Proceedings of the Institute of Acoustics

GENERALIZATION IN NEURAL SPEECH SYNTHESIS

Gavin C. Cawley (1) and Mike Edgington (2)

(1) B.T. Laboratories, Martlesham Heath, Ipswich, Suffolk, U.K.
(2) School of Information Systems, University of East Anglia, Norwich, Norfolk, U.K.

ABSTRACT

Previous research (e.g. Cawley [1, 2]) has demonstrated that artificial neural networks can be trained to generate the speech sounds corresponding to a sequence of phonetic tokens, including the effects of coarticulation required to produce natural sounding synthetic speech. The principal limiting factor in the performance of neural speech synthesizers has been found to lie in the amount of training data available. This paper presents the initial results of an investigation to determine the amount of training data required to reach optimal generalization in neural speech synthesizers, through an empirical exploration of the effects of the number of training patterns on test set error.

1. INTRODUCTION

Speech is produced as the result of a coordinated sequence of movements of the articulators, such as the lips, tongue and jaw. For a given language, there exists a set of elementary *linguistic* units, known as *phonemes*. The elementary *acoustic* unit of speech is the *allophone*, a symbolic representation of a number of subtly different speech sounds corresponding to a given phoneme. For example, the *light* l in "lemur" and the *dark*, or *syllabic* l in "eel" are both allophones of the phoneme l. The acoustic realization of an allophone in continuous human speech, referred to as a *phone*, varies greatly according to phonetic context. This is partly due to the physical inertia of the articulators themselves, and partly due to cognitive processes that seek to minimize the articulatory effort required to achieve error free communication. The sources of these variations are described by three terms:

- **Assimilation** is the process by which an allophone partially acquires the acoustic properties of adjacent speech sounds, to prevent the undue vocal effort required to articulate each sound distinctly.
- **Reduction** occurs when the principal articulator is unable to move with sufficient speed, without undue articulatory effort, and so undershoots its target position.
- **Coarticulation** describes the simultaneous movement of two, or more, articulators. This word is also used as a blanket term to describe the general merging of speech sounds in continuous human speech.

This research was supported by the British Telecommunications short term research fellowship programme and by a Nuffield Foundation award (ref: SCI/180/94/395/G).

While coarticulation is often assumed to be local to the immediate phonetic context, in some cases the effects can extend much further. For example in the phrase "the toucan", the lip rounding gesture required to produce the initial vowel sound in "toucan" can even cross a word boundary to occur during the articulation of the word "the". Coarticulatory effects carry little of the semantic meaning of an utterance. However, the human auditory system has adapted to expect these variations to be present in natural speech. As a result, synthetic speech where this variation is absent or inadequately modelled sounds stilted and unnatural. There are two basic approaches to speech synthesis, concatenative synthesis and synthesis by rule, that adopt different models to account for the effects of coarticulation.

1.1 Concatenative Speech Synthesis

Concatenative speech synthesis systems simply concatenate short, pre-recorded speech sounds to form the required utterance. The most frequently used speech unit is the diphone, consisting of the second half of an allophone and the first half of the subsequent allophone. The diphone captures the effects of coarticulation in the transition between the two allophones, and abut during the relatively steady state conditions in the middle of each allophone, so that the joins between diphones are less noticeable. Modern concatenative synthesizers, for instance B.T.'s *Laureate* system (Page and Breen [3]), often incorporate time domain algorithms to smooth the boundaries between diphones and to modify the duration and pitch of each allophone to model *prosodic* effects (e.g. Moulines and Charpentier [4]). Adding a new voice to a concatenative synthesizer requires a phonetically transcribed speech corpus, of sufficient size and diversity to form an adequate inventory of diphones. The phonetic transcription process has normally been performed manually. However, completely automated training of concatenative synthesizers may soon be practical, due to advances in automatic alignment techniques.

1.2 Speech Synthesis by Rule

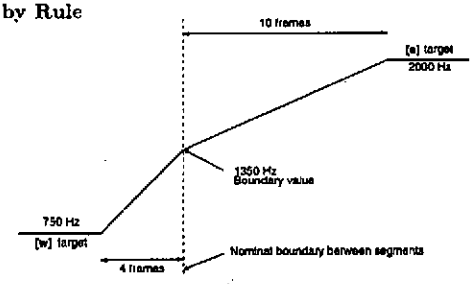

Figure 1: Second formant transition for the sequence we, using the Holmes-MattinglyShearme algorithm. After Holmes [5].

