Mansoura Engineering Journal

Volume 17 | Issue 1 Article 9

3-1-2021

Practical Study and Implementation of an Isolated Word Speech Recognition System.

Fayez Soliman
Communication Engineering Department., Faculty of Engineering., El-Mansoura University., Mansoura., Egypt.

Senot Shenodah
Communication Engineering Department., Faculty of Engineering., El-Mansoura University., Mansoura., Egypt.

R. El-Awady
Communication Engineering Department., Faculty of Engineering., El-Mansoura University., Mansoura., Egypt.

Recommended Citation

Soliman, Fayez; Shenodah, Senot; and El-Awady, R. (2021) "Practical Study and Implementation of an Isolated Word Speech Recognition System.," *Mansoura Engineering Journal*: Vol. 17 : Iss. 1 , Article 9.
Available at: https://doi.org/10.21608/bfemu.2021.167256

PRACTICAL STUDY AND IMPLEMENTATION OF AN ISOLATED WORD SPEECH RECOGNITION SYSTEM

FAYEZ W. ZAKI, SENOT D. SHENODAH, and R.M. EL-AWADY

Department of Communication Engineering, Faculty of Engineering
Mansoura University, Egypt.

دراسة عملية وبناء نظام تعرف علي الاصوات ذو الكلمات المنفصلة

خلاصة

هذا البحث يقدم تصميم وبناء نظام بسيط يعمل فى الزمن الحقيقى للتعرف على الكلمات المنفصلة، وفى هذا النظام يتم استخلاص الصفات الاساسية للكلمات المنطوقة من طريق بنك من اربعة مرشحات لتمرير الترددات الصوتية من ٢٠٠ الى ٣٢٠٠ ذبذبة فى الثانية ويتصل خرج هذه المرشحات بدوائر توحيد وتنعيم وذلك للحصول على قياس لطاقة الإشارة الصوتية فى مدى كل مرشح. بعد ذلك يتم توصيل قياسات الطاقة هذه الى حاسب شخصى مزود ببرنامج خاص للتعرف على الكلام وذلك بتحديد بداية ونهاية كل كلمة ثم مقارنة الصفات المدخلة اليه بمكتبة من الصفات المخزونة فى الذاكرة واختيار الكلمة ذات الصفات القريبة من الصفات المدخلة. وقد استخدمت فى هذه الدراسة مجموعتين من كلمات اللغة العربية- المجموعة الاولى هى(واحد-اثنين - ثلاثة-اربعة- خمسة- ستة-سبعة- ثمانية- تسعة- عشرة) والمجموعة الثانية هى (يمين - يسار-امام - خلف - فوق - تحت - سريع - بطئ- ابدأ- قف) واظهرت نتائج المحاكاة على الحاسب الآلى وكذلك النتائج المأخوذة من النموذج العملى أن النظام المقترح يتعرف على هذه الكلمات بنسب نجاح تتراوح بين ٧٦٪ الى ٩٤٪.

ABSTRACT

The art and science of speech recognition have been advanced to the state where it is now possible to communicate reliably with a machine by speaking to it in a disciplined manner using a vocabulary of moderate size. It is the purpose of this paper to present a hardware implementation of a simple real-time speech recognition system for isolated words. The system determines the energy of the spoken word through a bank of four channels band pass filters. The output from each filter is passed through a rectifier followed by a low pass filter to provide an average level which is proportional to the total signal energy within the particular filter band. The energy measurements from each channel is then passed to a personal computer where a recognition algorithm program is then executed. The program detects the starting and ending points of each spoken word, compares the energy features with the reference stored in the library, and chooses the most close pattern that matches the pattern of the spoken word. Results from experiments conducted by simulation and by the hardware prototype using Arabic speech show that the system provides recognition score between 76 % and 94 %.

I-INTRODUCTION

Speech recognition is the process of automatically extracting and determining linguistic information conveyed by a speech wave using computer or electronic circuits. Linguistic information, the most important information in a speech wave, is also called phonetic information.

Speech recognition can be classified into isolated word recognition, in which words uttered in isolation are recognised, and continuous speech recognition, in which continuously uttered sentences are recognised. Continuous speech recognition can be further classified into connected word recognition and conversational speech recognition. The former recognises a relatively small vocabulary aiming at recognising each word correctly. On the other hand, the latter recognises a relatively large vocabulary, but focuses on understanding the meaning of sentences rather than on recognising each word. In conversational speech recognition, also called speech understanding, it is very important to use sophisticated linguistic knowledge.

Speech recognition can also be classified from different points of view into speaker-independent recognition and speaker-dependent recognition systems. The former system can recognise speech uttered by any speaker, whereas, in the latter case, reference templates must be modified every time the speaker changes. Since speaker-independent recognition is much more difficult than speaker-dependent recognition, the former has been used only for isolated word recognition.

Various units of reference templates from phonemes to words have been used. When words are used as reference, the input signal is compared with each of the system's stored templates, i.e., sequences of values corresponding to the spectral pattern of a word, until one is found that matches. Conversely, phoneme based systems analyse the input into a string of sounds that they convert to words through a pronunciation based dictionary.

