NEURAL NETWORK SOLUTIONS FOR IMPROVING ENGLISH TEXT TO SPEECH TRANSCRIPTION

P.R. Gubbins and K.M. Curtis
Parallel Processing Specialist Group, Department of Electrical and Electronic Engineering, University of Nottingham, England

ABSTRACT

This paper describes the application of a novel hybrid architecture to the text to speech mapping problem. The hybrid architecture combines the two previously prominent techniques used in this field: a rule base and a neural network. It aims to improve on the success rate of the neural network solution whilst reducing processing and greatly speeding up learning rate. A further neural network application which evaluates the rule base part of the hybrid system is also discussed.

INTRODUCTION

The first stage in the high quality synthesis of speech from unrestricted text is the process of converting from a code representing language graphically to one which represents the sounds that make up the signals we can decode through our auditory system.

The fact that we are dealing with unrestricted text is important when considering the techniques we are to use. A synthesis system that only needs to reproduce words from a limited vocabulary will be able to store sound code for whole or large portions of words. As the required vocabulary of a system increases however, the storage capacity and processing for the retrieval of such large blocks of sound code becomes impractically large. The longer and hence more complex the unit of sound that is stored the greater the number that is needed to be stored to produce all combinations of speech sounds in a language. For a system working from an unrestricted vocabulary it is only practical to store the most fundmental units of sound code, i.e. phonemes. The transcription from text to phonemes is therefore the problem we are faced with.

The historical derivation of the English Language makes it a particularly difficult case for transcription from text to phonetic code. English is derived from several widely variant sources with words and therefore sounds coming from Germanic, Latin, and Scandinavian based languages. The problem is compounded by the fact that these sounds must be encoded graphically by a relatively sparse alphabet, lacking accents, which in other languages multiply the number of vowel sounds that can be represented. The result of this feature of english is that in many cases a relatively wide context of letters has to be considered around a target letter in order to correctly predict the resultant phoneme.

APPROACHES TO TRANSCRIPTION

Previous research has concentrated on two main methods for transcription of text to phonetic code. The first of these is a rule based algorithm. In an attempt to handle the complexity in English described above this approach has led to a three layered algorithm. The alternative approach that has been widely researched, notably with the "NETtalk" system[1], is the use of a neural network to learn and then repeat the mapping from graphical to phonetic code.

Rule based algorithm

Before other techniques were developed systems for converting from graphemes to phonemes were based on a set of rules that indicated which particular phoneme a letter (grapheme) should represent depending on the context of the other letters surrounding it. Much of the later work on rule based techniques has been based on the set of rules developped by Elovitz et al [2]. A three layered rule based system was developed at Nottingham University [3] as an extension to the Elovitz letter to sound rules. This system used three parallel, hirearchical algorithms to analyse the text input and produce phonemic output. The algorithms are, in order of precedence, an exceptions dictionary, a morphological decomposition algorithm and letter to sound rules. Morphological decomposition involves recognising common letter groups (morphs) at the beginning and end of words and applying phonemic and stress information to the pronounciation of the whole word where it can be deduced from the presence of a particular morph. Examples of morphs are the premorph "con-" and the postmorph "-ation".

Neural Network Solution

As research into neural networks developed the text to phoneme transcription task was seen to be a suitable task to apply a neural net. solution to. Artificial neural networks (ANN's) are good at picking up statistical regularities but are not impeded by occaisional inconsistencies. The NETtalk system was developed [1] and has become the basis for further research.

NETtalk is a software simulation of a three layered structure, the three layers being input, hidden and output. The network has as its input a seven letter window with the central letter being the target and the remaining six being the context. This translates to 203 (29x7) binary 'nodes' in the input layer where each of the 7 input letters is set as one of the 26 letters of the alphabet or one of 3 punctuation symbols. The output layer consists of 26 binary units representing 26 articulatory features. Each phoneme is produced by a unique combination of these articulatory features. Between these two layers is the hidden layer of 120 units. Each input unit is connected to every hidden unit and each hidden unit in turn to every output unit, giving a total of 27,000 connections. Each connection has a weight value associated with it and each node a transfer function which evaluates the sum of its weighted inputs and produces a resulting output. The network learns by being presented with an input and also the correct output. The weights in the network are adjusted by means of a back propogation algorithm for each presentaion of an input and correct output pattern. Eventually, after very many presentations (e.g. 20,000 words) the weights in the network will reach stable values and the network will be able to generalise i.e. produce an output pattern for any given input pattern.

THE NEED FOR AN IMPROVED SOLUTION

Both of the above approaches, rule base and neural net, have their limitations. In the case of the rule base the problem is the need to produce a set of rules that are correct in all cases for a very complex language or to compensate for deficiencies in the rules by means of the exceptions dictionary. In the latter case the exceptions dictionary tends to become prohibitively large. In the case of the neural network solution the problem is in controlling the learning such that apparent anomalies in pronounciation but nevertheless valid cases are catered for. Possible solutions are to increase the size of the input layer to give wider context or to increase the size of the hidden layer or indeed the number of hidden layers. These solutions lead to an increase in the already very large processing overhead. A further consideration is that although the number of connections provides no difficulty in software implementation it renders a hardware implementation completely impractical.

HYBRID ARCHITECTURE SOLUTION

Research at the University of Nottingham has shown that the use of a novel hybrid architecture can provide a solution to mapping problems with the potential to improve results whilst reducing the processing overheads [4],[5].

The hybrid architecture combines a rule base and a neural network in a parallel sturcture. Figure 1 shows a block diagram of the structure. Rather that the neural network learning the correct output pattern for a given input pattern it learns the difference between the output given by the rule base and the correct output pattern. In effect the neural network learns to correct the output of the rule base. When generalising the two units work concurrently, each being presented simultaneously with the same input vector.

