Military Technical College
Kobry Elkobbah,
Cairo, Egypt


2nd International Conference on
Electrical Engineering
ICEENG 99

# Performance Improvements of Isolated Word Recognizers Based on Hidden Markov Modeling

Mahmoud E. Gadallah
Assoc. Prof., Military Technical College

## ABSTRACT:

This paper introduces proposed solutions for increasing the efficiency of the isolated word recognizers (IWR) whose vocabularies comprise words with considerable difference in length. Most of the languages consist of words that have appreciable difference in the numbers of phonemes. In these cases, the recognition accuracy of the IWR systems that are based on Hidden Markov Modeling (HMM) is degraded because they usually use a fixed number of states for all the vocabulary words. These proposed approaches are originally initiated to overcome this problem for such types of vocabularies.

The proposed solutions, introduced in this work, have been developed in two stages: (i) In the first stage, the HMM is allowed to construct the words models with variable number of states (VNS-HMM) according to the length of the words. The application of this idea has shown better recognition accuracy over HMM with fixed number of states (FNS-HMM). (ii) Considerable improvement has been achieved in the second stage in which the reference words are categorized according to their lengths and divided into subvocabularies.
Due to the applications of these approaches, two other improvements have been obtained, namely, they are: (i) the ability to increase the vocabulary size; and (ii) decreasing the recognition time.

## KEY WORDS:

Hidden Markov Model (HMM), Isolated Word Recognition (IWR), Number of states of the HMM, Training phase, Recognition phase, Time compression. HMM with Fixed number of states (FNS-HMM), HMM with variable number of states (VNS-HMM).

## I. INTRODUCTION:

IWR systems still the most reliable approach among the different speech recognition techniques such as the connected word recognition (CWR) or the continuous speech recognition (CSR). According to the types of the reference patterns of the IWR systems, they can be classified into : (i) IWR with subword references and (ii) IWR with word references. Although the first type can be designed for larger vocabularies, its performance is lower than the second type because of the problems associated with the segmentation of speech into subwords. The major problem of the IWR with word references is the limitation of the vocabulary size.

Another problem which has been detected during this work, especially with those IWR systems based on HMM, is the degradation of the recognition accuracy when the vocabulary comprises

words with considerable differences in length. This problem has not been discussed before because most of the HMM-based IWR systems used to use fixed number of states for the HMM of all the reference words. This may work well when the vocabulary words are similar from the point of view of the number of phonemes [1]. The dependence of the recognition accuracy on the number of states of the reference models was studied only for the case of vocabularies that comprise words of similar lengths (e.g., the English digits). This study has shown that there is a certain number of states that gives the best recognition accuracy for the vocabulary under consideration only (local solution) [2].

In the normal case, when the vocabulary comprises words with considerable difference in the length, an appreciable degradation of the recognition accuracy has been noticed when using FNS-HMM. This phenomenon can be interpreted as, redundant information are used in constructing the models of short words while necessary information are highly compressed in the case of longer words. This problem had been solved using the Dynamic Time Warping (DTW) but this technique suffers from the huge amount of computations during the recognition phase and this limits its use to small vocabularies.

In this work, solutions to the two problems mentioned above are introduced by proposing a HMM-based IWR system that has been developed through two stages. In the first stage, a HMM with different number of states are used to construct the models of the reference words while keeping all of them in one vocabulary. In the second stage, the reference words are divided into subvocabularies, each one comprises those words of nearly similar phonetic construction (i.e., words whose number of phonemes is in a prescribed range). The solution introduced in the second stage allows increasing the vocabulary size of the system because only one of the subvocabularies will be used to recognize an input unknown word. In addition to that, this approach results in better recognition accuracy due to the separation of the vocabulary words according to their lengths (phonetic construction). Also, as a result of this approach, the response time of the recognizer decreases.

In the second stage, the vocabulary separation has been performed via three approaches, which are: (i) Direct classification depending on the word length; (ii) Direct classification depending on the word length with overlapping the subvocabularies. and (iii) Classification after time compression for the words to remove the unnecessary elongations.

This paper is organized as follows: In the next section, the IWR system based on HMM is introduced with the mathematics of the HMM. In section III, the idea of the VNS-HMM is given with its application to the IWR system introduced in section II. The division of the vocabulary into subvocabularies, is introduced in section IV. Section V presents the results obtained throughout this work. Finally, conclusions that can be deduced from this work are given in section VI.

