BRITISH ACOUSTICAL SOCIETY

"SPRING MEETING" at Chelsea College, London, S.W.3. on
Wednesday 25th April / Friday 27th April, 1973.

SPEECH AND HEARING: Session 'C': Speech Properties and Recognition

Paper No:

73SHC5

A Graph Theoretic Approach to Automatic Speech Recognition

John H. Warren

PTR Ltd., Taplow Court, Taplow, Maidenhead, Berkshire.

1. INTRODUCTION

It is the purpose of this paper to describe an unusual and novel technique for general pattern recognition, based on a graph theoretic approach, and to describe the application of this technique to automatic speech recognition.

There is an increasing awareness in almost all branches of pattern recognition that it is neccessary to consider the "structure" of patterns and not to regard a pattern as an assembly of statistically independent cells. This awareness leads to recognition processes characterised by firstly, a subdivision of the patterns into smaller units ("features" or pattern "primitives") and secondly, establishment of the (structural) relationships between these units. This provides a moderately compact "description" of the pattern in terms of the entities composing it; recognition is then accomplished by preparing a corresponding "description" of the unknown pattern and then comparing these descriptions. Such a process includes not only the capability of assigning the pattern to a particular class but also the capacity to describe aspects of the pattern which render it ineligible for assignment to another class (Reference 1).

Such structured approaches to pattern recognition have substantial commonality with computational linguistics and with finite state machine theory, and the generic term "syntactic pattern recognition", now generally accepted, recognises this. However, many problems still need to be solved; these include pattern description techniques, primitive selection, and recognition and inference procedures. It is the aim of this paper to contribute towards a solution to these problems.

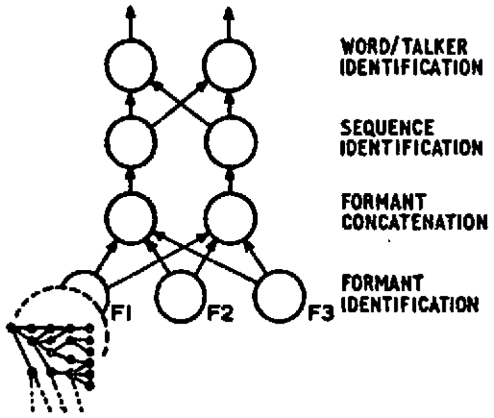
2. PATTERN DESCRIPTION

It has frequently been suggested that pattern descriptions should be hierarchic by nature, reflecting the hierarchically structured nature of the real world (which, in some abstract form, they represent). In such a scheme, pattern primitives are assembled (as nodes) into a single tree-like structure; it is then often alleged that the succession of nodes represents the necessary succession of hierarchic levels.

It is a contention of this paper that this approach perpetrates some measure of misconception and that a more accurate viewpoint would seek to represent the succession of hierarchic levels by a succession of tree-like structures. Furthermore, since many pattern recognition problems are characterised by the simultaneous presence of many different kinds of information, (speech, for example, containing information relating to speaker sex, identity, mood and regional accent as well as word identity), it seems logical to extend this first "succession" of structures so that it becomes one of many. Such a formulation of the pattern description problem enables several very significant problems to be overcome; it also offers a number of advantages not readily offered by the original formulation.

In such a system, recognition becomes very much a continuous process distributed across the network of graphs which represent the structure of the patterns being studied, rather than a discrete process based upon (for example) hypersurface discrimination.

Each structure (graph) can be constructed to extract primarily that information relevant to some particular object (eg: word identity, speaker identity) and to largely ignore information irrelevant to that object. The results of these decisions can be used to assist or enhance other decisions elsewhere in the network in a manner analagous to that which apparently occurs in the human brain.



TYPICAL GRAPH-SUPERGRAPH STRUCTURED RECOGNITION SYSTEM.

A graph theoretic approach to structured pattern recognition would seem to be advantageous since it offers an approach in which the structural relationships between pattern primitives can be abstracted, represented and manipulated; it may also offer a means whereby certain disadvantages of "syntactic" recognition schemes (notably, the limitation imposed by the use of single strings) could be overcome.

Formally, each of these individual structures is a graph; the recognition system is therefore based upon the use of a network of graphs, in which interconnection between the various graphs can occur. It should be noted that the complete recognition system can naturally also be regarded as a graph (at a higher level of abstraction) and that this "supergraph" forms a compact and unambiguous description of the overall structure of the system.

The use of a larger number of relatively small structures offers substantial advantages; it offers some form of inference (see Section 5) and permits very rapid searching to be undertaken.

