THE APPLICATION OF NEURAL NETWORKS TO COGNITIVE PHONETIC MODELLING

G.C. Cawley and A.D.P. Green

University of Essex, U.K.

ABSTRACT

A neural network is used to generate control parameters for a parallel formant speech synthesizer, corresponding to a sequence of allophone tokens. Training is to be accomplished using formant data obtained from both natural and synthetic speech. It is intended that theories of cognitive phonetics, currently being developed in the Department of Language and Linguistics at the University of Essex, will be used in order to improve the modelling of coarticulation.

INTRODUCTION

Speech is produced as the result of a coordinated sequence of movements of articulators such as the lips and tongue. For a given language there exists a set of discrete speech sounds known as phonemes which serve to distinguish one word from another. Each phoneme can be described by the set of articulatory gestures by which it is produced. Providing this does not conflict with the gestures required to produce the current phoneme, an articulator may begin to postion itself before the end of the current phoneme in preparation for production of the next. For instance the lip rounding gesture required to produce the vowel *u* in the word 'coup' (*ku*) intrudes into and may even begin before the preceding plosive *k*. This anticipatory movement of articulators and therefore of the formants (the spectral properties of the speech signal) is known as coarticulation. Coarticulation can also occur as a continuation of the motion required to produce the first phoneme into the second. An allophone is one of a set of speech sounds regarded as variants of the same phoneme, the differences in their realization being due to coarticulation. For instance both the aspirated /p/ in 'pot' and the unaspirated /p/ in 'spot' are both allophones of the phoneme *p*. The human auditory system has adapted to expect this variation to be present in natural speech, which when inadequetely modelled or absent, cause synthetic speech to sound unnatural [1]. Furthermore we speak more clearly when one word may easily be confused with another, than when additional clues are available in the semantic content of the message. This suggests that coarticulation is not merely the result of the inertia of articulators or the neurodynamics of the motor neurons by which the commands are transmitted, but is actively controlled as an aid to perception.

In conventional speech synthesis systems such as the JSRU synthesizer [2], the effects of coarticulation are modelled using the following procedure [3]: First, for each of a pair of allophones, target values for each parameter are selected from a table. An ad-hoc set of rules is used to modify these targets depending on the combination of allophones to be modelled. One of a set of templates is then used to interpolate these values to provide frames of formant data at the required rate. Only coarticulation occurring between individual pairs of allophones is normally considered. The effects of coarticulation can depend upon a much wider context, for example in the phase 'the toucan' the lip rounding in the first vowel of 'toucan' may cross the word boundary to the preceding 'the'.

NETWORK ARCHITECTURE

This paper proposes an alternative approach to phonetic modelling, in which computation of formant contours is performed by a mulit-layer backpropagation network (see Fig. 1) [4]. The network, along with a small set of interface programs, emulates the function of the lower phonetic task of the JSRU synthesis by rule system. The input to the system is in the form of a .plp [2] file containing a list of allophone names, with corresponding segment durations and fundamental frequency targets. The system output is a .soi file [2] containing frames of formant data for the Holmes parallel formant synthesizer.

The Input Layer

The input layer consists of two groups of neurons, each representing an allophone within the input file. The desired output of the network represents the change in the formant parameters during the transition between the allophones denoted by the neurons in the input layer.

At present a simple seven bit binary code is assigned to each allophone used in the JSRU synthesis by rule system. In future experiments, allophones will be encoded as a pattern of activation over a set

of neurons representing articulatory features such as the presence of voicing and place of articulation. In this way effects of coarticulation extending further than half way through an allophone will be modelled by the partial activation of the neuron corresponding to the coarticulated feature during the preceeding allophones.

The Hidden Layer

Currently a single hidden layer is used, consisting of an array of 64 neurons. Further experimentation will be needed in order to determine the optimal size of this layer. An acceptable compromise must be reached between the acurracy with which formant data is reproduced and the amount of generalization which takes place. If the hidden layer is too large, the network may simply act as a lookup table, a large training set will then be required as little phonetic knowledge is reproduced through generalization. If the hidden layer is too small, too much generalization will take place, whilst most allophones will be accurantly reproduced, important exceptions will be ignored.

The Output Layer

The output layer consists of ten groups of six neurons. Each group corresponds to one of the eleven variable control parameters of the Holmes parallel formant synthesizer. Fn, the frequency of the nasal formant is omitted as this may remain constant without adversely affecting the quality of the speech produced. The trajectory described by each parameter during the transition between successive allophones is encoded in the form of a set of six sample values. Three samples are taken at uniform intervals from adjacent halves of each allophone (see Fig. 2). This decision is based on the assumption that rapid parameter changes are more likely to occur during transitions to and from short allophones than long. In order to capture rapid parameter transitions the intervals between samples must be small, therefore a higher sampling rate should be used during short segments. The intervals between samples vary according to the duration of each segment, and so a cubic spline is used to interpolate between samples to generate frames of formant data at the required rate.

