SPEAKER RECOGNITION USING ILS

M.P. Moody, R. Prandolini*

* School of Electrical and Electronic Systems Engineering,
Queensland University of Technology

ABSTRACT *A means of identifying a speaker from recorded passages of speech using the signal processing package ILS (Interactive Laboratory System) is described, which is efficient and sufficiently accurate for use in legal proceedings. Statistical dependence on variables such as the length of the utterances, numbers and types of parameters used and the length of segments of speech is investigated with the aim of determining the confidence level dependence on these variations. Success rates of distinguishing speakers from similar populations are about 90%for reasonably short samples (a few minutes for references and a few seconds for test samples),while even quite shorter reference samples may result in sufficiently significant results to add to the weight of evidence.*

INTRODUCTION

The acceptance of recorded evidence in legal proceedings is an increasing trend in Australian law courts. Many recordings are made in situations where the identity of individuals is difficult to establish due to lack of supporting evidence, particularly in the case of telephone calls or recordings made covertly. For instance, in the case of obscene or extortionate calls, the only identifying evidence is usually the voice of the caller. In the case of audio surveillance of a premises, the occupants' identities may not be known. An efficient and accurate method of identifying a speaker is required which gives a statistically significant result and which cannot be 'tricked' easily by attempts to alter the voice artificially. Although the pitch of speech can easily be altered (by speaking in falsetto, for example), the resonant characteristics of the vocal tract of an individual is much more difficult to disguise. Extreme nasalisation or talking through a tube may have significant effects, but since most speakers do not realise this, these means are not often resorted to.

A method of statistical analysis of speech using the functions available in the Signal Processing Package ILS has been developed which compares sets of average reflection coefficients (used to model the vocal tract discontinuities) against those of known suspects and other similarly spoken test subjects. The results of some preliminary tests with a limited population of similar speakers are presented, showing the potential of the method to identify efficiently an individual speaker from a set of known references.

SPEAKER CHARACTERISTICS

A number of properties of speech are directly dependent on the physical characteristics of the speaker. In order to understand this, the processes involved in generating speech should be understood.

Speech can be divided into voiced and unvoiced sounds, generated by different but

related mechanisms. In the case of voiced sounds, the vocal cords are vibrated with a certain pitch and air is passed through them. The interaction of the vibrating cords and the passage of air produces excitation impulses which are then passed through the throat and mouth and to some extent the nose, constituting the vocal tract. The effect of the vocal tract is to impose certain spectral characteristics to the sounds, which depend upon the resonances of the vocal tract. The major resonances are called formants. In the case of unvoiced sounds and fricatives, the process is similar, except that the excitation function is random noise produced by passing air through a constriction in the mouth, usually produced by the tongue, the teeth and the lips. This air is then passed through a restricted vocal tract formed by the front part of the mouth and lips.

The shape of the vocal tract and the resonances caused by it in the case of either voiced or unvoiced sounds is very much a characteristic of the individual. Although the muscle control exercised during speech slowly alters the shape of the vocal tract, causing the differences in the various sounds being produced, the average shape of the tract is relatively fixed and quite hard to disguise. It is this characteristic more than any other which determines the basic characteristic and tonal quality of an individual's speech.

PARAMETERS and VARIABLES

The basic recognition parameters used to characterise speech are the reflection coefficients corresponding to the variations and discontinuities in the cross section of the vocal tract. These are calculated for independent test and reference passages of speech and the best match is determined using Euclidean Distance measures. The number of these, the length of speech segments for which these are calculated, and the total length of time over which the coefficients are averaged are independent variables which will affect the results of the comparisons of test and reference samples. For instance, if the segments of speech which are used to calculate the reflection coefficients are too short, the low frequency resonances will not be extracted properly. On the other hand, if the segments are too long, the speech will not be stationary over the length of the sample, resulting in formant averaging which will disguise the true formant frequencies.

The Number of reflection coefficients used is also important, since too few will result in incomplete specification of the vocal tract, while too many will result in inaccuracies due to an over-specified and hence ill-conditioned system. Since robust algorithms are used to calculate the coefficients, this latter problem is not serious, but the extra time taken to calculate the extra coefficients may become excessive.

The final variable is the total length of time over which the coefficients are averaged, that is, the number of sets of coefficients which are averaged to produce the reference and each test sample. In some cases, test samples of only a few seconds may be available (from a very short telephone conversation for example), and some notion of the degree of confidence with which matching can be made will be necessary.

