ON THE USE OF ACOUSTIC SEGMENTATION IN ISOLATED WORD RECOGNITION

Elijah Mwangi
Department of Electrical Engineering
University of Nairobi , Kenya.

ABSTRACT - An isolated word recognition system in which the broad acoustic structure of a word is used to supplement a coventional recognizer is described. The acoustic structure is the voiced, unvoiced , silence pattern of the word. Results obtained by computer simulation show an improvement in the recognition accuracy.

INTRODUCTION

A method for enhancing the recognition accuracy of a 50 word vocabulary of isolated words is presented in this paper. The method is based on the detection of the broad acoustic structure of a word, and then using the information to supplement a conventional recognizer. The acoustic structure obtained using the three voiced, unvoiced, and silence (VUS) classes, can give a strong indication as to the identity of an unknown word within the recognition vocabulary. For example, in a digit recognition system , if the unknown input word is detected to begin with an unvoiced fricative, then it is obvious that the word cannot be a "one", "eight" , or a "nine", and thus the pattern comparison would be limited to the other seven reference candidate words. The information on the acoustic structure can also be exploited in order to discriminate between words which have similar sounding regions. For example , the word set {x,six}, may result in very close distance measure in the LPC-based recognizer, and hence misclassification will occur , if say , an input word "six" has its "x" portion more similar to the reference word "six". Since "x" and "six" have quite different acoustic structures, it would be advantageous to use the acoustic structure as an aid in the recognition process.

Thus , in the proposed recognition system, a reference word pattern is represented both as a sequence of LPC vectors and by its broad acoustic structure. The acoustic structure is determined by first partitioning a speech pattern into a given number of regions using an optimization process. Each region is classified as voiced, unvoiced, or silence.

During the testing session , an input isolated word is expressed as a sequence of LPC vectors and applied to a conventional recognizer to obtain an output V_1. The same input word is also reduced to its broad VUS pattern, and compared with stored reference patterns. The set of words { V_2 }, whose VUS patterns are identical to that of the input word is thus obtained. The outputs of both the conventional recognizer, and the acoustic identifier, are passed over to a second stage. In this second stage only reference patterns in the set { V_1 U V_2 } are used. Each of the M regions of a reference pattern are represented by a

single LPC vector obtained from averaging the autocorrelation vectors in the region. The input word is likewise processed. The unknown input word is identified as the reference word in the set $\{ V_1 \cup V_2 \}$, which gives the minimum distance.

THE ACOUSTIC SEGMENTATION ALGORITHM

An optimum procedure for segmenting a speech utterance into a given number of regions is as follows (Bridle & Sedgewick 1977):

Let a speech utterance be represented by the discrete sequence of multi-dimensional feature vectors, $\{ a_1 , a_2 , \ldots , a_N \}$, that describe the short term spectral properties. It is desired to partition the utterance into M regions, where $M < N$. The speech pattern has (N-1) junctions, numbered $1, 2, \ldots, N-1$, between feature vectors where the boundary of the regions might be placed. Let the fixed boundaries preceeding the vector a_1, and after the vector a_N, be numbered 0 and N respectively. The division of the utterance into M regions now reduces to selecting the M-1 of the interior junctions $i_1 , i_2 , \ldots , i_{M-1}$, and keeping the fixed boundaries $i_o = 0$ and $i_M = N$.

A "Segment evaluation function:, f(i,j), is defined as the error introduced by representing the region of the utterance between junction i and junction j, as a single feature vector and is given by:

$$f(i,j) = \sum_{k=i+1}^{i} d(a_k , a_{ij}) \qquad (1)$$

where $a_{ij} = \sum_{k=i}^{j} a_k \ / (j-i)$,

is the mean vector in the region, and $d(a_k , a_{ij})$ is a spectral distance measure.

A global segmentation criterion, G, which is a function of the sequence of junctions chosen as the new boundaries is defined as the sum of errors introduced in each portion of the utterance, and is given as:

$$G (i_o , i_1 , \ldots , i_m) = \sum_{k=1}^{M} f (i_{k-1} , i_k k) \qquad (2)$$

Equation (2) can be defined recursively as:

$$G(i_o , i_1 , \ldots , i_m) = G(i_o , i_1 , \ldots , i_{m+1}) + f(i_{m-1} , i_m) \qquad (3)$$

The aim of the algorithm is to obtain the sequence ; $\{i_o , i_1 , \ldots , i_m\}$;which minimizes G. Let F(m,n) be the minimum value of G obtained in dividing the first n segments of the utterance into m sections. Then, equation (3) can be expressed as:

$$F (m,n) = \min_i \{ F (m-1,i) + F(i,n)\} \qquad (4)$$

Equation (4) allows the computation of the approximate error for

the best partition of the whole utterance into M sections. During the computation, the values F(1,N), F(2,N) ,..,F(M-1,N) which are the minimum errors if fewer sections are required, are also produced. At every stage in the computation, the junction number i, which minimizes equation (4) is stored in an array P(m,n). After F(M,N) is obtained, the optimal section boundaries can be recovered by starting with $i_M = M$, and then tracing back through the array P(m,n).