Speech recognition provides four specific advantages:

1-Speech input is easy to perform because it does not require a specialised skill as does typing or pushbutton operation,

2-Speech can be used to input information three to four times faster than typewriters and 8 to 10 times faster than handwriting,

3-Information can be input even when the user is moving or doing other activities involving the hands, legs, eyes, or ears, and

4-Since a microphone or telephone can be used as an input terminal, inputting information is economical, with remote inputting capable of being accomplished over existing telephone networks.

Automatic speech recognition methods have been investigated for many years aimed principally at realising typewriters and robots capable of recognising speech. The first technical paper to appear on speech recognition was published by Davis et al [1] in 1952. Research on speech recognition has since intensified, and speech recognisers for communication with machines through speech wave have recently been constructed, although remain of limited use. Conversation with machines can be actualised by the combination of a speech recogniser and a speech synthesiser.

In 1958, Dudley and Balashek [2] reported a ten digit recogniser for voice dialing of telephone numbers. In 1960, with the growth of digital computer technology, Denas and Mathews [3] proposed the first isolated-word speech recognition system using a digital computer. The digital computer offered a convenient means of applying a wide variety of digital signal processing techniques and testing recognition algorithms. In 1975, Itakura [4] proposed two speaker-dependent recognition systems for isolated words. The first system had 200 words of Japanese cities names in the system's vocabulary and provided a recognition rate of 97.3 %. The second system had a vocabulary of alphanumeric characters (A to Z and 0 to 9) and provided a recognition rate of 88.6 %. The two systems were based on LPC and loglikelihood distance measure. In 1983, Kowk et al [5] reported a recognition system using a bank of twelve 8-pole band pass filters and M6800 microcomputer. The features used were the amplitude and standard deviation of the filters output. The system provided a recognition rate between 57 % - 100 % for a vocabulary of ten words of Cantonese digits. In the same year 1983, Bui et al [6] reported an integrated speaker-dependent recognition system for isolated words using a bank of seven 4th. order band pass filters. The features used were the energy levels of the signal. Recognition rate of the order of 95 % was reported for a small vocabulary of fifteen words and ten digits. In 1985, Lau and Chan [7] proposed a speaker trained, isolated word recognition system based on the zero-crossing rate and energy level of the speech signal as features of the utterance. The recognition rate of the system was 97.5 % for a small vocabulary of ten words of Cantonese digits. Moreover, Morveit and Brodersen [8], reported a large vocabulary (1000 words) recognition system based on special purpose integrated circuit and a general purpose microprocessor. The recognition algorithm was carried out using a bank of 16 band pass filters and dynamic time warping(DTW) technique. Recognition rates between 82.5 % and 92 % were reported. In 1988, Wei et al [9] proposed three recognition systems for Chinese diphones. Instead of dynamic time warping, linear matching at a few fixed points of word duration was performed. The first system had a vocabulary of 59 Chinese phonetic units and provided a recognition rate of 76.3 %. The second system had a vocabulary of 40 monophones and provided a recognition rate of 95 %. The last system had a vocabulary of 100 diphones and provided a high recognition rate of 99.5 %.

In 1990, Garas et al [10] reported a speaker-dependent recognition system for isolated words using a bank of seven 4th. order band pass filters. The features used were the zero-crossing rate of the signal. Recognition rate of the order of 85 % to 90 % was reported for a small vocabulary of ten Arabic words.

In 1991, Adznan [11] reported a recognition system for Malay digits (0 - 9) using bank of 8 band pass filters. Each filter was constructed using switched capacitor 4th. order chip MF10. Based on the energy levels of the outputs from the filter bank, an average recognition rate of 92 % was reported for a small vocabulary of ten digits.

II-RECOGNITION SYSTEM ANALYSIS

Speech recognition is, in its general form, a conversion from acoustic waveform to a written equivalent of the message information. The nature of the speech recognition problem is heavily dependent upon the constraints placed on speaker, speaking environment, and message context.

Typically, speech recognition problem is treated as a classical pattern recognition problem. This involves comparing the parameters or features representing the incoming utterance with the previously stored prototype reference patterns of each of the words in the vocabulary. A block diagram of a general speech recognition system is shown in Fig. 1. The function of each block is now explained.

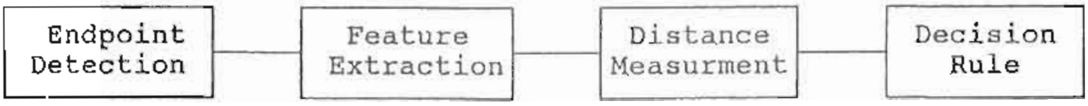


Fig. 1, General Block Diagram for Speech Recognition System.

1-End-Point Detection

The main function of the end-point detector is to specify the region of the speech utterance to be recognised. Most end-point detectors depend on amplitude functions and/or zero crossing rates to perform their tasks. The goal of the technique used in this work implies that only simple measurements can be made on the speech waveform as a basis for the decision. If speed and simplicity were not major issues, far more sophisticated processing could be used to give better and more accurate results. With the above consideration in mind, the end-point location algorithm [12] that was implemented is based on simple energy measurements and uses simple logic in the final decision algorithm.

