AN OVERVIEW OF THE SPEAKER VERIFICATION PROJECT AT THE UNIVERSITY OF QUEENSLAND

Tom Downs, Ah Chung Tsoi, Mark Schulz, Brian Lovell,
Michael Barlow, Ian Booth, David Shrimpton, Brett Watson.
Intelligent Machines Laboratory,
Department of Electrical and Computer Engineering
The University of Queensland

ABSTRACT

This paper gives an overview of a project on speaker verification at The University of Queensland which is funded under a Syndicated Research and Development Program.

INTRODUCTION

In 1991, The University of Queensland initiated a number of projects that were funded under an external Syndicated Research and Development program. This program was established by private investors who were interested in investing in pre-competitive, but potentially commercial, research projects. One of these projects concerns speaker verification and is being conducted in the Department of Electrical and Computer Engineering at The University of Queensland. This paper gives an overview of the project.

AIMS OF THE PROJECT

The major aim of the project is to develop a system which can verify a speaker's identity in a non-ideal (i.e. potentially noisy) environment. Such systems have been developed in the past (see for instance Atilli et. al., 1988) but they have met with only limited success. Previous systems have been based upon conventional computing approaches, using the Von Neumann computer architecture and signal processing algorithms developed for that architecture. Our approach is based upon artificial neural networks and, as is explained below, it is our expectation that a superior and commercially-viable system will be developed.

Speaker Verification

A major aim of current speech research is the development of systems for speech recognition, i.e. machines that can recognise the spoken word. A long term goal is to develop speaker independent systems, i.e. systems that perform with high accuracy regardless of differences in speaker characteristics. In contrast, in the case of speaker verification, it is the differences in speaker characteristics that we seek to exploit in order to distinguish between different speakers.

In speaker verification, the speaker informs the system of his/her identity and the system then attempts to verify this claim. Note that the speaker verification problem is somewhat simpler than the problem of speaker *identification*. In the latter, the speaker's identity is not known, but the speaker is assumed to belong to a known population. Thus, a speaker identification system recognises a person's identity by comparing that person's speech characteristics with the characteristics of each member of the known population. In speaker verification, the speaker asserts his/her identity and the system seeks to match the speech characteristics of the speaker with the stored characteristics of the person whose identity has been asserted.

Besides being somewhat simpler than speaker identification, a speaker verification system has more immediate commercial applications. Foremost among these are telephone transactions involving finance or the exchange of sensitive information where one or both parties to the transaction are required to identify themselves. Often PIN numbers or similar devices are employed to support the claimed identities, but such devices are not entirely secure. For instance, PIN numbers can be lost, stolen or forgotten. A reliable speaker verification system would clearly provide additional security or support in cases such as this. Other applications, such as the gaining of access to sensitive areas or secure buildings, also clearly exist.

Artificial Neural Networks

Artificial neural networks are learning systems that are based on some quite simple processes that are believed to underlie learning in the brain. The brain has billions of neurons that are connected together via synaptic junctions to form a network. The synaptic junctions provide connections of varying strengths and it is believed that when learning takes place, the connection strengths (or network weights as they will be called from here on) undergo changes. This learning mechanism is employed in artificial neural networks (ANNs).

The first proposal for ANNs (McCulloch & Pitts, 1943) pre-dates the Von Neumann computer. Developments such as the perceptron (Rosenblatt, 1958) in the 1950s created considerable excitement but the field lost momentum in the 1960s because of an inability to apply the above learning mechanism to anything other than very simple networks and also because of the primitive nature of integrated circuit technology. During the 1970s a very widely applicable learning procedure for ANNs was developed (backpropagation: Rumelkart et. al., 1986) and VLSI technology also began to emerge. With these developments in place the main limitations of ANN technology had been overcome and the potential for ANNs to implement sophisticated tasks could be thoroughly investigated. During the 1980s, these investigations have proceeded apace and it has become clear that ANNs have at least two major advantages over conventional computing systems. Both have major implications for our project and are as follows :

1. An ability to generalise
2. An ability to learn complex mappings (even when the true nature of the mapping is unknown).