SYSTEM DESCRIPTION

The proposed recognition system consists of two parallel sections,i.e a conventional word recognizer and the VUS -based recognizer. The block diagram given in Figure 1 illustrates the various sections in the system.

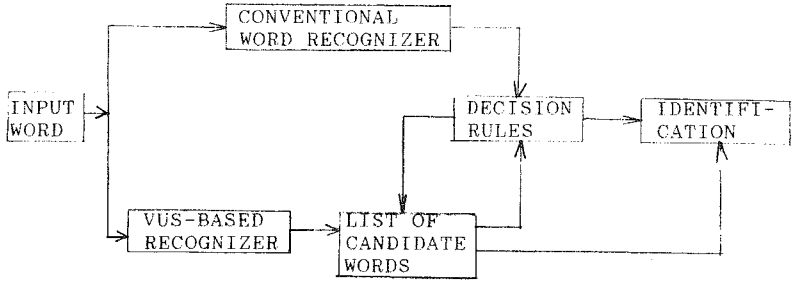

Figure 1: The recognition system

The VUS based recognizer

An input word is segmented into suitable time frames and the LPC coefficients computed. Each frame is also classified as voiced, unvoiced, or silence. This is accomplished by extracting, from the speech segment, the five parameters:
(i)The zero-crossing rate count (ii)The energy (iii) The unit delay autocorrelation coefficient (iv)The normalized prediction error (v)The first LPC coefficient. The five parameters are used in a fuzzy decision process to determine whether the frame is voiced, unvoiced or silence (Mwangi & Xydeas 1985). The next step in the VUS-based recognizer, involves the division of the input word pattern into broad regions. Bridle's algorithm is employed to locate the junctions in the sequence where the region boundaries are to be placed. Since the speech pattern is also expressed as a sequence of LPC vectors, the segment evaluation function in the algorithm employs the gain-normalized Itakura-Saito distance measure (Itakura 1975). Each region is then classified as voiced, unvoiced, or silence, according to the identity of the majority of the segments contained within the region. The result of the above procedure, is that the input word is expressed as a sequence of voiced, unvoiced or silence labels that indicate the broad acoustic structure of the word. This broad VUS pattern of the input word is compared with reference VUS patterns of vocabulary words generated in a similar manner during a training session. All the reference words, whose VUS

patterns have the same structure as that of the input word, are identified as potential candidates.

The conventional word recognizer
In the recognizer, Vector Quantization (VQ) techniques are employed, thereby avoiding the use of dynamic time warping. The recognizer is shown diagramatically in Figure 2 (Mwangi 1987).

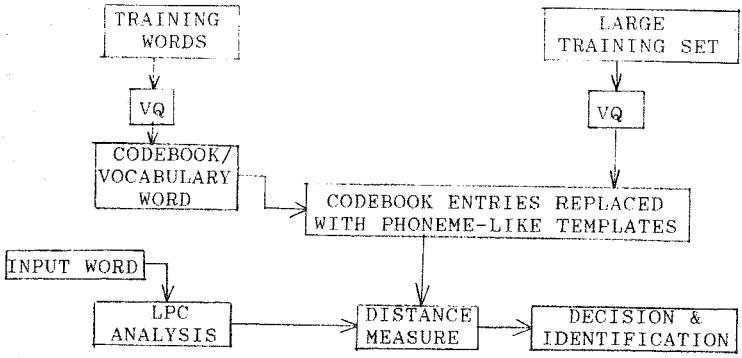

Figure 2: The Conventional word recognizer

A separate reference codebook per vocabulary word is used. Each codebook is based on one word spoken a number of times by different speakers, during a training session. A codebook of M pseudo-phonemes, is also generated by a full search VQ process using a long training sequence of vocabulary words. In addition, the entries of each reference codebook are replaced with the nearest of the M pseudo-phonemes. That is, given an R word vocabulary, the r th codebook C_r of size N_r;

$$C_r = \{ C_1, C_2, \ldots, C_j, \ldots, C_{Nr} \} \quad (5)$$

where C_j is the j th entry.
and given also the phoneme-like templates codebook X;

$$X = \{ X_1, X_2, \ldots, X_k, \ldots, X_K \} \quad (6)$$

The modified codebook $\bar{C}r$ is defined as :

$$\bar{C}r = \{ \bar{C}_1, \bar{C}_2, \ldots \bar{C}_j, \ldots \bar{C}_{Nr} \} \quad (7)$$

where $C_j = X_k$, and the index k minimizes the distortion measure:

$$d(C_j, X_k) \text{ for } k=1,2,\ldots,K \quad (8)$$

An input word, A, which is expressed as a sequence of I, LPC vectors is compared with each of the modified reference codebooks $\bar{C}_r$ to give a distance d_r ,r=1,2,...,R i.e.

$$d_r = (1/I) \sum_{i=1}^{I} \underset{1 \leq j \leq N}{MIN} d(a_i, \bar{C}_j) \quad (9)$$

where $d(a_i, \bar{C}_j)$ is the gain - normalized Itakura- Saito distance measure between the i th LPC vector of A , and the jth entry of $\bar{C}_r$.