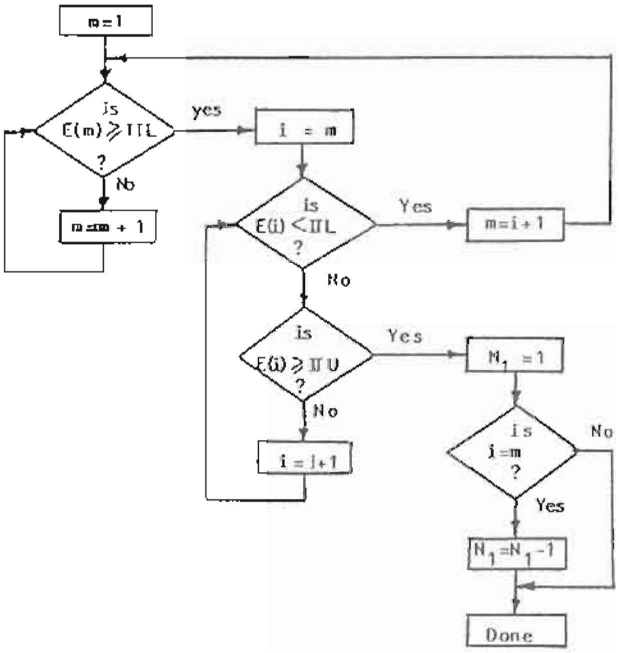
Fig. 2, shows flowcharts for the end-point location algorithm used. The energy function for the entire speech utterance E(n) is computed. The peak energy (IMX) and the silence energy (IMN) are used to set two thresholds ITL and ITU according to the rule

$$I_1 = 0.03(IMX - IMN) + IMN \quad (1)$$

$$I_2 = 4\ IMN \quad (2)$$

$$ITL = \min\ (I_1\ ,\ I_2) \quad (3)$$

$$ITU = 5\ ITL \quad (4)$$



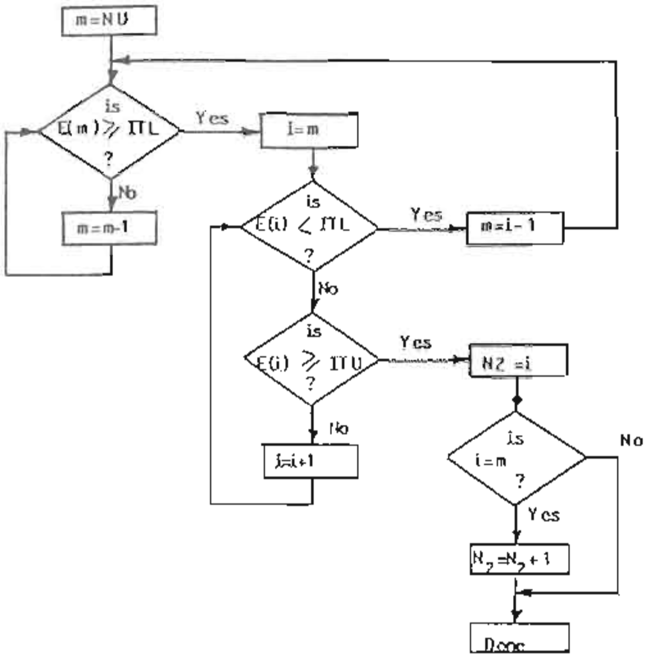


Fig. 2(a), Flowchart for the Beginning Point Initial Estimate.

Fig. 2(b), Flowchart for the Ending Point Initial Estimate.

Eq.(1) shows I_1 to be a level which is 3 % of the peak energy (adjusted for the silence energy), whereas Eq.(2) shows I_2 to be a level set at four times the silence energy. The lower threshold (ITL) is the minimum of these two conservative energy thresholds, and the upper threshold (ITU) is five times the lower threshold.

The algorithm for a first guess at the beginning point location is shown in Fig.2(a). The algorithm begins by searching from the beginning of the interval until the lower threshold is exceeded. This point is preliminarily labeled the beginning of the utterance unless the energy falls below ITL before it rises above ITU. Should this occure, a new beginning point is obtained by finding the first point at which the energy exceeds ITL, and then exceeds ITU before falling below ITL. A similar algorithm shown in Fig. 2(b) is used to define a preliminary estimate of the ending point of the utterance. The beginning and ending points are called N_1 and N_2 respectively.

2-Feature Extraction

The simplest form of representation of a speech signal in a digital computer requires limiting the spectral bandwidth of the signal, sampling it at appropriate rate (typically 8000 to 16000 sample/sec.), and storing each sample with an adequate resolution (8 to 12 bits). For speech recognition, it is desirable to eliminate redundancy in the speech signal to permit efficient representation of the essential aspects in the form of parameters and to simplify data manipulation. The relevant parameters for speech recognition must be consistent across speakers, thus they should yield similar values for the same phonemes uttered by various speakers, while exhibiting reliable variation for different phonemes. The feature extraction is basically a data reduction technique whereby a large number of data points (samples of the speech signal) are transformed into a small set of features.