3. PRIMITIVE SELECTION

It has been suggested (Reference 2) that the human perceptual mechanism recognises clusters on a local distance and nearest-neighbour basis. There is also some evidence that the human clustering mechanism is adaptive, in the sense that although some stimuli will invariably be classified uniquely, others may have a classification which is dependent upon their immediate environment. The use of graph-theoretic techniques (Reference 2) to locate clusters by fragmentation of the minimal spanning tree covering the point set has been investigated in detail from an automatic speech recognition viewpoint and can provide an effective means of segmenting isolated utterances into smaller units which correlate well with the sounds of speech as perceived by a phonetician, as well as providing a measure of environment dependent classification where necessary (Section 5).

The process is completely general, in that it does not assume any pre-defined constraints and any set of patterns could be offered. This generality means that the process could be used repeatedly, first for segmenting elementary speech sounds from isolated utterances and subsequently for dividing these in any appropriate manner.

4. PATTERN REPRESENTATION

The generalised representation of a pattern class in terms of the primitives selected and the relationships between these is not a trivial problem. A number of approaches to finite state machine synthesis are known but these in general require enough information to be included in the problem statement to ensure a unique solution. For many pattern recognition problems this is an unduly rigorous criterion and an alternative approach to the problem has been adopted. This approach is based upon, but supersedes, an approach adopted earlier (without success) by Steingrandt and Yau (Reference 4).

The existence of a set of patterns for "training" is assumed and it is also assumed that a set of primitives have been defined from these. The problem is then to construct a graph (finite state machine) to represent the pattern class from the samples of the class, as exemplified by the training patterns.

The obvious need is to minimise the size of the graph, in terms of the number of nodes and arcs in the graph (since this will maximise the use of each primitive). It can be shown (Reference 5) that additions to the graph will be necessary if the primitive does not appear in the graph but may not be necessary if the primitive is already present. A solution to the problem of additions has been demonstrated and an algorithm written which implements an optimum form of this solution.

5. PATTERN RECOGNITION

A variety of recognition strategies can be implemented using a syntactic recogniser. In all cases, it is essential that the unknown pattern be given a description in terms of the primitives of which it is composed and that this description be compared with the representation of the pattern class already derived. This implies searching the graphs in the network using some metric or similarity measure to define the result of the comparison process. None of this is difficult, although it may be tedious.

The use of many graphs, rather than one, provides a facility whereby a form of inferential classification can be achieved. If it becomes increasingly unlikely that the current path being followed in the graph at level N is the correct one, then this path can be retraced. If an earlier decision at some node was in fact wrong this must arise because of an erroneous output from the graph at level (N-1). The correct output can be predicted from level N, and it may well prove that this correct output is entirely acceptable on a similarity basis even if it is not the most highly similar decision possible.

This approach bears some intuitive similarity to human linguistic error correction, and a more direct similarity to the decoding processes of convolutional codes; this similarity is regarded as encouraging.

6. EXPERIMENTAL RESULTS

A group of computer programs have been written to implement a graph theoretic recognition system and these have been used for automatic recognition of spoken words. So far, only one fairly simple structure has been used (two layers, with a single graph in each layer) but even this shows a performance substantially better than the comparative performance associated with template matching, with error rates reduced by 2:1; more complex structures (involving three layers and five graphs) are being studied.

Performance levels of about 99% are currently being achieved for a small range of talkers using a spoken digit ("ZERO" - "NINE") vocabulary, "training" being based initially on five utterances of each word. More training than this seems to be valuable since at this level some utterance descriptions are neither encountered nor deducible from the training patterns.

7. ACKNOWLEDGEMENTS

This paper describes work carried out under the Joint Speech Research Unit (Ruislip) and their permission, as well as that of The Plessey Co. Ltd., to publish this paper is acknowledged. Acknowledgement is also due to Mr. J. N. Holmes and Mr. S. R. Hyde for the benefit of advice and discussion during the course of this work.

References:

1. Fu, K.S. "Introduction to Special Issue on Syntactic Pattern Recognition", Pattern Recognition, Vol. 3, No. 4, November 1971.

2. Zahn, C.T. "Graph Theoretical Methods for Detecting and Describing Gestalt Clusters", I.E.E.E. Trans. Computers, Vol. C-20, No. 6, January 1971.

3. Biermann, A.W. and Feldman, J.A. "On the Synthesis of Finite State Machines from Samples of their Behaviour", I.E.E.E. Trans. Computers, Vol. C-21, No. 6, June 1972.

4. Steingrandt, W.J. and Yau, S.S. "Sequential Feature Extraction for Waveform Recognition", Proc. SJCC, 1970, pp65-70.

5. Pynn, Mrs. C. Unpublished manuscript.

Footnote: Reference 3 provides additional material relevant to the synthesis of finite state machines described in Section 4.