The Generalisation Capability

The ability of ANNs to generalise has been a major reason for the exceptionally high level of interest in the subject in the last 10 years. As an illustration, it is not uncommon for ANNs to be trained as expert systems. They learn all the rules and when put into operation, if confronted by a situation for which they have not been given a rule, they tend to give an acceptable response by, in some sense, generalising on the rules upon which they have been trained. This is a clear improvement on conventional expert systems and has implications for speaker verification. Two basic problems with the development of a speaker verification system arise from the variable nature of the voice characteristics of an individual speaker and from the variable nature of the environment in which speech samples might be provided. If such variabilities are incorporated into the training set of speech samples, the resulting verification system should show a tendency to generalise and hence accommodate variations in speech characteristics and the speaking environment. This is not, however, a clear-cut issue. Only a very primitive theory of generalisation exists so far (see, for instance, Baum & Haussler, 1989; Ehrenfeucht et. al., 1989; Bartlett & Williamson, 1991) and we can only speak in terms of possibilities rather than guarantees. Experience gained over the coming months should give us an indication of the degree of generalisation to be expected when the system is operating in the field.

The Ability to Learn Complex Mappings

Another major reason for the high level of interest in the subject of ANNs is their ability to learn complex mappings, especially in situations where the processes involved are not fully understood. A well-known example of this is the NETalk system developed by Sejnowski and co-workers (Sejnowski & Rosenberg 1987). This is a system that learns to pronounce (through a voice synthesiser) arbitrary English text. The performance of NETalk is comparable to that of DECtalk which was developed by Digital Equipment Corporation. But because of the complexity of the mapping from text to speech, the DECtalk system took

10 years to develop and drew upon the expertise of a large number of linguists. In contrast, the NETalk system, which is ANN-based, took just a matter of hours to learn a substantial vocabulary.

It is clear that the process of speaker verification involves a highly complex mapping (from speech sample to verification of identity). The speech signal is both non stationary and nonlinear, but ANNs can accommodate both of these complications in a quite straightforward fashion. The neural elements in an ANN have nonlinear input/output functions (usually the tanh or logistic function is used) so that when an ANN learns a mapping (and usually it learns a set of mappings) any nonlinearities in the mapping are accommodated quite naturally. And if non stationary signals are involved in the learning process, a variety of suitable ANN architectures are available. The best known among these are the recurrent net (Pineda, 1989) and the time-delay neural network (Waibel et. al., 1989). The capabilities of both of these architectures (as well as others) in application to the speaker verification problem are currently being investigated and this is reported on elsewhere in these proceedings (Shrimpton & Watson, 1992).

RESEARCH AND DEVELOPMENT STRATEGY

The major and minor objectives of the project were listed as a number of quarterly milestones so as to keep track of progress. Some of the major milestones were as follows:

1. Obtain or collect a small speech database suitable for verification
2. Build a conventional verification system based on dynamic time warping
3. Implement and test other verification systems based on hidden Markov models, vector quantisation, and various neural network architectures
4. Obtain or collect a much larger high quality database to evaluate the algorithms
5. Collect a very large low quality database to evaluate the performance of the algorithms in "real-life" situations

We would have liked to purchase a dynamic time warping verification system rather than develop one in house, but there were no state-of-the-art systems readily available; most commercial systems were based on early and inferior algorithms which were not suitable for our project. The dynamic time warping system was required as a performance benchmark against which we could compare the performance of all the other techniques. A small high-quality database was collected to evaluate the performance of the verification systems.

At the end of 1991 (the first year of the project), five of the best students from our graduating class were employed for several months to develop other verification techniques based on hidden Markov models, vector quantisation and several neural network architectures. Some of these students have since enrolled as post-graduates and have continued their association with the project in 1992. The hidden Markov model and vector quantisation techniques were implemented both as conventional computing algorithms and using ANNs. Neural net implementations of these techniques are discussed in Bridle (1989) and Naraghi-Pour et. al. (1991). Most of the new techniques compare favourably with the dynamic time-warping method and each seems to offer advantages in certain situations. At the time of writing, these new techniques are still being evaluated and improved.