Methods for feature extraction can be divided into parametric analysis (PA), e.g. analysis by synthesis, linear predictive coding LPC, maximum likelihood, etc., and nonparametric analysis (NPA), e.g. short-time autocorrelation, short-time energy, short-time zero crossing rate, etc. In the PA, a model which fits the objective signal is selected and applied to the signal by adjusting the feature parameters representing the model. On the other hand, NPA methods can generally be applied to various signals, since they do not model the signal. The two most common techniques of NPA are now considered.

a)Short-Time Energy

Most short-time processing techniques produce parameter signals of the form:

$$Q(n) = \sum_{m=-\infty}^{\infty} T\,[s(n)]w(n-m) \qquad (5)$$

where the speech signal s(n) undergoes a transformation T, is weighted by the window function w(n) and is summed to yield a signal Q(n) at the original sampling rate. The signal Q(n) represents some speech properties, corresponding to the transform T, averaged over the window duration. To the extent that w(n) represents a low pass filter, Q(n) is a smoothed version of T[s(n)]. Q(n) in Eq.(5) corresponds to short-time energy if T is a squaring operation, whereas it corresponds to average magnitude if T is an absolute magnitude operation. The energy measure emphasises high amplitudes, since the signal is squared, while the average magnitude measure avoids such emphasis and is easier to calculate. For isolated word recognition, Q(n) can aid in accurate determination of the end-points for a word surrounded by pauses. In particular for voiced segments Q(n) is generally much higher than that for unvoiced segments.

For speech signal, the short-time energy calculation may be carried out as:

$$E(n) = \sum_{m=-\infty}^{\infty} s^2(m)h(n-m) \qquad (6)$$

where h(n) is the impulse response of a low pass filter given by

$$h(n) = \begin{cases} \propto^{n-1} ; & n \geq 1 \\ 0 ; & \text{otherwise} \end{cases} \qquad (7)$$

and $0 < \propto < 1$ for stability of the low pass filter. Applying Eq.(7) into Eq.(6), gives

$$E(n) = \sum_{m=-\infty}^{n-1} s^2(m) \qquad {}^{n-m-1} \tag{8}$$

It can be shown that E(n) in Eq.(8) satisfies the difference equation

$$E(n) = \alpha E(n-1) + s^2(n-1) \tag{9}$$

Eq.(9) is suitable for simulation operation.

b)Short-Time Average Zero Crossing Rate

The zero crossing of a discrete-time signal is said to occure if successive samples have different algebraic signs (i.e. it crosses the time axis). The zero crossing rate can be used to provide adequate information at low computational cost. Thus the average zero crossing rate provides a reasonable way to estimate the characteristics of the speech signal.

The average zero crossing rate may be expressed as

$$Z(n) = \sum_{m=-\infty}^{\infty} \left| \mathrm{sgn}(s(m) - \mathrm{sgn}(s(m-1)) \right| w(n-m) \tag{10}$$

where

$$\mathrm{sgn}(s(n)) = \begin{cases} +1 & ; \quad s(n) \geq 0 \\ -1 & ; \quad s(n) < 0 \end{cases}$$

$$w(n) = \begin{cases} 1/2\ N & ; \quad 0 \leq n \leq N-1 \\ 0 & ; \quad \text{otherwise} \end{cases}$$

and N is the window length used.

Since high frequencies imply high zero crossing rates, and low frequencies imply low zero crossing rates, there is a strong correlation between zero crossing rate and energy distribution with frequency for unvoiced speech. This is due to the fact that most of the energy is found at higher frequencies for unvoiced speech. A reasonable generalisation is that if the zero crossing rate is high, the speech signal is unvoiced, while if the zero crossing rate is low, the speech signal is voiced.

3-Time Alignment and Distance Measurements

The next step in the recognition system of Fig. 1, is to determine similarities between test and reference patterns [13,14]. Because speaking rates vary greatly, pattern matching involves both time alignment and distance measure. In practice, these two operations are performed simultaneously.

The function of time alignment between test pattern T(t) and reference pattern R(t) is shown in Fig. 3. The goal is to find an alignment function w(t) which maps R(t) into the corresponding parts of T(t). The criterion for correspondence is that some measure of distance between the functions D(T,R) be minimised by the mapping function w(t). Thus the problem is to seek w(t) such that

$$D(T,R) = \min_{w(t)} \int_{t_0}^{t_1} d(t,w(t))G(t,w(t),w'(t)) \quad dt \tag{11}$$

where w(t) is the set of all monotonically increasing, continuous differentiable functions, w'(t) is the derivative of w(t), d(t,w(t)) is the pointwise distance from R to T, and G is a weighting function. The problem of

Eq.(11) is not, in general, solvable so this equation can be discretised by letting

$$T = \{ T(1), T(2), \ldots\ldots T(NT) \} \tag{12}$$

and

$$R = \{ R(1), R(2), \ldots\ldots R(NR) \} \tag{13}$$

The optimum time alignment path is a curve relating the m time axis of the reference pattern to the n time axis of the test pattern, of the form

$$m = w(n) \tag{14}$$

where the constrained beginning and ending points of Fig. 3 can be expressed as constraints on w(n) as

$$w(1) = 1 \tag{15}$$

$$w(NT) = NR \tag{16}$$

Several techniques have been proposed for determining the alignment path w, including linear time alignment, time event matching, correlation maximisation, and dynamic time warping [13]. In this study, we adopted the linear time alignment of the form

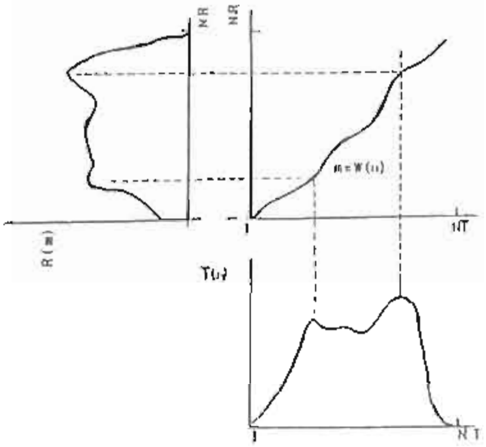


Fig. 3, Simple Time Alignment

$$\begin{aligned} m &= w(n) \\ &= (n-1)\,\frac{(NR-1)}{(NT-1)} + 1 \end{aligned} \tag{17}$$