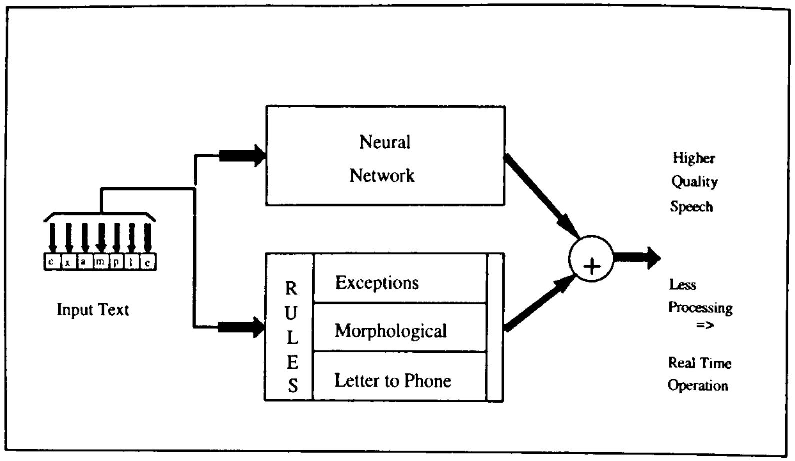

Figure 1. Block Diagram of The Hybrid Text to Phoneme Transcription Architecture

The system output is obtained by summing the output of the rule base and the ANN. The network complements the rule base in mapping the areas that it doesn't model. This can be said to model the way that we learn the relationship ourselves in that we are taught some rules when we initially learn to read, but do not consciously refer to them as our skill progresses, rather our brain develops subconscious patterns relating letter combinations to sounds. The contribution of the neural network allows the rule base to be very much cut down, therefore overcoming the need to specify an exhaustive set of rules. Similarly, due to the presence of the rule base the learning process of the neural net is greatly accelerated when compared with a purely network solution. As well as reducing the learning overhead this leads to the possibility of reducing the size of the network and hence the number of connections. As mentioned above this is a desireable situation when considering a hardware implementation of a text to speech system.

SELF ORGANISING NEURAL NETWORK USED TO EVALUTE RULE BASE REDUCTION

One of the major problems to be faced when investigating the use of the hybrid architecture for the text to phoneme transription task is deciding on the most appropriate way to cut down the rule base so as to maximise the performance of the hybrid system. This leads to a secondary use of neural networks in this research. A self-organising network, using a radial basis function algorithm, is used to examine the output of the rule base system. Whereas supervised networks such as NETtalk learn to map inputs to outputs by means of many correct examples of the mapping, unsupervised or self organising networks simply group inputs within the space provided by the network according to similar properties[6],[7]. Analysis of the way in which the output phonemes of the hybrid system group themselves in such a network will enable the selection of the best way to cut down the original rule base to give the optimum results in the combined system. It has been found that certain mapping phenomena, such as the difference between initial 's' and 'z' sounds, are differentiated only very late in the learning process of the neural net alone solution [8]. With the knowledge of such effects the choice of which rules are to be retained will enable these phenomena to be counteracted.

CONCLUSIONS

Extensive investigations are currently being carried out into the effectiveness of the hybrid architecture with many different reduced rule base and reduced size neural network combinations. Initial results indicate that a system with a very greatly reduced rule base and the original sized network can acheive correct transcription rates as good as NETtalk (91% after training on a set of 20,000 words) after substantially fewer training iterations (circa 50%). It is envisaged that once results of the self organising network have been applied to the system the success rate will increase. Trials involving reducing the network size have yet to be initiated.

The application of either a rule based or a neural network based solution to text to speech transcription requires a great deal of initial work to be carried out either in the form of producing a set of rules or in preparing an extensive training database of correct phoneme transcriptions. The use of a hybrid sytem as described potentially reduces these tasks as only simplified rules are required and training data requirements are reduced. A further feature of the english language to be considered is its very wide usage around the world. This leads to widely variant pronounciations as the language has followed separate development paths in different nations. The reduced requirements of the hybid solution put forward will facilitate the application of speech synthesis systems to different regional and national accents and indeed different languages.

REFERENCES

[1] Sejnowski, T.J. and Rosenberg, C.R., (1987), "Parallel netwoks that learn to pronounce English text", *Complex Systems 1*, pp145-168.

[2] Elovitz, H.S., Johnson, R., McHugh, A and Shore, J.E., (1976), "Letter-to-sound Rules for Automatic Translation of English Text to Phonetics", *IEEE Trans. Accoustics, Speech and Signal Processing*, 24(6)

[3] Asher,G.M., Curtis, K.M., Andrews, J and Burniston, J, (1990), "A Parallel Multialgorithmic Approach for an Accurate and Fast Text to Speech Transcriber", *ICSLP(90)*, Kobe, pp813-816.

[4] Burniston, J.D., Curtis, K.M. and Craven, M., (1992), "A hybrid rule based/ rule following parallel processing architecture", *PACTA'92*, Barcelona.

[5] Burniston, J.D. and Curtis, K.M., (1994), "A hybrid neural network/ rule based architecture for diphone speech synthesis", *ISSIPNN'94*, Hong Kong.

[6] Kohonen, T, (1989), *"Self-Organisation and Associative Memories"*, Springer Verlag.

[7] Wilde, S.A. and Curtis, K.M., (1994), "A Transputer Based Mixed Supervised/Unsupervised Neural Network For Speech Recognition", *Transputer Applications and Systems '94*, IOS Press.

[8] Craven, M.P., (1993), *Inter chip communication in an analogue neural network utilising frequency division multiplexing*, PhD Thesis, University of Nottingham