It was clear from the start of the project that the successful testing and evaluation of verification systems depended critically on the quality and size of the speech database. There was a need for a large high quality database to evaluate the performance of the systems on clean speech. Yet there was also a need for a still larger database of low quality speech to test the robustness of the system to "real-life" voice signals.

The smaller high-quality database is being collected under supervised conditions using a close-talking (head-set) microphone (to minimise the effect of mouth to microphone distance variations) connected directly to a digital audio tape machine. The larger low-quality database is being collected in an unsupervised manner by a voice-actuated entry system which is used to control undergraduate access to a computer laboratory.

DISCUSSION

Our major aim is to develop a competitive working prototype by the end of 1993. The results we have obtained from our conventional system and from other, more novel, techniques are highly encouraging.

It is particularly important, in attempting to develop a speaker verification system for application in a commercial environment, to gain some appreciation of the likely psychological response of users when they are first introduced to the system and later, as they become better acquainted with it. With our voice-actuated entry system for controlling undergraduate access to a computer laboratory, we are well-placed to gain this kind of information. The laboratory houses about 60 personal computers and is frequently very busy. Currently the entry system is being used simply as a means of collecting voice samples, but even in this simple mode of operation, we have encountered minor problems. When there is the delay caused by the requirement of providing voice samples can lead to considerable impatience. In an attempt to eliminate this problem, we have had to reduce the size of the speech sample requested. Note, however, that this does not indicate any potentially serious problem when our speaker verification system is implemented in a commercial environment because, prior to our introduction of the voice-controlled access system, students simply walked into the laboratory, and this is the source of their impatience. If a speaker verification system is introduced into, say, some form of banking transaction, the fact that a small delay is incurred in processing the transaction should not cause the user any great concern. Certainly queues at automatic teller machines are quite common and have been accepted as a necessary reality by users. Once we have gathered a sufficiently large low-quality database by means of our voice-controlled access system, the system will be changed so that speaker verification will be implemented. The manner in which the students adapt to this change will provide us with important information regarding the likely psychological response of future users of a commercial system.

Our final system will be based around artificial neural networks. Each speaker, when introduced to the system, will train an ANN to recognise his/her individual voice characteristics. The differences in these characteristics will be reflected in different strengths of synaptic connections in the neural network. The strengths of synaptic connections in ANNs are called network weights and, when a speaker seeks to have his/her identity verified, it is simply a matter of downloading the recorded weights of that speaker and doing the appropriate comparison. Thus, importantly, as the population of users increases, the complexity of the speaker verification problem will increase only linearly.

REFERENCES

Atilli, J.B., Savic, M., & Campbell, J.P. (1988) *A TMS32020 Based Real-Time Text Independent Automatic Speaker Verification System*, Proceedings IEEE Conference on ASSP, New York, 599-602.

Bartlett, P.L., & Williamson, R.C. (1991) *Investigating the Distribution Assumptions in the Pac Learning Model*, Proceedings 1991 Workshop on Computational Learning Theory, Morgan Kaufmann.

Baum, E.B., & Haussler, D. (1989) *What Size Net Gives Valid Generalization?*, Neural Computation, Vol 1, 151-160.

Bridle, J.S. (1990) *Alpha-Nets: A Recurrent 'Neural' Network Architecture with a Hidden Markov Model Interpretation*, Speech Communication, 9, 83-92.

Ehrenfeucht, A., Haussler, D., Kearns, M. & Valiant, L. (1989) *A General Lower Bound on the Number of Examples Needed for Learning*, Information and Computation, Vol 82, 247-261.

McCulloch, W.S. & Pitts, W. (1943) *A Logical Calculus of the Ideas Immanent in Nervous Activity*, Bulletin of Mathematical Biophysics, Vol 5, 115-133.