After the time alignment is applied to normalise the reference and test patterns, a frame by frame distance measure is carried out to make a decision about the spoken word. The most common techniques for distance measure are Euclidean distance, covariance weighting, spectral distance, aned LPC loglikelihood measure [14]. The Euclidean distance measure used in this work may be expressed as

$$\begin{aligned} D(T,R) &= \| T - R \| \\ &= \sum_{i=0}^{p} (T_i - R_i)^2 \end{aligned} \tag{18}$$

where T_i and R_i are the ith components of the vectors of the test and reference patterns respectively.

4-Decision Rule for Recognition

The last step in the speech recognition system of Fig. 1, is the decision rule which chooses the most close reference word that match the unknown test word. Although a variety of rules are applicable, only two decision rules have been widely used in most applications. These are the nearest neighbor (NN) and the K-nearest neighbor (KNN) rules.

The NN rule operates as follows: assume the reference pattern vector R_i, i=1,2,....V, where V is the vocabulary size, and for each pattern, the average distance score obtained by the distance measure is D_i, then the NN rule is simply

$$i^* = \underset{i}{\operatorname{argmin}} \ \{D_i\} \tag{19}$$

i.e., choose the pattern R_i^* with smallest average distance as the recognised pattern.

The KNN rule is applied when each reference word is represented by two or more reference patterns. Thus if there are P reference patterns for each of V reference words, and we denote the jth occurrence of the ith pattern as $R_{i,j}$, $1 < i < V$, $1 < i < P$, then if we denote the distance measure for the jth occurrence of the ith pattern as $D_{i,j}$, and if we reorder the P distance of the ith word so that

$$D_{i,1} \leqslant D_{i,2} \leqslant \ldots\ldots \leqslant D_{i,P}$$

then for the KNN rule we compute the average distance (radius) as

$$r_i = 1/K \sum_{j=1}^{K} D_{i,j} \tag{20}$$

and we choose the index of the recognised poattern as

$$i^* = \underset{i}{\operatorname{argmin}} \ \{ \ r_i \ \} \tag{21}$$

III-SYSTEM SIMULATION

To design and construct the recognition system introduced in this work, computer simulation has been carried out first. The results obtained from this simulation provided the base plate for the hardware implementation of the next section.

The data used in simulation experiments were prepared as follows: analog speech signal from a dynamic microphone is band pass filtered from 200 to 3200Hz, sampled at 16 KHz, converted into digital form using 12-bit high speed A/D converter, and stored on floppy disks. Two sets of Arabic words were used. The first set consists of the Arabic numbers (واحد-اثنين-ثلاثة-أربعة-خمسة-ستة-سبعة- ثمانية-تسعة-عشرة)and the second set consists of the Arabic words(امام ، خلف ، يمين ، يسار فوق ، تحت ، سريع ، بطيء ، إبدأ ، قف) .The duration of each word is limited to one second. For a given speaker, three versions of each word were stored on the floppy diskettes, so that a vocabulary size of 60 words was available. The simulation procedure is shown in Fig. 4.

Here, for each utterence of speech, an end-point detection program is called to specify the starting and ending points of the utterance. In the next step, the speech data is applied to a bank of band pass 4th order Butterworth digital filters, rectifiers, and smoothing low pass filters for feature extraction. The last phase of the simulation procedure is carried out through the recognition algorithm (Euclidean distance and decision rule) program.

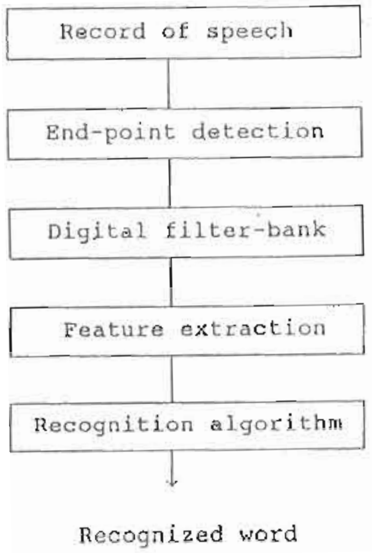


Fig. 4, Sequence of Simulation Procedure

A series of simulation experiments was performed using different feature sets (e.g. energy, absolute magnitude, zero crossing, and combination of both), variety of filter bank order, and filter spacing. The results are shown in Table 1. From which, it can be concluded that the recognition system based on short-time energy or short-time absolute magnitude provides the highest recognition rate as compared to systems based on either zero crossing rates or hybrid combination of both energy and zero crossing rate. Although, the 15th order filter bank provides the highest recognition rate, it requires very high computation complexity. The 4th order filter bank with an octave filter spacing requires much less computation complexity and provides a comparable recognition rate to the 15th order filter bank.

Table 1, Recognition Systems Performance as a Function of Filter Bank Order, Filters Spacing, and Feature Sets.

