SPEAKER VERIFICATION USING ARTIFICIAL NEURAL NETWORKS

Brian C. Lovell and Ah Chung Tsoi

Department of Electrical Engineering
University of Queensland, Australia.

ABSTRACT – Speaker verification is a process by which a machine authenticates the claimed identity of a person from his or her voice characteristics. A major application area of such systems would be providing security for telephone-mediated transaction systems where some form of anatomical or "biometric" identification (which cannot be lost, forgotten or stolen) is desirable. Due to the great potential shown by artificial neural networks (ANNs) in the field of speech recognition, we evaluate the performance of a variant of the multi-layer perceptron ANN in the task of speaker verification. To prove the concept, the technique is applied to the classification of 2 speakers using a single utterence. A clustering algorithm partitions the input and output synaptic weights of the trained networks according to a Euclidean distance measure and it is found that the input synaptic weights appear to effectively characterize the speaker. The results demonstrate that the chosen ANN model can be used for speaker identification and verification purposes.

INTRODUCTION

Speaker verification is a form of biometric personal identification. The term "biometric" is used to distinguish identification methods based on intrinsic characteristics of a person — such as voice, signature or fingerprints — from methods that use artifacts like keys and passwords. Biometric identification is generally considered to be more reliable than artifact identification because it is based on intrinsic characteristics of the individual which are difficult, if not impossible, to mimic.

The task of speaker verification is much less difficult than speaker identification since, in the verification task, the claimed identity is provided by the candidate and the system only has to confirm or reject this identity, whereas, in the identification task, the system must establish the true identity from the voice characteristics alone (Doddington [1985]). Thus a speaker verification system is analogous to verification using handwritten signatures since, in legal documents, the handwritten signature is always accompanied by a printed version of the person's name.

Although many speaker verification systems have been developed in the past and some have attained reasonable performance on high quality speech signals, their performance has generally been degraded when used on telephone quality speech. However, low quality, inconsistent voice channels should no longer be a problem once the Integrated Services Digital Network (ISDN) is fully established. This raises the real possibility of a bank or some other institution using a single speaker verification system to securely identify all of its customers via the telephone. The added security of voice verification would allow the banks to offer a far wider range of telephone services — effectively turning every telephone into an electronic teller.

Clearly, there is an enormous potential market for a reliable verification system but how should it be implemented?

CONVENTIONAL SYSTEMS

Over the years, researchers have designed a number of systems based on conventional speech processing techniques such as linear predictive coding (LPC), dynamic time warping and hidden Markov models with a reasonable degree of success. Fortunately, speaker recognition is one area of artificial intelligence where machine performance can exceed human performance (O'Shaughnessy [1987]). The system proposed by Atilli, Savic and Campbell (Atilli, Savic and Campbell [1988]) rejected legitimate candidates about 1.5 % of the time and accepted fraudulent candidates about 0.5 % of the time when operated in a text-dependent manner. Although these results are encouraging, it is uncertain whether conventional verification systems will ever come close to the levels of performance demanded by financial institutions.

ARTIFICIAL NEURAL NETWORKS (ANNs)

Recently there has been a great deal of interest in the application of artificial neural nets (ANNs) to problems of pattern and speech recognition. In these particular areas, ANNs frequently outperform sophisticated conventional algorithms and this is one reason to investigate the application of ANNs to the speaker verification task.

Another reason is due to an argument based on modelling strategy as follows: The traditional LPC-based technique is linear in nature, i.e., it assumes that the underlying signal is both linear and stationary. Since it is well known that a speech utterance is non-stationary, the conventional technique for handling this problem is to process the signal as a series of short segments. The LPC model coefficients for each segment are extracted and then dynamic time warping or hidden Markov models are used to match the extracted coefficients to some template model coefficients. This roundabout way of analysis is necessary because of the linearity assumption of the LPC technique.

On the other hand, the ANN technique uses a nonlinear signal model and, in a certain sense, it may be viewed as a nonlinear extension of the LPC technique. Because of this nonlinear model, it is expected that the range of validity of the signal model would be considerably wider than that of LPC-based methods and the model could remain valid for far longer speech segments.

In this paper, we conduct a preliminary study of the application of a particular ANN model to speaker recognition.