## II. A HMM-BASED IWR SYSTEM WITH FIXED NUMBER OF STATES:

Figure 1 shows the basic block diagram of an IWR system based on template matching technique using HMM. The following sub-sections introduce a brief description for the functions of the main components of this system.

## 2.1 Signal Preprocessing

The speech signal is sampled and digitized with 10 kHz sampling rate and 8 bits resolution. A technique for End-Point Detection (EPD) [3] is used to remove the background noise before and after speech intervals. This technique is based on the principle of Fuzzy logic to discriminate speech signals from noise using energy and zero-crossing rate. The speech signal is framed using Hamming window of length 20 ms. and successive frames are overlapped by 0.5 frame length. A first order high pass filter is used as a pre-emphasis filter to limit noise and to enhance the higher frequency components.

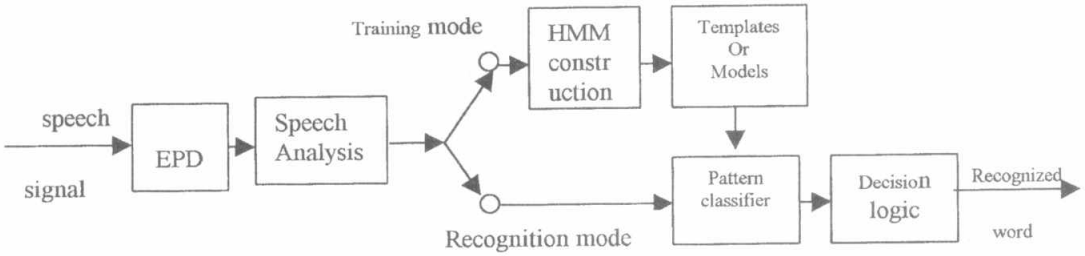



**Figure 1. The Block diagram of an IWR system**

## 2.2 Feature Extraction of Speech

There are many signal processing techniques for feature extraction such as filter bank; linear predictive coding (LPC) and cepstrum analysis. Recently, in the work reported in [4], many processing techniques are compared from the point of view of the recognition accuracy in a noisy environment, the perceptual LPC (PLP) [1],[5] and [6] has shown the best results. Thus, in this work, the Perceptual Linear Prediction analysis of Speech is used for feature extraction. The PLP technique incorporates three basic concepts about the human hearing system. These three basic concepts are the critical band resolution curve, the equal-loudness curve, and the intensity-loudness power-law relationship. These three physical components of the hearing process, combined with an autocorrelation model of the speech signal help to create a better speech coding technique with immunity to noise.

## 2.3 Speech Recognition System:

As shown in Figure 1, the IWR system has two phases: the training phase and the recognition phase. These two phases are described below.

### *(i) The Training Phase:*

The training algorithm can be summarized in the following steps [6]:

1. Read the observations of the words, which are used for building the model. These observations are a set of indices resulting from vector quantization unit.
2. Considering the number of states N=6, initialize the model using the state transition probability matrix A as:

$$A = \begin{bmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

If M is the number of distinct observation symbols per state, the observation symbol probability matrix B is:

$$B = \begin{bmatrix} 1/M & 1/M & \dots & 1/M \\ 1/M & 1/M & \dots & 1/M \\ 1/M & 1/M & \dots & 1/M \\ 1/M & 1/M & \dots & 1/M \\ 1/M & 1/M & \dots & 1/M \\ 1/M & 1/M & \dots & 1/M \end{bmatrix} \quad (2)$$

And the initial model distribution as:

$$\pi = [1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] \quad (3)$$

3. Calculate the probability $P(O/\lambda)$ using the forward and backward algorithms where O is the observation sequence and $\lambda i$ represents the HMM model parameters $(A,B,\pi)$.
4. Using Baum-Welch algorithm to find new estimate values of Aest, Best, $\pi$est.
5. Calculate the likelihood of the model, which is given by:

$$P(O/\lambda_{est}) = \prod_{x=1}^{L} P_x$$

$$P_x = P(O^x/\lambda_{est}) \quad (4)$$

where L is the number of the repetitions of each word that will are used in the training.