Filter Bank		Recognition Rate %		
		Feature Sets		
No. of Channels	Filter Spacing	Energy & Absolute Magnitude	Zero Crossing	Energy & Zero Crossing
3	Uniform	71 %	60 %	53 %
4	Uniform	82 %	77 %	67 %
5	Uniform	88 %	60 %	67 %
10	Uniform	88 %	60 %	71 %
15	Uniform	94 %	50 %	67 %
30	Uniform	88 %	45 %	67 %
4	Octave	88 %	60 %	88 %
7	1/3 Octave	88 %	60 %	76 %
12	Nonuniform	88 %	60 %	71 %

Taking advantage of the above results, a speech recognition system based on short-time absolute magnitude and 4 channel octave spacing filter bank is devised for implementation. The cut off frequencies for this filter bank are shown in Table 2.

Table 2, Filter Bank Cut Off Frequencies

Channel No.	Cut Off Frequencies	
	f_1 (Hz)	f_2 (Hz)
1	200	400
2	400	800
3	800	1600
4	1600	3200

IV-HARDWARE IMPLEMENTATION

A block diagram for the hardware implementation is shown in Fig. 5. As shown the system designed consists of three parts, analog signal processing card, A/D card, and personal computer to run the recognition algorithm program.

1-Analog Signal Processing Card

Fig. 6 shows a circuit diagram for the analog signal processing card. In this figure, input speech signal from a microphone is amplified to the required level through a two-stage preamplifier built around the dual Op Amp IC LF 353. This amplifier also provides isolation buffering between the microphone and the filter bank. The output from the preamplifier is then applied to the filter bank and the rectifier. The filter bank (four channels each consists of a 4th order multiple feedback band pass filter [15]) is one of the NPA techniques mentioned previously. The band pass filters in the filter bank are arranged so that the center frequencies are distributed with equal intervals on the logarithmic frequency scale (see Table 2), so that the -3 dB points of the adjacent filters coincide. The output of each band pass filter is rectified, smoothed by an RC low pass filter, and then applied to the A/D converter. The RC time constant is chosen to be 0.01/π sec. to satisfy a 50 Hz cut off frequency for the L.P.F.

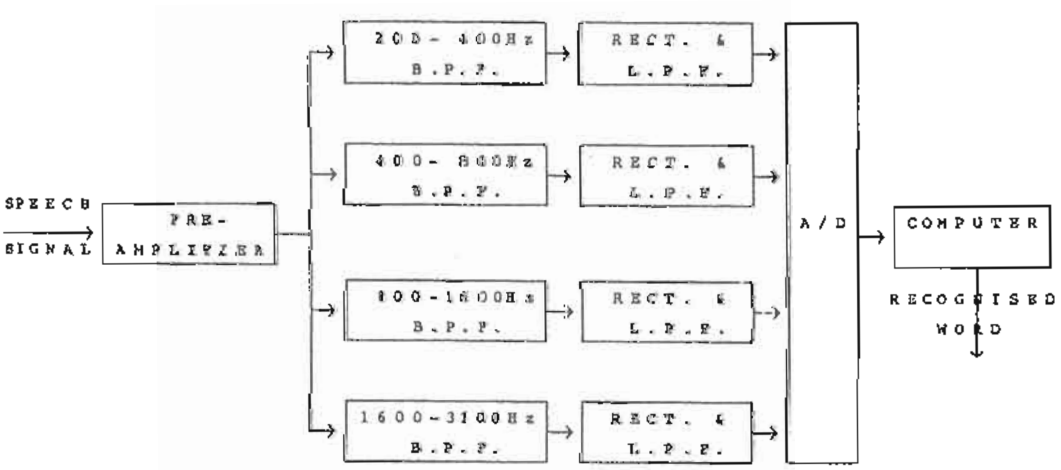


Fig. 5, Block Diagram for the Speech Recognition System.