An ANN-based signal model

Consider a signal $y(t)$, $t = 1, 2, \ldots, N$. This signal can be nonlinearly modelled in the following manner using a multilayer perceptron with 1 output unit, p hidden units and m input units:

$$y(t) = \boldsymbol{k}^T \boldsymbol{z}(t), \quad \text{and} \quad \boldsymbol{z}(t) = f(\boldsymbol{A}\boldsymbol{u}(t) + \boldsymbol{\theta})$$

where

T denotes the transpose of the vector,
$\boldsymbol{k}$ is an $p \times 1$ vector,
$\boldsymbol{z}$ is a $p \times 1$ vector,
$\boldsymbol{A}$ is a $p \times m$ matrix,
$\boldsymbol{\theta}$ is an $p \times 1$ vector,
$\boldsymbol{u}(t)$ is an $m \times 1$ vector,

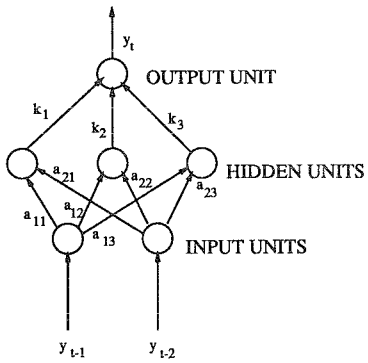

Figure 1: Multi-layer perceptron with a linear output unit, 2 input units and 3 hidden units.

and

$$\boldsymbol{u}^T(t) = [y(t-1), y(t-2), \ldots, y(t-m)].$$

The function $f(.)$ is nonlinear. In our case, it is defined by

$$f(\alpha) = \tanh(\alpha).$$

This model is a variant of the multi-layer perceptron ANN as illustrated in Figure 1. It is inspired by the ways in which the neurons in our brains appear to function. Vector $\boldsymbol{z}(t)$ represents the hidden layer neurons and vector $\boldsymbol{\theta}$ is known as the threshold vector. The nonlinear function is a way to abstract the nonlinear function of neurons. Matrix A is known as the synaptic connection matrix. Note that if $f(.)$ were replaced by a linear function, the ANN model becomes equivalent to the usual LPC model. Thus the ANN model may be considered as a nonlinear generalization of the conventional LPC model.

The unknown variables, A, $\boldsymbol{k}$ and $\boldsymbol{\theta}$ can be determined by minimizing the mean square cost criterion, J, defined by

$$J = \sum_N [y^d(t) - y(t)]^2$$

where $y^d(t)$ is the desired value of $y(t)$.

The unknowns can be obtained by differentiating J with respect to each of A, $\boldsymbol{k}$, and $\boldsymbol{\theta}$ and the parameters are updated by passing a given utterence through the ANN several times (typically about 15) and using a gradient descent algorithm as follows:

$$\boldsymbol{k}^{\text{new}} = \boldsymbol{k}^{\text{old}} + \eta \frac{\partial J}{\partial \boldsymbol{k}}, \quad A^{\text{new}} = A^{\text{old}} + \eta \frac{\partial J}{\partial A}, \quad \boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta}^{\text{old}} + \frac{\partial J}{\partial \boldsymbol{\theta}},$$

where η is a constant known as the step size.

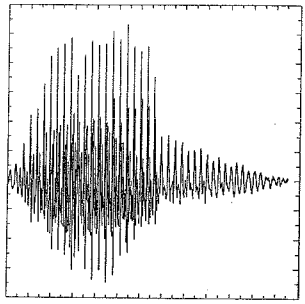

Figure 2: The original speech utterance "one".

$$\frac{\partial J}{\partial \boldsymbol{k}} = -[y^d(t) - y(t)]\boldsymbol{z}(t),$$

$$\frac{\partial J}{\partial \boldsymbol{A}} = -[y^d(t) - y(t)]\boldsymbol{\Lambda}\boldsymbol{f}'\boldsymbol{u}^T(t),$$

$$\frac{\partial J}{\partial \boldsymbol{\theta}} = -[y^d(t) - y(t)]\boldsymbol{\Lambda}\boldsymbol{f}',$$

where $\boldsymbol{\Lambda} = diag(\boldsymbol{k})$ and $\boldsymbol{f}'$ denotes the vector derivative of f evaluated at $\alpha = \boldsymbol{A}\boldsymbol{u}(t) + \boldsymbol{\theta}$ and is given by

$$\boldsymbol{f}' = \text{sech}^2(\boldsymbol{A}\boldsymbol{u}(t) + \boldsymbol{\theta}).$$

Note that it is possible to include a momentum term in the above update equations, to facilitate some leakage (forgetting of the past data). Notice also that in the ANN model, the number of hidden units is a variable. There are few theoretical results to guide us in the selection of this variable, so it is usually obtained by trial and error. From our own experience, the choice of the number of hidden units does not seem to be a great problem. It seems that, provided it is larger than a certain number, the ANN model will perform satisfactorily.

Application of the ANN model to a single speech utterance

We have applied the ANN model to analyze a single speech utterance — the word "one." Figures 2 and 3 show the results of applying such a model. The waveform corresponding to the word "one" is shown, together with the output from the ANN model. The ANN model is assumed to have 5 inputs, i.e., the past five outputs of the time series, and 10 hidden units. We settled on 10 hidden units after some initial experimentation.