Speech synthesis by rule systems incorporate a model of coarticulation based on the interpolation of formant parameters, according to a fixed template, for example the Holmes-Mattingly-Shearme (HMS) algorithm (Holmes *et al.* [6]) employed in the Joint Speech Research Unit (JSRU) synthesizer (Lewis [7]), as shown in figure 1. Tables are compiled containing interpolation parameters for each speech parameter, for each allophone. A large set of context sensitive rules may also be necessary to achieve acceptable speech quality. The compilation of tabulated interpolation parameters and rule base involves extensive manual comparison of human and synthetic speech spectra. As a result *revoicing* a speech synthesis by rule system is an expensive operation.

1.3 Neural Speech Synthesis

As coarticulation is largely the result of the physical and cognitive processes, it seems sensible to suggest that the speech sounds associated with a given allophone will vary with phonetic and prosodic context in a systematic, *generalizable* manner. The artificial neural network has been demonstrated to have the ability to generalize knowledge extracted from a set of representative examples. The use of neural networks in speech synthesis has been investigated independently by a number of researchers (e.g. Cawley [2], Teurk *et al.* [8, 9], Scordilis and Gowdy [10] and Karaali *et al.* [11]). It is hoped that neural models of coarticulation can be trained without the extensive manual effort required to voice a synthesis by rule system, but without incurring the high storage requirements of concatenative synthesizers.

the remainder of this paper is organized as follows: Section 2. describes the implementation of neural speech synthesizers and describes the method used to estimate the amount of training data required to reach optimal generalization. Section 3. presents initial results obtained for a single allophone **h**, and section 4. concludes.

2. METHOD

Figure 2 shows the basic neural architecture employed in this research. The input layer of the network contains three groups of neurons representing the current and left and right context allophones, according to a vector of articulatory and prosodic features. The input layer forms a "sliding window", similar to that used in the Net Talk system (Sejnowski and Rosenberg [12]), over a stream of phonetic symbols corresponding to the desired utterance. The phonetic symbols move from left to right across the window, at each step the network generates an appropriate sequence of speech parameters to synthesize the current allophone, including the effects of coarticulation consistent with the immediate phonetic context. In order to generate the sequence of speech parameters, first the appropriate pattern of activation is applied to the three groups of neurons corresponding to the current and context allophones. A continuous value, representing the normalized duration of the current allophone is applied to the allophone duration input neuron, and a ramp input applied to the *time index* neuron. The network is trained so that as the input to the time index neuron steadily increases, the output units trace out the appropriate sequence of speech parameters. A discussion of the interpolation properties of these networks can be found in Cawley [2].

The most direct method to determine the optimal size of the training set, and the approach adopted here, is simply to train a large number of neural speech synthesizers with training sets of different sizes and to record the minimum root-mean-square error over a test data set achieved by each network. A scatter plot of the resulting data, against the size of the training set could be expected to have an exponential decay characteristic, as shown in figure 3. Clearly a network trained with a very small data set is unlikely to generalize well as the training set is unlikely to provide adequate coverage of the sounds corresponding to an allophone in a range of different phonetic contexts. The rate at which the generalization error is reduced will fall as training set grows in size. This is because a new pattern introduced into a large data set is less likely to be different to an existing pattern than for a new pattern introduced to a small data set. This suggests an exponential decay in test set error as the size of the training set increases.

If the exponential curve exhibits significant downward slope for networks trained on large data sets this suggests that the data set is too small, and a useful improvement in test set error might result if more training patterns were available. Conversely if the exponential curve rapidly becomes near horizontal, the neural network generalizes well given only a small sample of the available data, and so the data set too large in the sense that unnecessary effort was expended in its collection. Ideally the curve should display