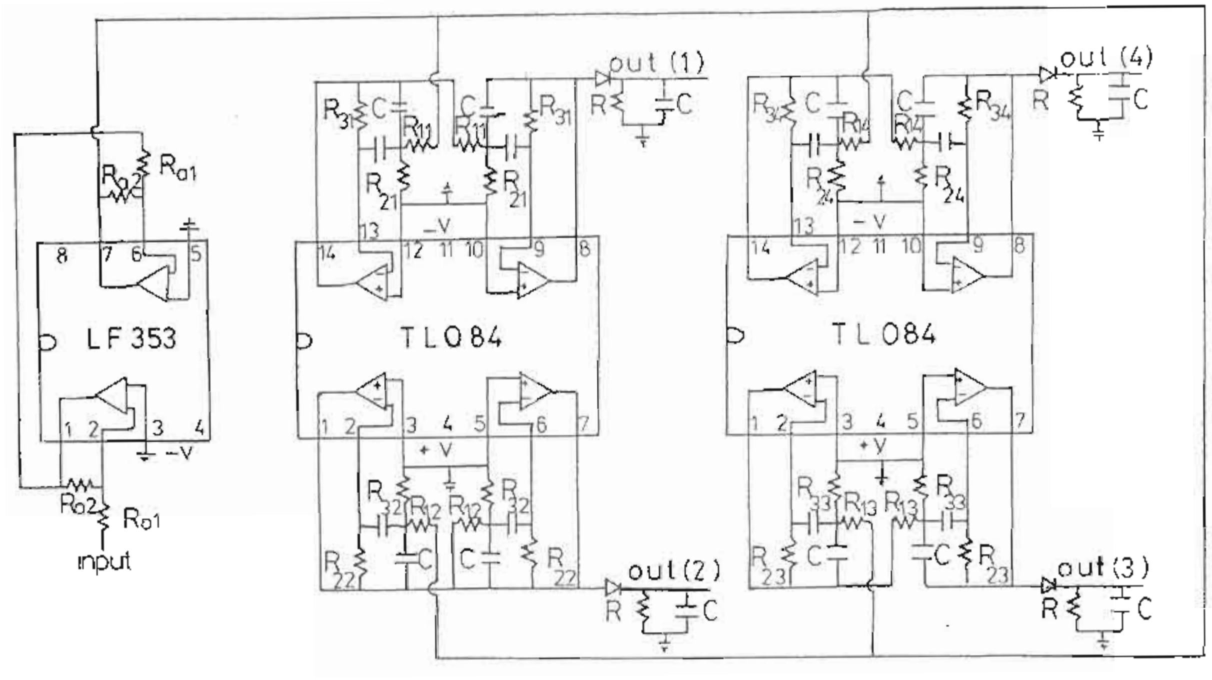


Fig. 6, Circuit Diagram of the Analog Signal Processing Card.

2-A/D Interface Card

The second part of the recognition system is the A/D interface card. This card is used to convert the energy measure information to a suitable form for processing using a personal computer. The A/D card used is the uCDAS-8PGA [16] from Metra Byte. It is a high speed 12-bit successive approximation A/D converter with conversion time of 25 µsec.(35 µsec. typical). It has 8 analog channels, only 4 of them are used. The full scale input voltage ranges from -5 V. to +4.99 V., which is suitable to the output from the filter bank.

3-Microcomputer and the Recognition Algorithm

The final part of the system implementation is the design of a computer program to perform the function of the recognition algorithm. All programs used are written in Quick Basic [17]. The first program reads the data from the A/D at a rate of 1024 sample/sec. for each channel and store it in the computer memory as a two dimension array. The second program performs the end-point detection discussed previously. Due to variation in speaking time of each version of the same utterance, resegmentation of the data is necessary. In this process, a fixed number of segments is assigned to each word (typical values used were 128 seg./word).

Finally, the Euclidean distance measure is applied to compare the features of a test word with a set of reference features stored in the memory. The program concerned estimates the minimum distance between the test word and the reference words, then decides which word was spoken. A flowchart for the recognition algorithm software is shown in Fig. 7

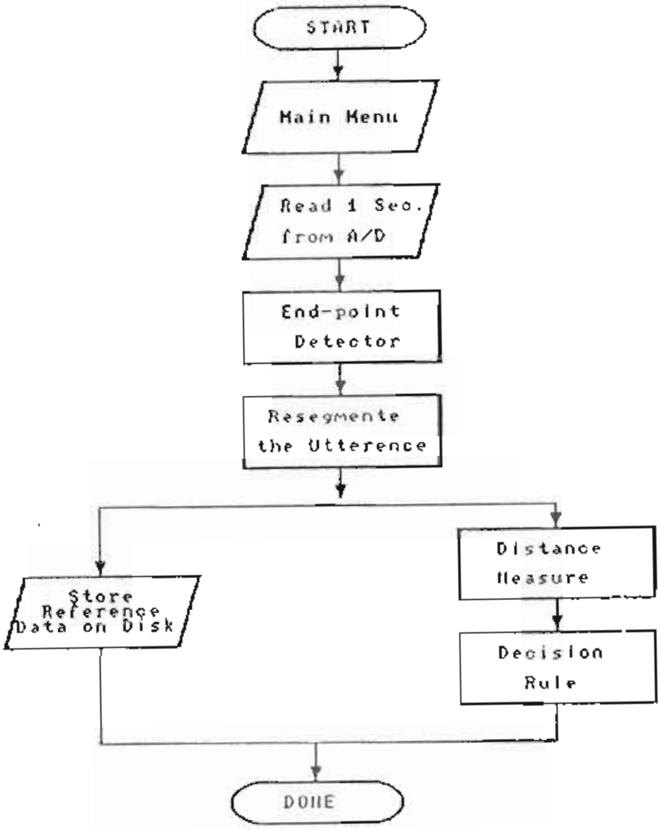


Fig. 7, Flowchart for the Recognition Algorithm Software

IV-EXPERIMENTAL RESULTS

Two sets of vocabulary were stored in the memory. These are the Arabic numbers (واحد- اثنين - ثلاثة- اربعة- خمسة- ستة- سبعة- ثمانية- تسعة- عشرة) and the Arabic words (أمام ، خلف ، يمين ، يسار ، فوق ، تحت ، سريع ، بطيئ ، إبدأ ، قف) For each word in the first vocabulary set, thirty test were conducted, whereas twenty five tests were conducted for each word in the second set. The results are shown in Table 3.