6. Calculate the distance, dis, which is given by

$$dis = \frac{P_i - P_{i-1}}{P_{i-1}} \quad (5)$$

where $P_i$ is the present likelihood and $P_{i-1}$ is the previous likelihood. This value is compared with a threshold value $\varepsilon$. If it is larger, the procedure 3-6 are repeated until dis is less than $\varepsilon$ which is chosen to be 0.0005.

### *(ii) The Recognition Phase:*

The recognition algorithm is as follows [1]:

1. Read the observation sequence, $O=\{o_1 o_2 \dots o_T\}$, of the unknown word, T is number of frames of the word.
2. Start with the first reference model $\lambda_1 = (A_1,B_1,\pi_1)$; compute the probability of this word being generated by this model, $P(O/\lambda_1)$, using alternative Viterbi algorithm.
3. Repeat step 2 for all the stored models. The unknown word is assumed to be generated from the model that gives the highest probability.

## III HIDDEN MARKOV MODELING WITH VARIABLE NUMBER OF STATES:

The main difference between the VNS-HMM and the FNS-HMM is in the training phase. In the case of VNS-HMM, the number of states is not determined in the beginning (as in the case

of FNS-HMM), but the model is constructed with number of states (N) which varies between minimum and maximum values (i.e., $N_{min} < N < N_{max}$). Then the optimum number of states of the model is selected as follows:
Calculate the likelihood of the model for all the possible numbers of states as:

$$P_s = P(O \setminus \lambda_S)$$

Where s is the number of states of the model, $N_{min} < s < N_{max}$.
Then select the best number of states is chosen as:

$$s_{sel} = \arg\{(P_s)_{max}\}.$$

This idea is simply implemented by allowing the word-modeling algorithm, explained above, to construct the word model with a number of states that depends on the word length.

## IV. VOCABULARY SEPARATION:

The idea of vocabulary separation has been thought in order to achieve three objectives, namely, they are:

- Obtaining sub-vocabularies that comprise words of similar lengths.
- Consequently, decreasing the number of computations required for recognizing an unknown word.
- The ability to increase the overall vocabulary size.

The number of computations required for an unknown word recognition is given as [1]:

$$C = V.N^2.T \qquad (6)$$

Where
V ... is the vocabulary size;
N ... is the number of states of the reference models;
T ... is the number of observations of the unknown word (the word length).

It is clear from equation (6) that, by vocabulary separation, V will decrease because the number of the words of the total vocabulary is distributed to the subvocabularies. Each subvocabulary contains $V_s(i)$ words (i is the number of the subvocabulary), where $V_s(i) < V$. On recognizing a word, only one of the subvocabularies is considered. Thus, the number of computations will decrease.

The following subsections introduce the three approaches applied for Vocabulary separation.

### 4.1 Direct vocabulary separation:

In this approach, the vocabulary is divided directly according to the lengths of its words into three subvocabularies: long words; medium words; and short words as shown in figure 2.The models and the codebook of each of these subvocabularies are stored separately. For an unknown input word, the subvocabulary with which it will be compared is selected after the EPD directly according to its length.

The problem of this approach is the ambiguity between some of the words from the length point of view. This ambiguity occurs due to non-standard way of speaking. In another words, the speaker may elongate a short word or shorten a long word and so on. In such case, the input unknown word will be matched with a wrongly selected subvocabulary. To overcome

this problem, the system shown in figure 3 has been applied. In this system, an overlap between the subvocablaries has been implemented. The overlap between subvocabularies means that, there are reference models of some words in two subvocabularies (e.g., short and medium / medium and long). Although this system requires more storage as well as more recognition time, it shows better recognition accuracy.

