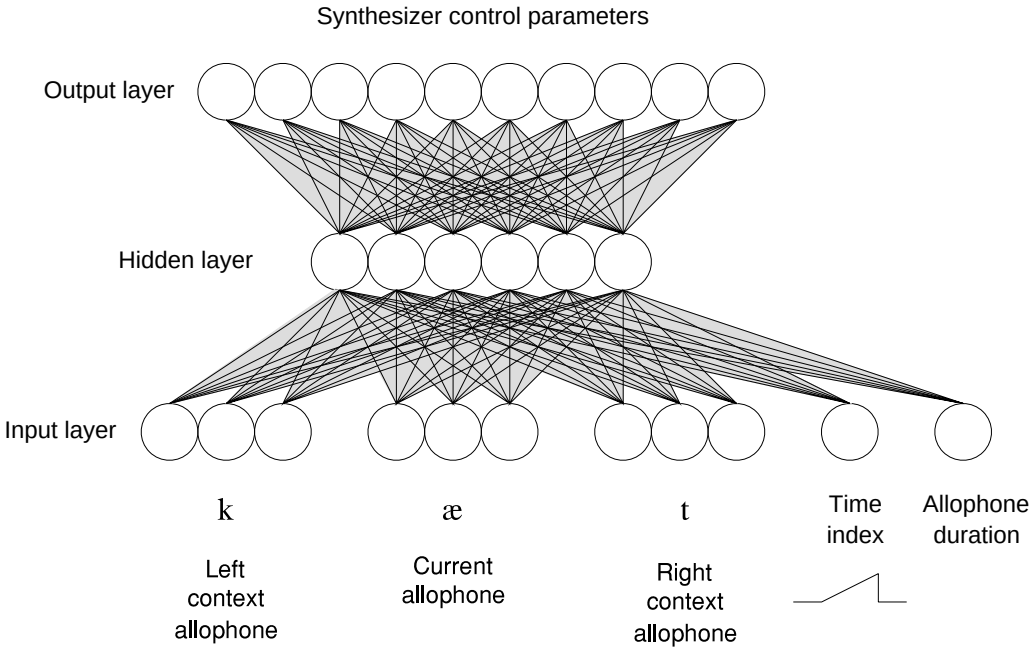

Figure 1: Schematic drawing of network architecture

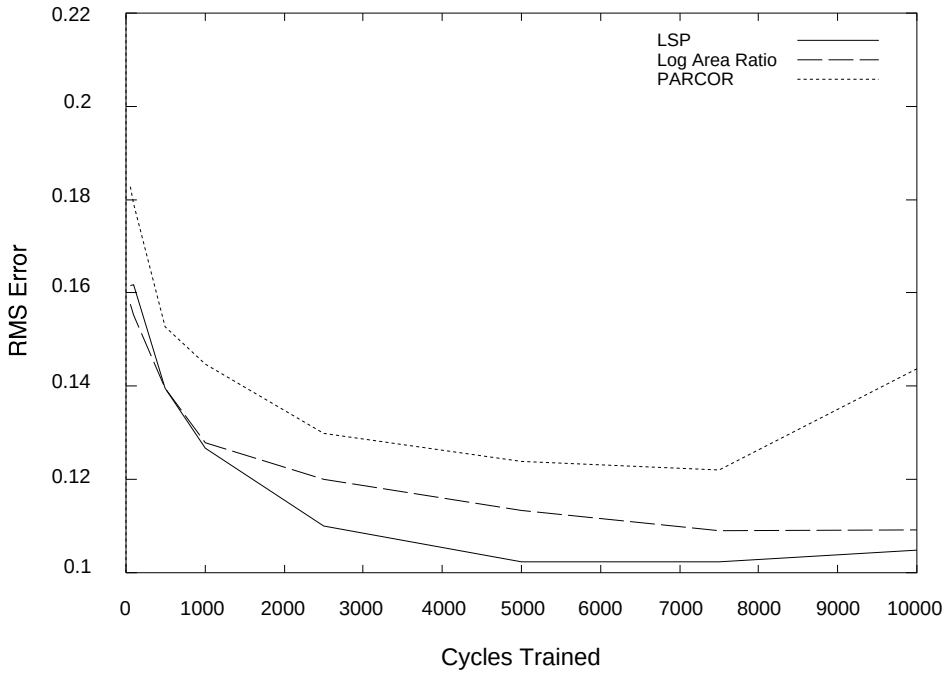

Figure 2: Graph of RMS error against cycles trained

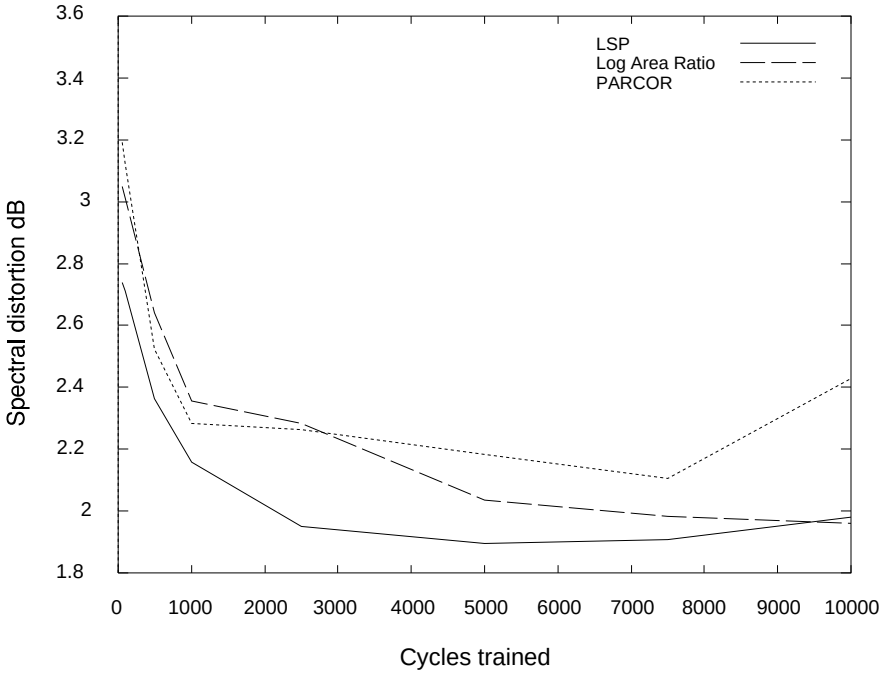

Figure 3: Graph of spectral distortion against cycles trained