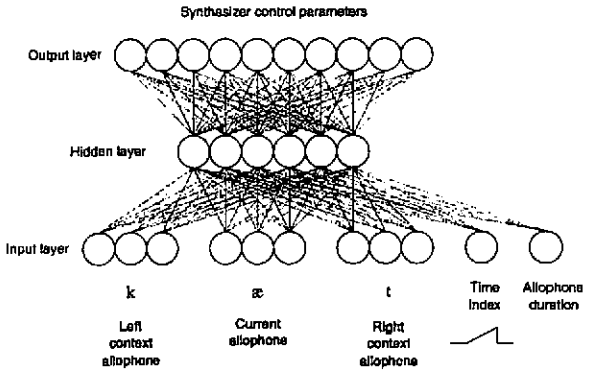

Figure 2: Basic neural architecture employed in this research.

steady reduction in test set error, leveling off only as the proportion of patterns used to form the training set approaches 90%. This would imply that the neural network is able to extract all of the generalizable information from the training set, but that adding further training patterns would have little effect on the test set error.

3. RESULTS

A pitch synchronous, twelfth order line spectral pair analysis (Sugamura and Itakura [13]) was first performed on a corpus of some 230 phonetically balanced sentences (approximately 10000 phonemes) of English speech, spoken by a male speaker with a received pronunciation accent. The resulting data were then partitioned to form a data set representing each allophone. A separate neural network was then trained on each data set. Previous experiments indicated a hidden layer of 16 neurons to be more than sufficient for networks trained on 90% of the data, larger numbers of hidden units providing only a minimal improvement in test set error. One hundred trials were performed for each data set, using between 10% and 90% of the available patterns to form the training set. In each case, the remaining patterns were used to form a test set. It should be noted that if the training set is large, fewer patterns will be left to form the test set. This implies that the measure of generalization will be much less reliable for networks trained using large training sets. A small data set may not be sufficiently large to be statistically representative of the underlying distribution and will also be sensitive to outliers or artifacts introduced by the random partition of the data between test and training sets. We should therefore expect to see much greater variance in the test set error for networks trained with large training sets than those trained with small training sets (and therefore a more substantial test set).

The neural networks were trained using an implementation of the back propagation algorithm (Rumelhart *et al.* [14]), written in the C programming language using the Parallel Virtual Machine (PVM) package [15], running in parallel on a network of 10 Linux workstations. Figure 4 shows a scatter plot of root-mean-square test set error for a neural network trained to produce the allophone **h** in different phonetic contexts, trained using between 10% and 90% of the 163 examples of this phoneme contained in the speech corpus.

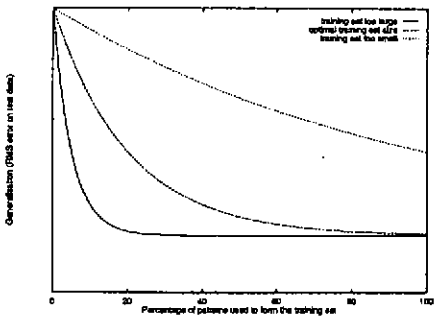

Figure 3: Expected exponential decay characteristic for scatter plots of root-mean-square error over the test set against the proportion of the available patterns used to form the training set.

A model of the observed exponential reduction in test set error with increasing size of the training set is also shown. A model of the form

$$y = A + Be^{Cx}$$

is used, where and A, B, and C are constants found using a least squares optimisation procedure. Table 1 shows the coefficients obtained for the allophone h. The expected exponential decay characteristic of the scatter plot is clearly evident. It can be seen that the slope of the graph still decreases significantly for large training sets, suggesting that while the corpus used approaches the optimal size for a neural network with 16 hidden layer units, a larger corpus may yield an improvement in generalization. As expected, the variability of the test set error when large training sets are used (80–90% of the available patterns) is very high, as fewer patterns are available to form the test set.

Coefficient	Value
A	0.0895844
B	0.0563094
C	0.0305952

Table 1: Coefficients obtained for an exponential model of test set error against the proportion of patterns used to form the training set for the allophone **h**.