The input word is recognized as the word which corresponds to that codebook giving the minimum weighted average distance , D_m:

$$\text{i.e.} \quad D_m = \underset{1 \leq r \leq R}{\text{MIN}} \{ d_r / L_r \} \qquad (10)$$

where L_r is the number of distinct entries used in the r th codebook to obtain d_r. Codebooks of size 16 were used in the proposed recognizer.

Decision rules and identification

During the testing session, an input word , A, is applied to the conventional recognizer to obtain an output , say V_1. The same word is partitioned into regions using Bridle's algorithm, and its VUS sequence identified .The input word, as a VUS pattern, is applied to the VUS - based recognizer, which gives the output set of words $\{V_2\}$, whose VUS patterns are identical to that of the input. The outputs of both the conventional and the VUS-based recognizers, are passed over to the decision stage.

The input word is identified as V_1 if $V_1 \in \{V_2\}$, otherwise a feedback to the VUS recognizer is made. The following rule was used to identify the input word in the set $\{V_1 \cup V_2\}$:

Let the word pattern A be expressed as the discrete sequence of I , LPC vectors:

$$\text{i.e } A = a_1, a_2, \ldots, a_i, \ldots, a_I \qquad (11)$$

The sequence A is partitioned into M regions. Let the mth region, where $1<m<M$, contain the L , LPC vectors:

$$\{ a_i, a_{i+1}, \ldots, a_{i+L-1} \} \qquad (12)$$

The mth region is then represented by the vector a_m , obtained from the autocorrelation coefficients vector $\bar{R}_m$ given by:

$$\bar{R}_m = (1/L) \sum_{j=1}^{L} R_{i+j-1} \qquad (13)$$

where R_i is the autocorrelation coefficient vector which gives the LPC vector a_i.

Thus, the whole utterance A, is represented by M , LPC vectors $\{a_1, a_2, \ldots, a_M\}$. During a training session the reference patterns are also reduced to sequences of M , LPC vectors. A value of M=4 was found suitable for the vocabulary words used in the recognizer.The input pattern is compared with reference patterns in the set $\{V_1 \cup V_2\}$ and identified as the reference word which gives the least distance.

RESULTS

The performance of the proposed recognition system was assessed by computer simulation . The reference patterns were determined using a speech training sequence of 50 isolated words, each word uttered twice by three male and two female speakers. The speech signal was bandlimited to 5kHz was sampled at 10kHz. The 12 bit

per sample, digitized signal was segmented into 25.6 mSec frames, from which LPC coefficients were obtained by a 15th order autocorrelation analysis on the pre-emphasized and Hamming windowed speech every 12.8 mSec.

The speech data base was recorded in a silent chamber and was composed of the following words set:

{ delete, 9, input, f ,o, w, z, k, 3, zero, write, end, 6, j, d, s, load, n, 1, add, m, h, b, set, control, 4, store, l, g, a, v, y, no, e, i, q, 5, read, u, x, 2, p, 8, c, t, yes, r, 7, multiply, output}

The input words, spoken by a subject who did not contribute to the generation of reference patterns, were used for testing the recognition system. The recognition results obtained , as a percentage of correct identification of input words, was 95.3% . Without the VUS based recognizer, the conventional recognizer gave a recognition accuracy of 92.6%.

CONCLUSION

For recognition purposes , it would be desirable to obtain a general acoustic structure of the speech utterance, from its VUS classification of fine segments. This was achieved by using Bridle's algorithm of segmenting an utterance into a specified number of regions. The identification of the VUS pattern of a word was usefully exploited in enhancing the accuracy of an isolated word recognition system.

REFERENCES

Bridle, J. S. and Sedgewick, N.C. (1977) "A method for segmenting acoustic patterns with application to automatic speech recognition". Proc IEEE Inter Conf ASSP pp 656- 659.

Itakura, F. (1975) "Minimum prediction residual principle applied to speech recognition." IEEE Trans ASSP Vol 23 pp67 -72.

Mwangi, E. (1987) "Speaker independent isolated word recognition". PhD Thesis . Loughborough University of Technology, UK.

Mwangi, E. & Xydeas, C.S. (1985) "Voiced - Unvoiced - Silence classification of speech using the fuzzy set theory". Proc IEEE/ Mediterranean Electrotechnical Conf. pp123 -126.