We see that the ANN model can model the speech waveform rather well. Note that if we were to use an LPC-based method, it is unlikely that we could use a single LPC model over the entire speech utterance.

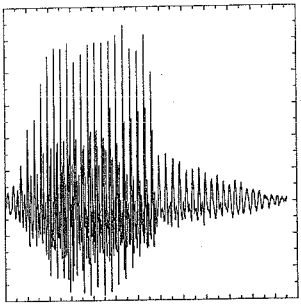

Figure 3: The estimated output from an ANN model.

To ensure that the estimated signal obtained from the ANN truly resembles the original utterance, we synthesized the signal, and played it back on a simple audio system. The utterance generated by the ANN sounds like "one", except that it has a slight metallic sound. This metallic sound might be attributable to the system used in synthesizing the sound rather than the signal itself, but at the moment we are not entirely sure.

Application of the ANN model to speaker verification

In our intended application, we wish to use the ANN to model a word spoken by a particular speaker. In traditional speech recognition, one is required to use some kind of matching algorithm which must ensure that the signal is correctly aligned in time with the template, eg., the dynamic time warping algorithm. But, in the ANN model, we no longer need to perform this alignment as we can use a single model to analyze the entire utterance.

In our case, we have decided to consider the coefficients of the ANN model and see how closely they are related to one another. Towards this end, we have separated the input synaptic weights (the matrix A), and the output synaptic weights (the vector k). We considered each input to the hidden unit to be a vector, and investigated its closeness to the synaptic weights obtained using another utterance of the same word by the same or a different speaker. A clustering algorithm is used to investigate the grouping of the vectors.

Figure 4 shows that the input synaptic weights of the ANN models corresponding to speaker a align quite well with one another and are well-separated from the input synaptic weights corresponding to speaker b. Some alignment can also be seen when the output synaptic weights are used as shown in Figure 5. Thus, the ANN model appears to be able to discriminate between the speakers.

CONCLUSIONS

In this paper, we have explored the possibility of using an artificial neural network (ANN) model to perform the task of speaker verification. The chosen implementation is a variant of the standard multi-layer perceptron, with linear output units. From cluster analysis, we found that the ANN-based model can indeed discriminate between different speakers. Our degree of confidence in the

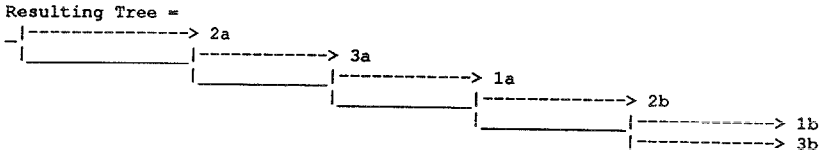

Figure 4: The clustering diagram of the input synaptic weights using three utterences from speakers a and b.

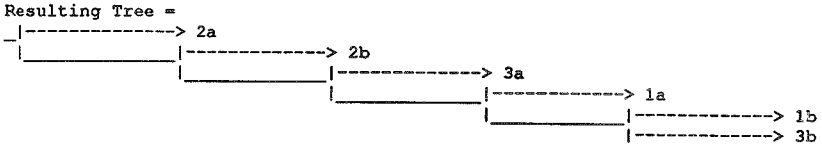

Figure 5: The clustering diagram of the output synaptic weights using three utterences from speakers a and b.

modelling of single utterances by the ANN was strengthened by listening to the estimated output from the ANN model, which was found to resemble the original utterance.

Although we have demonstrated that there is some merit in the ANN technique which we have adopted, this study must be considered as very preliminary. We have tested the method on only one utterance from different speakers. To ensure that the approach merits further attention, we will need to apply this technique to a larger vocabulary and a larger speaker community. This is work which we hope to report in a future publication.

REFERENCES

J.B. Atilli, M. Savic and J.P. Campbell [1988], *A TMS32020-based real time, text-independent, automatic speaker verification system*, Proc. IEEE Int. Conf. on ASSP, New York.

George R. Doddington [November, 1985], *Speaker Recognition — Identifying people by their voices*, Proc. IEEE 73, 1651–1664.

A. Lapedes and R. Farber [1987], *Nonlinear signal processing using neural networks: Prediction and signal modelling*, Los Alamos National Laboratory, Tech Report LA-UR87-2662.

R.P. Lippmann [1987], *An introduction to computing with neural networks*, IEEE ASSP Magazine 4, 4–22.

Douglas O'Shaughnessy [1987], *Speech Communication*, Addison-Wesley.