The tests presented below represent some preliminary trials to determine the sensitivity of the method to these variables. A single test using up to seven minutes of speech and twenty-nine reflection coefficients for each subject can take considerable time on an IBM386 personal computer with co-processor. For this reason, only preliminary results are available at this stage, comparing three

speakers with similar speech characteristics (Australian males, reading 15 minutes of independent passages of text).

RESULTS

The output of the matching process is a confusion matrix as shown in Figure 1, together with lists of Euclidean distances. The columns of the confusion matrix represent the reference passages of the three speakers (initials MM, RP, JG), and the rows represent the test passages. The sum of the row values gives the number of independent tests performed in the experiment (50 in the example below).

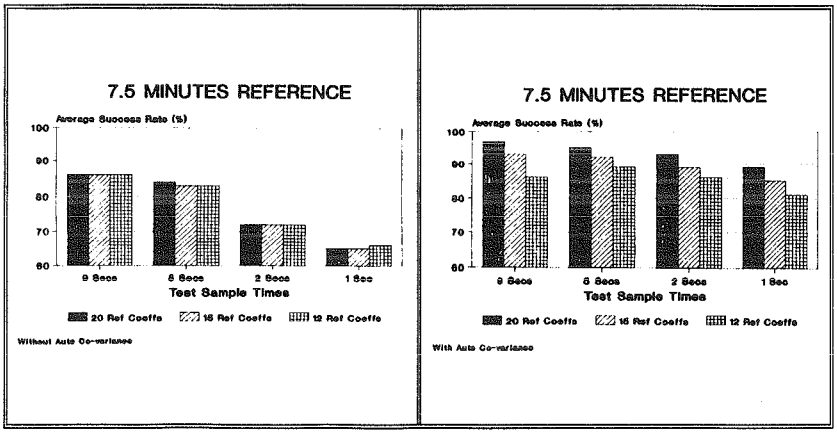
	MM	RP	JG
MM	48	0	2
RP	1	50	1
JG	1	2	47

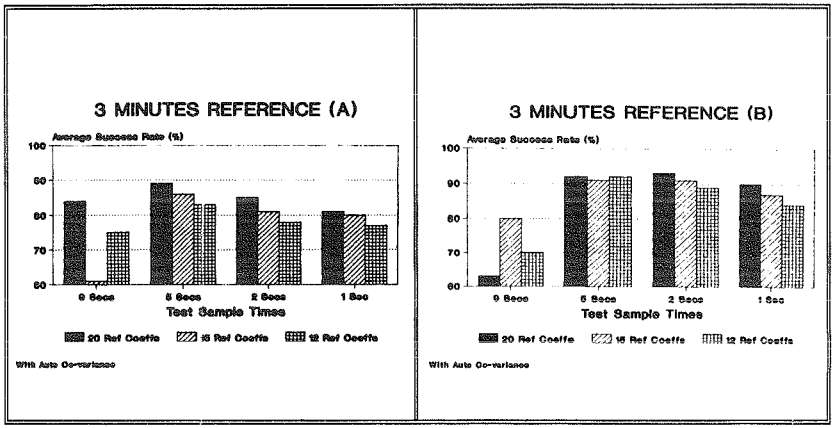
Figure1 Confusion matrix (50 tests)

An arbitrary measure of success of the matching process is the geometric mean of the diagonals of the matrix as a percentage of the number of tests. For example, the above matrix gives a success index of 97%. This simplistic measure is used to quantify the effects of the variables in the results below. From these tests, a recommendation for the optimum and minimum values of the variables can be made.

RESULTS

The graphs below give results for various combinations of the variables. The graphs are self explanatory, but it must be remembered that they are limited in extent. The number of test samples is taken such that the whole sample of 7.5 minutes of speech is used. The results appear to be sufficiently significant to be useful. It would appear that an optimum reference time is about 5 minutes of speech (normally available from police interviews) and test samples in of 2 seconds or longer are sufficient. At least 50 test samples are desirable, requiring a minimum of about 2 minutes of test speech.





