DIPHONE SYNTHESIS USING A NEURAL NETWORK

G.C. Cawley and P.D. Noakes

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

Abstract

A neural network is used to produce formant data for the Holmes parallel formant speech synthesizer [2] from an allophonic transcription of plain english text. This paper presents results obtained from training a back propagation neural network using speech generated by a conventional speech synthesizer. The network is able to learn to reproduce the basic form of formant transitions between allophones. Use of formant data obtained from natural speech is planned.

1 INTRODUCTION

Conventional speech synthesis systems typically consist of three functional modules: (i) a phonological element, which converts plain text into phonemes (an abstract representation of the basic speech sounds), (ii) a prosodic component, which generates pitch and timing information to add intonation and (iii) a phonetic component, which deals with the realisation of the selected phonemes as speech, usually in the form of control parameters for a formant synthesizer.

The phonetic component first selects an appropriate allophone for each phoneme according to context. An allophone is one of a set of speech sounds which can be regarded as variants of the same phoneme. The initial l in light has a different qualities to the word final l in eel, but both have the same linguistic meaning, and so are assigned different allophones accordingly (the l in light is a 'clear' l, whereas the l in eel is pronounced with the back of the tongue higher in the mouth producing a 'dark' l). The differences in the realisation of a phoneme with context is due to co-articulation, an anticipatory movement of the articulators towards the position required for the next phoneme. Although allophone selection provides coarse modelling of co-articulation, to model more subtle aspects, formant parameters are interpolated using simple straight lines from target values stored for each allophone which are modified according to context.

While a rule based approach is well suited to the symbolic processing involved in the phonological module, lower level functions of the prosodic and phonetic modules are less easily expressed in the form of explicit rules, and so an assembly of neural networks may be a more appropriate paradigm in these areas.

2 NEURAL ARCHITECTURE

A multi-layer perceptron is trained to generate control parameters for the Holmes parallel formant synthesizer from a list of allophones provided by the JSRU speech by rule system [3]. The network considers two allophones at a time, producing formant data for the diphone formed by their adjacent halves.

Both arbitrary (seven bit) binary coding and representations based on articulatory features such as place of articulation and lip-rounding (there are twenty two in all) have been used. In addition two extra neurons are used to indicate the duration of each allophone as this affects the behaviour of the formants during transitions.

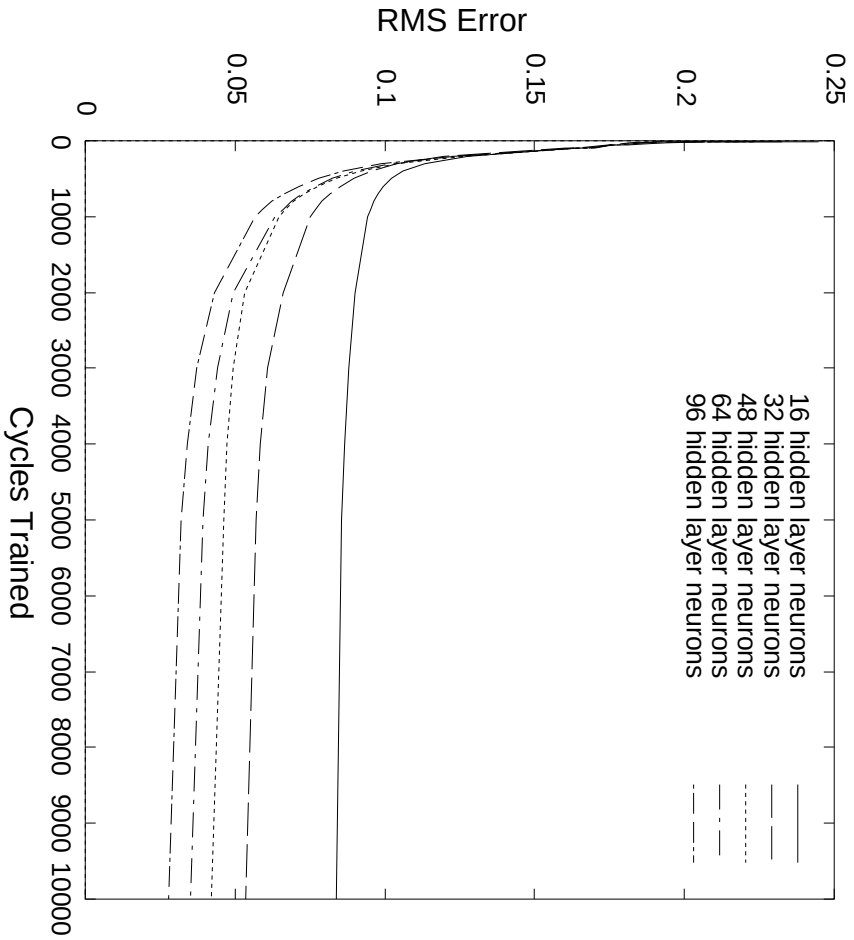

Figure 1: Graph of total rms error against cycles trained for 16, 32, 48, 64 and 96 hidden units.

