SPEECH RECOGNITION BASED UPON A SEGMENT CLASSIFICATION AND LABELLING TECHNIQUE AND HIDDEN MARKOV MODEL

W. A. Mahmoud and L. A. M. Bennett

Department of Electrical and Electronic Engineering
University College of Swansea,
Swansea SA2 8PP U.K.

1 Abstract

A new structure for isolated-word speech recognition via vector quantisation (VQ) is described, namely the segment classification and labelling technique (SCLT). The proposed recognizer requires the generation of separate codebooks for the acoustically dissimilar events and then the merging of them to produce a single reference codebook. Three major acoustic events were considered, namely voiced, unvoiced and silence (V/U/S). The results show that the proposed structure has the capability of reducing the degradation of VQ in speech recognition and provides a better set of observations for the hidden Markov model (HMM).

2 Introduction

Two very important speech modelling techniques have been applied to speech recognition. They are vector quantisation (VQ) of the linear predictive coding (LPC), which is used for representing the short-term spectral characteristics of speech, and the Hidden Markov Model (HMM), which can be used for representing the long-term statistical characteristics of speech. The VQ generates an ordered set of reference codewords, referred to as the codebook, which represents a partitioning of the acoustic space in the domain of the speech being quantized. The HMM treats any speech utterance as a sequence of random observations generated according to a particular underlying law of the HMM. The underlying law is estimated in the form of the generation of a given utterance from a given set of observations by making a maximum likelihood estimation. The random observations can be in various forms, one of which is quantised LPC vectors.

While enjoying certain advantages, however, VQ has the drawback of reducing recognition accuracy. Recently the authors successfully proposed a method for effectively reducing this degradation called the Segment Classification and Labelling Technique (SCLT) [1]. The SCLT classifies the training data into three classes; voiced, unvoiced or silence. Then it generates a separate codebook for each of these classes before producing a single reference codebook. It is interesting to use these codebooks as the random observations for HMM. Various codebook sizes (16,32,64,128 and 256) have been used for quantizing the LPC vectors and testing the performance of our systems. The performance is also compared with both VQ/DTW and SCLT/DTW alpha-numeric recognition systems which share the same LPC quantizer and testing data.

3 The Segment Classification and Labelling Technique (SCLT)

In the first step of the SCLT, the training speech sequence is required to be classified into three major classes namely, V/U/S. In the second step, the separate data of each class is used to generate the corresponding codebook using the VQ algorithm [1]. In the final stage of this technique a reference codebook of desired size will be formed from the three separate codebooks following a combination (merging) criterion.

A novel approach for detecting the VUS classes was used, in which a spectral characterization of each of these signals was obtained during clustering of the training data, using a K-mean algorithm similar to the VQ algorithm. This method of classification was used for the following reasons: (1) Since it uses the VQ algorithm it does not need to implement a new algorithm for the application under consideration. (2) It does not require the calculation of any other feature other than that used in the analysis of the application. (3) It gives an acceptable discrimination accuracy.

Since the aim was to apply the SCLT to speech recognition, then the table look-up method used here necessitated the need for a criterion for merging these codebooks. Thus such a combination criterion should result in a single reference codebook, so allowing the calculation of the distance of the matrix for its codewords in the usual way of VQ. In such a criterion, codebooks of different codeword counts were combined to form the desired reference codebook. The question that arises now is how to make the most efficient use of this combination. From the actual counts, it was observed experimentally that the number of voiced vectors was approximately twice that of each of unvoiced and silence. This population of voiced vectors satisfied the principle of giving a higher representation for them in the reference codebooks. Therefore, in the following tests a voiced codebook of twice the size of the unvoiced and silence was attempted. Thus to form a reference codebook of size 64, a voiced codebook of size 32 was combined with a codebook of 16 unvoiced codewords and a codebook of 16 codewords of silence.

4 The Hidden Markov Model (HMM)

The idea of representing speech events by HMM's has been used in several speech processing systems. In the HMM we assume that each word model has N-states (where N=5 is used here) and is characterised by a state-transition matrix A and a symbol-probability matrix B. The model parameters (i.e. A and B elements) are estimated from a training sequence of two versions of the vocabulary for each speaker and used to calculate the probability of the observation set given a particular model M. Re-estimation formula due to Baum-Welch was used to iteratively adjust the A's and B's elements until the probability of the observation sequences conditioned on the parameter values stopped increasing significantly, or when some other stopping criterion is met (e.g. the number of iteration exceeded some limit). The recognition procedure used was the Viterbi algorithm.