4. CONCLUSIONS AND FURTHER WORK

At the time of writing only results for a single allophone (**h**) are available. These results suggest, for at least this allophone, that while the corpus used in this research is sufficient to obtain meaningful results, neural speech synthesis systems may benefit from a somewhat larger speech database. It should be noted that the multi-layer perceptrons used in this research contained a hidden layer consisting of a relatively small number of units. Clearly in order to reliably estimate the amount of generalizable information that can be extracted from the corpus, either a very large hidden layer in conjunction with regularization, or a

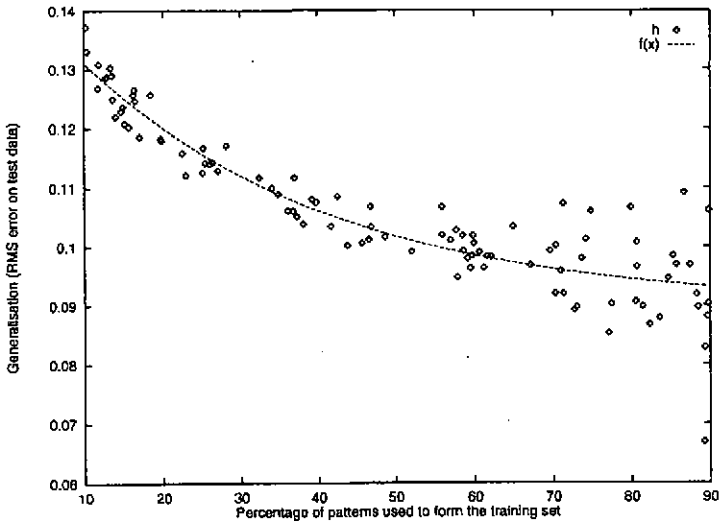

Figure 4: Scatter plot of test set root-mean-square error against the proportion of available patterns used to form the training set, for the allophone **h**.

constructive training algorithm, should be used, to ensure that hidden layer size is not a limiting factor in minimizing test set error.

5. REFERENCES

[1] G. C. CAWLEY & P. D. NOAKES, Allophone Synthesis Using a Neural Network', In *Proceedings of the World Conference on Neural Networks*, pp 122–125, Portland, Oregon, USA, 1993.

[2] G. C. CAWLEY, *The Application of Neural Networks to Phonetic Modelling*, PhD thesis, Department of Electronic Systems Engineering, University of Essex, Colchester, Essex, U.K., March 1996.

[3] J. H. PAGE & A. P. BREEN, 'The Laureate text-to-speech system - architecture and applications', *BT Technology Journal*, 14(1):57–67, (January 1996).

[4] E. MOULINES & F. CHARPENTIER, 'Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones', *Speech Communication*, 9:453–467, (1990).

[5] J. N. HOLMES, *Speech Synthesis and Recognition*, Van Nostrand Reinhold (UK), 1988.

[6] J. N. HOLMES, I. G. MATTINGLY, & J. N. SHEARME, 'Speech Synthesis by Rule', *Language and Speech*, 7:127–143, (1964).

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

[8] C. TUERK, P. MONACO, & A. ROBINSON, The Development of a Connectionist Multiple-Voice Text-to-Speech System', In *Proceedings of the I.E.E.E. International Conference on Acoustics, Speech and Signal Processing*, pp 749–752, 1991.

[9] C. TUERK & A. ROBINSON, Speech Synthesis using Artificial Neural Networks Trained on Cepstral Coefficients', In *Proceedings of the European Conference on Speech, Communications and Technology (EuroSpeech-93)*, September 1993.

[10] MICHEAL S. SCORDILIS & JOHN N. GOWDY, Neural Network Based Generation of Fundamental Frequency Contours', In *Proceedings of the I.E.E.E. International Conference on Acoustics, Speech and Signal Processing*, pp 219–222, 1989.

[11] O. KARAALI, G. CORRIGAN, & I GERSON, Speech synthesis with neural networks (invited paper)', In *Proceedings of the World Conference on Neural Networks*, pp 45–50, San Diego, California, USA, 1996.

[12] T. J. SEJNOWSKI & C. R. ROSENBERG, 'Parallel Networks that Learn to Pronounce English Text', *Complex Systems*, 1:145–68, (1987).

[13] N. SUGAMURA & F. ITAKURA, 'Speech Analysis and Synthesis Methods Developed at ECL in NTT -from LPC to LSP-', *Speech Communication*, 5:199–215, (1986).

[14] D. E. RUMELHART, J. L. McCLELLAND, & the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1: Foundations, The MIT Press, 1988.

[15] A. GEIST, A. BEGUELIN, J. DONGARRA, W. JIANG, R. MANCHEK, & V. SUNDERAM, *PVM: Parallel Virtual Machine — A Users' Guide and Tutorial for Networked Parallel Computing*, MIT Press, 1994.