CONCLUSIONS

The above results indicate a promising method for determining the identity of a speaker by matching samples of speech with those of a suspect and other reference subjects. Further tests with larger populations of different kinds (female, ethnic etc.) need to be made to determine proper levels of confidence. Limits on the Euclidean Distances must be set for the cases where the test samples do not come from the set of reference subjects, but these limits have not yet been determined. Limits will be necessary to prevent the samples from being 'assigned' to a certain test subject, since the method itself gives simply a 'best match'. The Euclidean distances themselves will give a measure of the confidence with which a match can be made.

The method uses readily available software, and can be produced in a 'turn-key' form for use directly by police, although some knowledge of the method is desirable to interpret the results with confidence. The method is superior to other methods such as those involving spectrograms, since it gives a single distance measure, and removes the subjective measures normally employed.

ACKNOWLEDGEMENTS

Acknowledgements are made to Support Staff and Post-Graduate Students who are associated with the Signal Processing Group. Although too numerous to mention by name, their efforts and assistance are greatly appreciated.

Using Probabilistically Conditioned Neural Networks to achieve Speaker Adaptation

David Bijl and Frank Fallside
Cambridge University Engineering Department
Cambridge, England.

15 October 1990

Abstract

Speaker Adaptation using Neural Networks is generally difficult because network weights are adjusted in accordance to a whole training set. Introduction of new adaptation data provides a problem, because back-propagation training would converge exactly on that test data, throwing away previously learnt information.

If a neural network is formulated via a probabilistic approach, it is possible to use concepts of maximum likelihood to adapt the parameters of the network so as to accommodate changes without discarding valuable information generalized from initial training.

Here, a probabilistic approach is demonstrated which allows speaker adaptation in automatic speech recognition. The units of speech used are phonic and prosodic.

1 Introduction.

Neural networks have provided performance in speech recognition tasks that compares more than favourably with the traditional statistical technique based on Hidden Markov Models (HMMs) [7].

Given a large database of speech, a network can be trained such that the weights defining the network yield minimum error pattern classification.

Every person has a more or less distinctive voice, which makes possible the identification of a particular person from their voice, albeit prone to error [4]. In speech recognition, it is useful to exploit this property by allowing adaptation of a speech recognizer to each speaker.

In order to achieve an understanding of what network adaptation might achieve, it is useful to relate a traditional minimum risk classifier to a neural network. Consider a set of independent classifiers, whose output $g()$ is some function of

an input vector x and a set of learnt parameters p. An input is deemed to belong to a class if the classifier modelling that class outperforms other classifiers.

$$g_i(x, p_i) = max_j\,(g_j(x, p_j)) \Rightarrow \text{ x belongs to class i}$$

Consider the output as being connected to a set of internal nodes and inputs. If a node is only connected to the nodes closer to the input than itself (ie to lower nodes), the network is said to be feed forward. Such networks may be trained by the back-propagation algorithm [8]. The input to hidden and output nodes may include both the actual network inputs and lower node values. Most commonly, only two layers are used. The first layer is a set of N_{hd} hidden nodes connected only to the N_{in} inputs. The second layer is a set of output nodes, which connect only to the hidden nodes. Connection is conventionally achieved by taking a linear weighted sum of the inputs.

$$f(p_i.x) = f\left(\sum_j^{N_{hd}} p_{ij} x_j\right)$$

$$= \sum_{j=0}^{N_{hd}} p_{ij}^{hd} \sum_{k=0}^{N_{in}} f\left(x_j p_{jk}^{hd}\right)$$

A network need not be constructed by a linear sum. Indeed more appropriate connection may be considered to be a functional approximator, represented by a polynomial or even transcendental function. A quadratic or higher order polynomial function provides a more general input space descriptor than a linear function, though orders higher than quadratic are not usually practicable because the number of terms required becomes very large.

A quadratic operator is easily described by a matrix notation. Let a_{ij} weight the contribution of the product of the i^{th} and j^{th} node, and define these weights by a real symmetric matrix A^{out}. Similarly, let the i^{th} hidden output be controlled by weights defined by matrix A_i^{hd}.

$$h_i = f\left(x^t A_i^{hd} x\right) \tag{1}$$

$$output_i = f\left(h^t A^{out} h\right) \tag{2}$$

Back propagation is easily adapted for this type of network [7], making training of the network possible.

2 The Aim of Adaptation

Consider the i^{th} output of a multi-layer perceptron. Let it be connected to the outputs of a set of lower nodes h_j, some of which may be input nodes, some of which may be hidden.