The Holmes parallel formant synthesizer accepts frames of eleven parameters, updated every 10ms. The frequency of the nasal formant, FN, may remain constant for a given speaker without affecting the subjective quality of the speech, a value of 250Hz is used in these experiments. The pitch or fundamental frequency, F0, is supplied by the JSRU speech by rule system which also provides the segmental duration. This leaves nine parameters to be generated by the network. Each parameter is sub-sampled such that six samples are taken during each phonetic element. At each step each neuron in the output layer generates one of six samples for each of the nine parameters over the transition between the phonetic elements presented at the input layer. Three samples are allocated to the adjacent halves of each allophone.

3 TRAINING

In order to test the sub-sampling and interpolation scheme used an initial experiment was conducted to train a neural network to pronounce the set of minimal pairs pit, pat, pot, put, put and bit, bat, bot, but, butt. The results of this experiment were promising and demonstrated that using sub-sampling and interpolation procedure it was possible to capture formant transitions accurately (for a discussion of this work see [1]). A much larger experiment was then performed using a training set of 866 vectors extracted from formant data generated by the JSRU speech by rule system from a corpus of 25 sentences.

3.1 Input coding

It was found that training times for networks using an input coding based on articulatory features and those based on an arbitrary binary coding were approximately equal. The use of a representation based on articulatory features is more computational expensive as it requires many more input neurons than a binary coding. However such a representation may allow us to model the effects of co-articulation extending beyond the boundaries between adjacent allophones, through the partial activation of those input neurons corresponding to the co-articulated feature during production of previous or subsequent allophones.

3.2 Hidden layer dimensions

A hidden layer consisting of around 64 neurons was found to provide a good compromise between error and complexity. Hidden layers as large as 128 and 256 neurons have been used, but proved computationaly prohibitively expensive for only a small reduction in the total rms error. Figure 1 shows a graph of total rms error against cycles trained for networks with hidden layers consisting of 16, 32, 48, 64 and 96 neurons.

4 EVALUATION OF EXPERIMENTAL RESULTS

A subjective assessment of the effects of formant parameter quantization and bandwidth limitation on voiced speech, performed by Rosenberg, Schafer and Rabiner [5], indicates that formant synthesis should be relatively insensitive to the average level of error in the parameters generated by the network. However some training patterns still exhibited large maximum errors after a large number of training cycles. Analysis of these patterns indicates that most errors fall into three of basic categories, suggesting a number of improvements which could be made:

4.1 Sub-sampling errors

The sub-sampling procedure used in creating the training data was responsible for a number of the patterns exhibiting large maximum error at the boundary between allophones. This effect was especially noticeable during diphones composed of very short phonetic elements such as the release or post-release phases of plosives. Most of these errors were in the ratio of voiced to unvoiced speech, which tends to change abruptly between its extreme values at the boundaries between voiced and unvoiced allophones. The use of network architectures which eliminate the need for complicated sub-sampling and interpolation schemes are currently being evaluated.

4.2 Synthesis of glottal fricative h

The network had considerable difficulty in synthesizing the glottal fricative h and other allophones particularly susceptible to the effects of co-articulation. Roach [4] describes h as "phonetically a voiceless vowel with the quality of the voiced vowel that follows it". The effects of this co-articulation are present throughout production of the h. When producing the initial half of the h, a diphone synthesizer is unaware of the identity of the following vowel which will greatly affect its spectral properties. This implies that allophone synthesis should be employed, in which the neural network produces formant data for a complete allophone given both right and left context, allowing us to model the effects of co-articulation which extend beyond half way through the adjacent allophone.

4.3 Formant frequency errors during silent segments

During transitions into silent segments at the end of sentences a number of substantial errors were found in the parameters relating to formant frequencies. However since formant amplitudes are set to minimum at these times these errors are not audible. Pre-processing of the training

data to ensure that formant frequencies are handled uniformly, would not affect the quality of the speech produced, but may reduce training time as the error for such patterns would be more easily minimised and would disrupt the learning of other patterns less.

5 CONCLUSIONS

This paper discusses the use of neural networks to generate formant data for the Holmes parallel formant speech synthesizer. It is shown that the system works successfully, but results indicate that some improvements are necessary in both the structure of the network and in the way in which data is presented to the network. An investigation of the number of neurons in the hidden layer suggests that 64 units is adequate to produce acceptable modelling of formant behaviour. Further work is under way to refine the neural architecture and to investigate the use of formant data obtained from human speech. It will be interesting to see if a neural network will be able to interpret the variability in continuous human speech in a meaningful way.

References

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

[2] J. N. Holmes. Formant synthesizers: Cascade or parallel? In *Speech Communications*, volume 2, pages 251–273, 1983.

[3] E. Lewis. *A 'C' implementation of the JSRU text-to-speech system.* Computer Science Department, University of Bristol, August 1989.

[4] P. Roach. *English Phonetics and Phonology — A Practical Course.* Cambridge University Press, 1983.

[5] A. E. Rosenberg, R. W. Schafer, and L. R. Rabiner. Effects of smoothing and quantizing the parameters of formant coded voiced speech. *The Journal of the Acoustical Society of America*, 50(6):1532–1538, 1971.