TRAINING

Extensive use is made of formant data obtained from synthetic speech as this may easily be generated in large quantities. From this the network will be able to learn the basic form of formant transitions between pairs of allophones. Data obtained from natural speech is a difficult and time consuming procedure [5, 6]. Formant transcriptions of utterances of natural speech are required in order for the network to learn the effects of coarticlation which are not adequately modelled in synthetic speech.

Training is performed using Ansim, a neural network training package, running on an SAIC Sigma neurocomputor workstation, comprising of a 486 PC compatible microcomputer and an SAIC delta 2 floating point array processor. This provides a throughput during backpropagation learning of about .4 million connections per second. The normalization facility provided by Ansim was used to normalize the input and output data before training, to values between -0.5 and 0.5.

RESULTS

Initial results indicate that the effects of coarticulation can be modelled effectively by a neural network such as that presented in this paper.

A neural network has been trained to pronounce the words pit, pat, pot, put, putt, bit, bat, bot, but and butt. The subjective quality of the speech produced is similar to that of the JSRU speech synthesis by rule system from which the training data was obtained (see Fig. 3). This experiment provides a good test of the network's ability to model coarticulation, as the difference between allophones /p/ and /b/ are quite subtle.

Work is currently under way to extend the training data to provide modelling of other allophones in order to produce a more general speech synthesizer.

References

[1] M. A. A. Tatham. The representation and accessing of linguistic knowledge in simulations of language behaviour, a tutorial, 1989.

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

[3] J. N. Holmes. *Speech Synthesis and Recognition.* Van Nostrand Rienhold (UK), Wokingham, England, 1988.

[4] D. E. Rummelhart, J. L. McClelland, et al. *Learning Internal Representations by Error Propagation*, chapter 8. Massachusetts Instute of Technology, U.S.A., 1989.

[5] R. W. Schafer and L. R. Rabiner. System for automatic formant analysis of voiced spech. *J. Acoust. Soc. Am*, 47(2):634–648, 1970.

[6] J. D. Markel. The SIFT algorithm for fundamental frequency estimation. *IEEE Trans. AU*, 20(2):129–137, 1972.

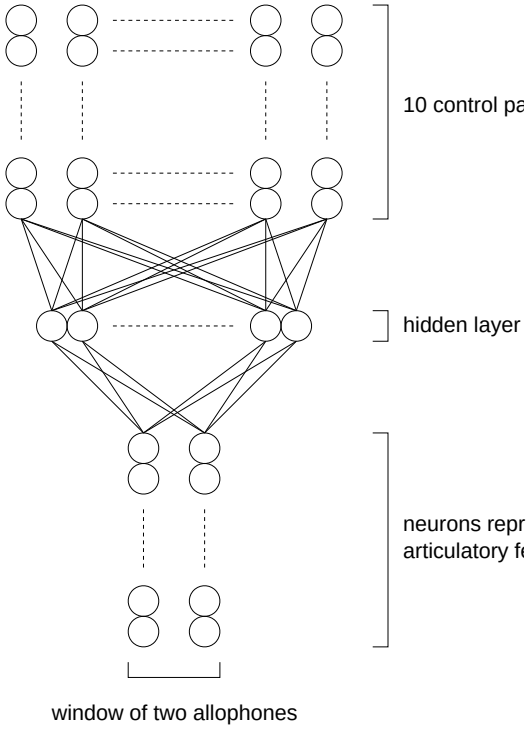

Figure 1: Schematic drawing of network architecture.

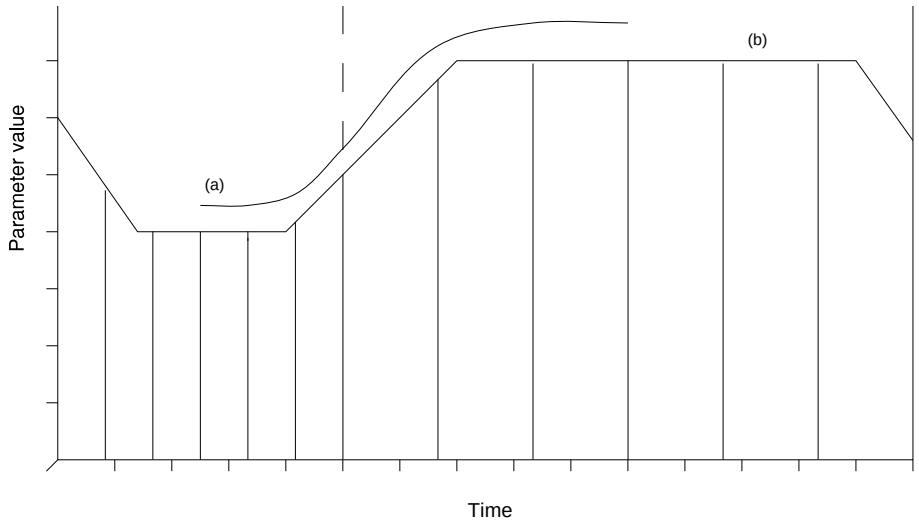

Figure 2: Example of a parameter contour for the transition between adjacent allophones, generated using a cubic spline (a) to interpolate between samples taken from the output of the JSRU synthesizer (b).