Table 3, Performance of the Hardware Prototype Recognition System

Vocabulary Set I			Vocabulary Set II		
Word	No. of Tests	Aver. Recognition Rate	Word	No. of Tests	Aver. Recognition Rate
واحد	30	86.7 %	أمام	25	84.0 %
إثنين	30	93.3 %	خلف	25	80.0 %
ثلاثة	30	86.7 %	يمين	25	76.0 %
أربعة	30	93.3 %	يسار	25	84.0 %
خمسة	30	90.0 %	فوق	25	80.0 %
ستة	30	86.7 %	تحت	25	84.0 %
سبعة	30	83.3 %	سريع	25	88.0 %
ثمانية	30	83.3 %	بطيئ	25	80.0 %
تسعة	30	90.0 %	إبدأ	25	84.0 %
عشرة	30	86.7 %	قف	25	76.0 %

It can be seen from table 3, that the recognition rate for the first vocabulary set varies between 83.3 % to 93.3 % with an average of 88 %. On the other hand, the recognition rate for the second vocabulary set varies between 76 % to 88 % with an average of 81.6%. These results are close to those obtained by computer simulation (table 1). The marginal difference between the results obtained from the hardware prototype · and the computer simulation may be due to variations in recording environment, time alignment, and tolerance in hardware components

V-CONCLUSIONS

An isolated word speech recognition system based on different feature sets have been considered in details in this paper. The system based on measurements of energy has been demonstrated to be superior to those systems based on zero crossing rate and hybrid combination of energy measure and zero crossing rate. Moreover, the features are extracted using a bank of band pass filters, rectifiers, and low pass filters. Four channels, 4th order band pass filter bank with an octave spacing is found to provide the highest recognition rate and lower cost as compared to those systems reported in [10] and [11].

Computer simulation as well as design and implementation of the recognition system were developed. The implementation consisted of an analog signal processing card for features extraction, a 4 channel A/D converter to transfer the features to the microcomputer, and a computer program to perform the end-point detection, distance measure, and decision rule. The recognition rate obtained from the hardware prototype varied between 76 % to 93.3 % when two sets of small vocabulary size Arabic words were used.

Since it is a speaker dependent recognition system, it may be suitable for applications in machine control by voice, aids for handicapped, and in consumer products like watches, clocks or toys.

REFERENCES

1-Davis, K.H., Biddulph, R., and Balashek, S. "Automatic Recognition of Spoken Digits," Journal of the Acoustic Society of America, Vol. 24, No. 6, pp 637-642, Nov. 1952.

2-Dudley, H., and Blashek, S., "Automatic Recognition of Phonatic Patterns in Speech," Journal of the Acoustic Society of America, Vol. 30, No. 8, pp721-732, Aug. 1958.

3-Denes, P., and Mathews, M.V., "Spoken Digit Recognition Using Time Frequency Pattern Matching," Journal of the Acoustic Society of America, Vol. 32, No. 11, pp1450-1455, Nov. 1960.

4-Itakura, F., "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. ASSP-23, No. 1, pp67-71, Feb. 1975.

5-Kwok, H.L.,Tal,L.C., and Fung,Y.,"Machine Recognition of the Cantonese Digits Using Band Pass Filters," IEEE Trans. ASSP-31, No. 1, pp220-222, Feb. 1983.

6-Bui, N.C., Monbaron, J.J., and Michel, J.G., "An Integrated Voice Recognition System," IEEE Trans. ASSP-31, No. 1, pp323-328, Feb. 1983.

7-Lau, Y.K., and Chan, C., "Speech Recognition Based on Zero Crossing Rate and Energy," IEEE Trans. ASSP-33, No. 1, pp320-323, Feb. 1985.

8-Murveit, H., and Brodersen, R.W., "An Integrated Circuit Based Speech Recognition System," IEEE Trans. ASSP-34, No. 6, pp1465-1472, Dec. 1986.

9-Wei, L.Y., Lee, J.P., and Lee, C.C., "Recognition of Chinese Diphones," IEEE Trans. ASSP-36, No. 10, pp1684-1687, Oct. 1988.

10-Garas, E., Hamed, A.M., Elghonemy, M.R., and Somaie, A.A., "Limited Vocabulary Isolated Word Recognition System," Proc. of the 7th National Radio Science Conf., M.T.C., Cairo, Feb. 1990.

11-Adznan, B.J., "Recognition System of Malay Digits Using Bank of Band Pass Filters," Proc. AEC 91, Al-Azhar Eng. 2nd Int. Conf.,Faculty of Eng., Al-Azhar Univ., Cairo, pp357-369, Dec. 21-24, 1991.

12-Rabiner, L.R., and Sambur, M.R., "An Algorithm For Determining The End Points of Isolated Utterances," Bell Syst. Tech. Journal AT&T, Vol. 54, No. 2, pp297-317, Feb. 1975.

13-Rabiner, L.R., Rosenberg, A.E., and Levinson, S.E., "Consideration in Dynamic Time Warping For Discrete Word Recognition," IEEE Trans. ASSP-26, pp275-282, Dec. 1978.

14-Rabiner, L.R., and Levinson, S.E., "Isolated and Connected Word Recognition: Theory and Selected Applications," IEEE Trans. COM-29, pp621-659, May 1981.

15-Berlin, H.M., "Design of Active Filters with Experiments," Haward W. Sams & Co. Inc., 1977.

16-"µCDAS-8PGA User's Manual," MetraByte, Part #24863, Jan. 1988.

17-Shenodah, S.D., "A Study of Speech Recognition and its Applications to Arabic Speech," M.Sc. Thesis, Dept. of Communication Eng., Faculty of Eng., Mansoura University, Egypt, 1991.