Naraghi-Pour, M., Hegde, M. & Bourge, F. (1991) *A Comparison of Two Neural Network Architectures for Vector Quantisation*, Proc. IJCNN Washington, Vol 1, 391-396.

Pineda, F.J. (1989) *Recurrent Back-Propagation and the Dynamical Approach to Adaptive Neural Computation*, Neural Computation, Vol 1, 161-172.

Rosenblatt, F. (1958) *The Perceptron : a Probabilistic Model for Information Storage and Organization in the Brain*, Psychological Review, Vol 65, 368-408.

Rumelkart, D.E., Hinton, G.E., & Williams, R.J. (1986) *Learning Internal Representations by Error Propagation*, chapter in "Parallel Distributed Processing, Vol 1", D E Rumelhart and J L McClelland (eds.), MIT Press.

Sejnowski, T.J., & Rosenberg, C.R. (1987) *Parallel Networks that Learn to Pronounce English Text*, Complex Systems, Vol 1, 145-168.

Shrimpton, D. & Watson, B. (1992) *Comparison of Recurrent Neural Network Architectures for Speaker Verification*, Proc. Fourth Aust. Int. Conf. Speech Science and Technology.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. & Lang, K. (1989) *Phoneme Recognition Using Time-Delay Neural Networks*, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol 37, 328-339.

HMM SPEAKER VERIFICATION WITH SPARSE TRAINING DATA ON TELEPHONE QUALITY SPEECH

M.E. Forsyth A.M. Sutherland J.A. Elliott and M.A. Jack
Centre for Speech Technology Research
UNIVERSITY OF EDINBURGH

ABSTRACT—Speaker verification experiments using discrete and semi-continuous HMMs with telephone quality isolated digits are reported. The models were trained with varying numbers of tokens, giving equal error rates of 14% and 16% respectively on single isolated digits, and 4% and 2% on a sequence of 7 isolated digits.

INTRODUCTION

The tasks of speech recognition and speaker verification have much in common. Currently the most widely used technique used in speech recognition is that of hidden Markov modelling (HMM). There are three main forms of HMMs, discrete (DHMM), continuous (CHMM) and semi-continuous (SCHMM) (also known as tied-mixture continuous). All three techniques can be applied to speaker verication.

Although verification and recognition are similar tasks, there is a basic difference. In speaker independent speech recognition the objective is to avoid discriminating between different speakers saying the same word. In verification the objective is to maximise this discrimination. Use of minimal training data per speaker is a perogative.

The first section describes the parameters of the HMM system used, including the performance measure used to evaluate the system. The second section contains the results obtained using a conventional DHMM for speaker verification using 5 and 10 tokens to train isolated word models. The third section shows the comparative performance of a SCHMM system using 10 training tokens.

THE SYSTEM

Database

A database of isolated digits recorded over the U.K. telephone network was used for all experiments. There are 12 digits, consisting of the digits 1 to 9 plus 'zero', 'nought' and 'oh'. The data was sampled at 8 kHz and was divided into blocks of 5 utterances of each digit. Three such blocks (A, B, C) were used in these experiments. Each block comes from a different recording session. The data were automatically endpoint-detected to remove excess silence, and to reduce the amount of memory required by the database.

An independent set of 20 speakers was used for training speaker independent codebooks. The first utterance of each word from the A block data of each of these speakers was used to train a discrete codebook of 256 codewords and a semi-continuous codebook of 50 probability density functions (pdfs). The codebook data consisted of a total of 35,000 cepstral vectors (20 ms frames with 15 ms overlap). A set of 12 cepstral coefficients were used as the parameter set for all experiments.

A second set of 11 speakers, 6 male and 5 female was used for training and an additonal 9 impostor speakers added for the verification tests. The 5 token models were trained with the A block data and the 10 token models were trained with the A and the B block data. In both cases the models were tested against the C block data, giving 5 true speaker tokens and 95 imposter tokens for each digit for each speaker. Results were collected over all speakers to obtain 55 true scores and 1,045 impostor tests for each digit.