5 The Database used in the Evaluation

Ten speakers, five male and five female, generated the database. Each speaker was asked to read out as isolated words a list of five versions of the alphabet in random order and ten versions of the randomly ordered English digits (0-9). The VQ and the SCLT algorithms training data were collected from one version of the vocabularies for each speaker. A Hamming window of 256 points at 75% overlap was used. The isolated words are first processed by a 12 poles LPC analysis using the autocorrelation method and Durbin's recursion to form sequences, of LPC vectors. These sequences are then quantized by a VQ and an SCLT. The distance measure used is the minimum prediction residual of Itakura. The outputs of the VQ and SCLT algorithms are then divided into two exclusive sets, one for training the HMM and the other for testing.

6 Comparison of the Performance of VQ and SCLT Recognizers

To evaluate the effectiveness of the SCLT-produced codebooks a series of isolated-word recognition tests were carried out in independent mode for the digit vocabulary and in adaptive mode for the alphabet vocabulary. 50 versions of each word, with an equal number of male and female speakers, were used in creating two reference templates for the independent mode, where a new method of creating reference templates was used [2]. To make the most use of the alphabet data, the letters of each version were assumed to be templates and compared to the letters of the other versions of the vocabulary that were assigned as test words.

Fig. 1 compares the recognition results for the different codebooks, generated by the VQ and SCLT using the method of combination described before, for the digits and the alphabet vocabularies. An examination of these results show that; first, the SCLT reference codebooks gave a lower recognition error rate in comparison with all VQ conventional codebooks for both vocabularies. Second, from Fig. 2 for the Hidden Markov Model recogniser, it is clear that the SCLT codebooks have lower error rates in comparison with the VQ codebooks of the same size. Thus, the SCLT reference codebooks provides a better observation sequence for the HMM than that of VQ codebooks. Generally speaking, the above results suggest that it may be better to quantise other acoustically dissimilar events in addition to V/U/S with a codebook that is formed from separate codebooks.

7 References

1. Mahmoud, W.A. and Bennett, L.A.M., "The Distortion Measure of the Segment Classification and Labelling for different window lengths" Proc. of IEEE International Electrical and Electronics Conference (ELECTRONIC'85), Canada, 7-9 October 1985.

2. Mahmoud W.A. and Bennett, L.A.M., "Creating Reference Patterns Via a Vector Quantization Algorithm", Proceeding of the 2nd International Conf. on Advances in Pattern Recognition and Digital Techniques, India, pp 31-45,6-9 Jan. 1986.

8 Acknowledgement

The authors wish to acknowledge the kind assistance and helpful discussion of Dr. S. Levinson, AT and T Bell Laboratories in USA.

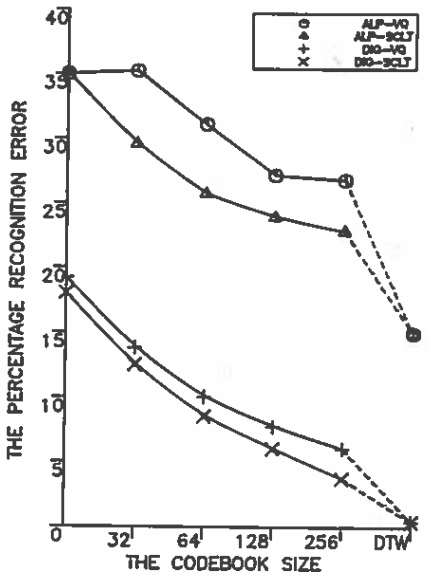

Fig.(1) Average Recognition Error for both vocabularies using different VQ technique as a function of codebook size.

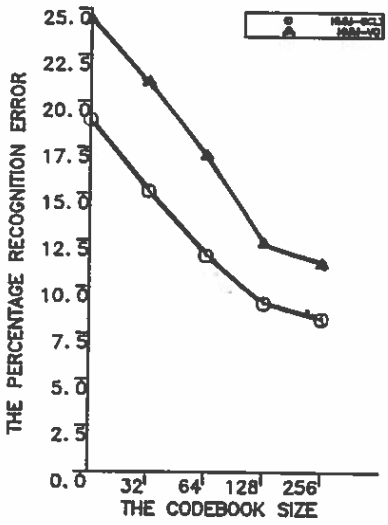

Fig. (2) The average recognition Error for the digit vocabulary using HMM on different VQ techniques.