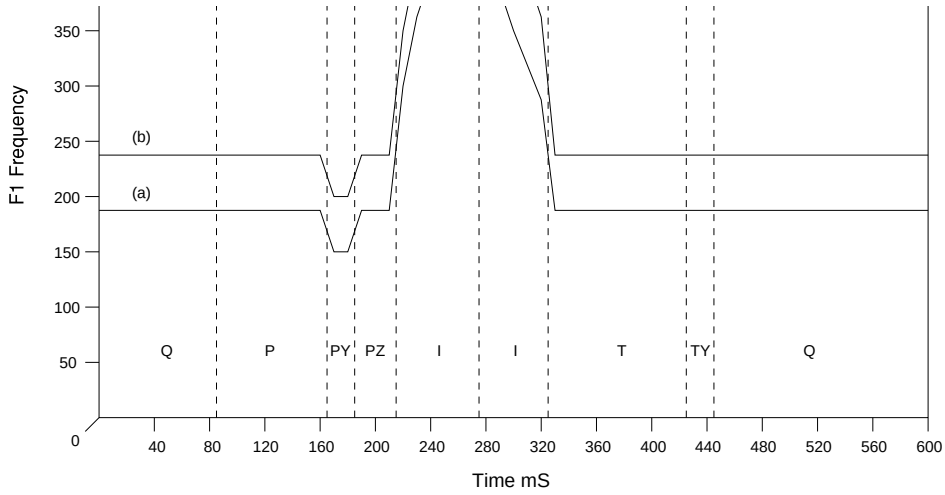

Figure 3: Frequency contour of the first formant for the word 'pit' (a) as produced by the JSRU synthesizer and (b) as reconstituted using a cubic spline from samples generated by the network (shown displaced by 50Hz for clarity).

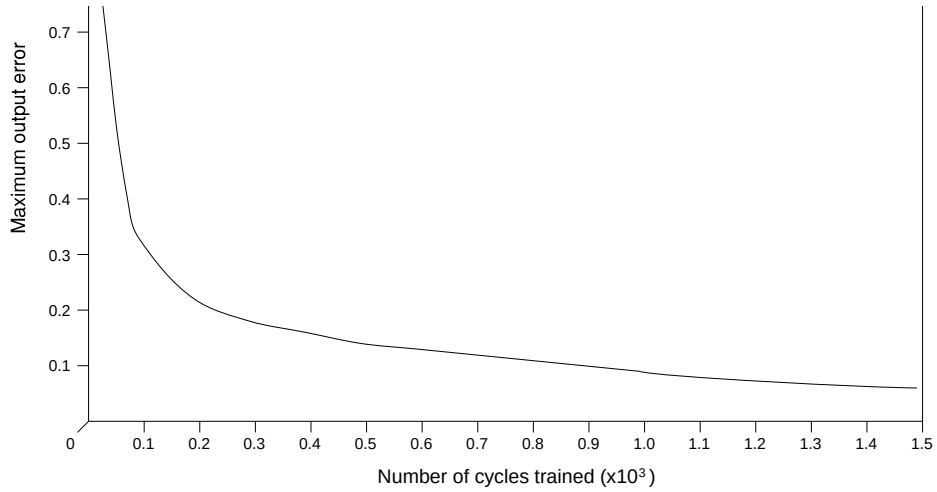

Figure 4: Graph of maximum output error against number of cycles through the training set.

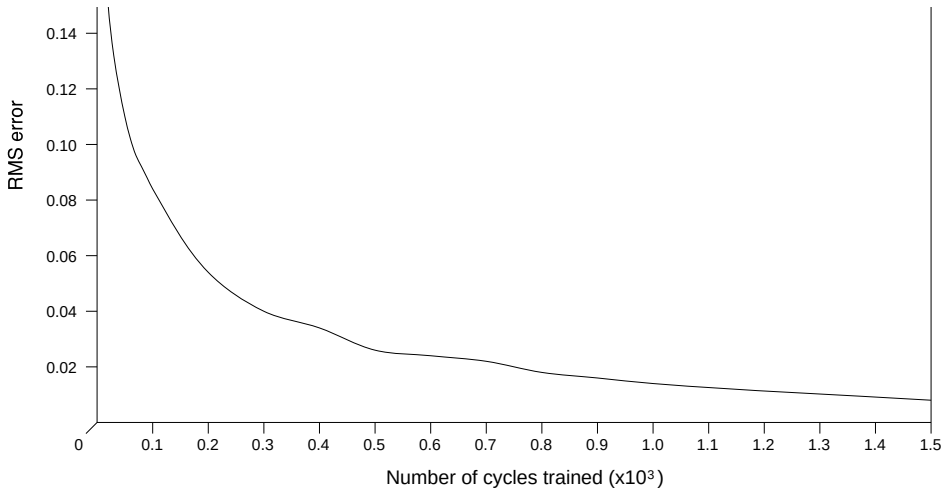

Figure 5: Graph of RMS error against number of cycles through the training set.