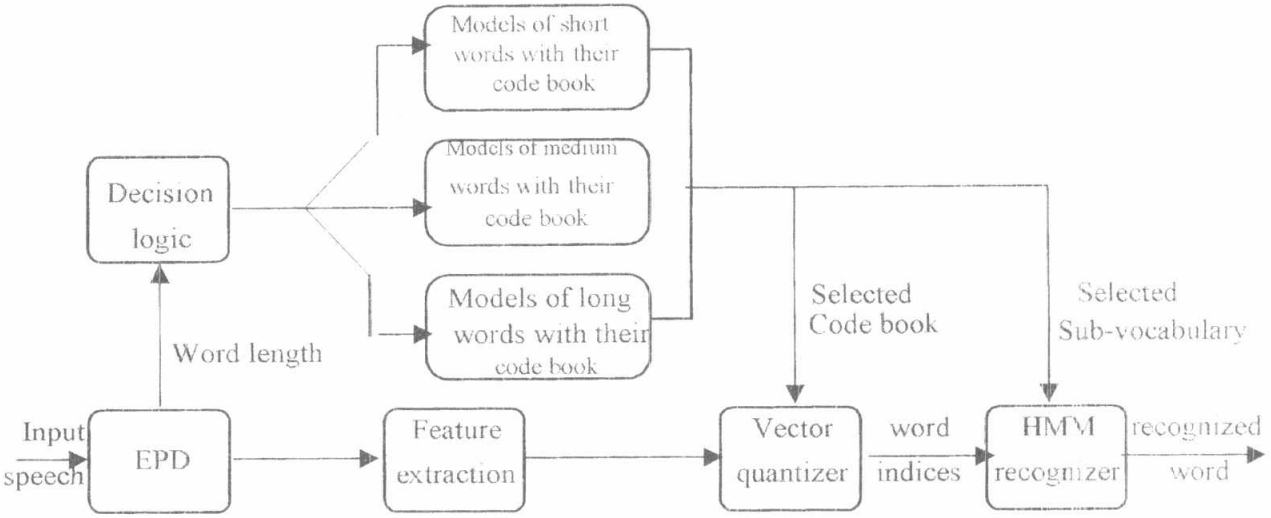



Figure 2. Separation of words according to their lengths

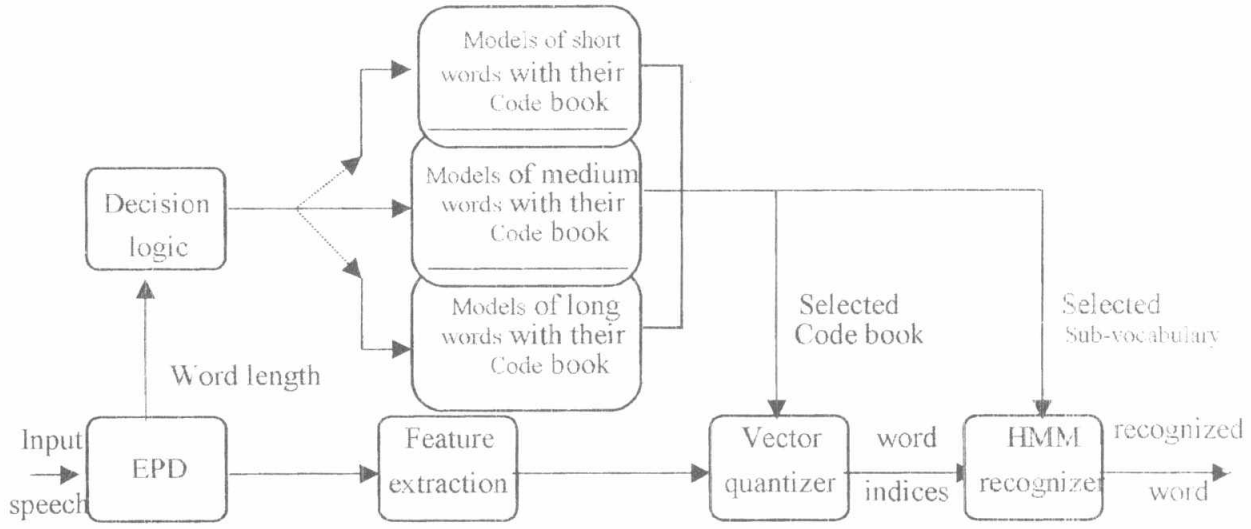



Figure 3. Separation of words according to their lengths with overlap between the sub-vocabularies

## 4.2 Vocabulary Separation After Time Compression:

To overcome the ambiguity that occurs due to the non-standard way of speaking, time compression for both the reference words and the input unknown words has been applied.

Time compression functions to remove the non-necessary elongation of the words. Thus, separation of the vocabulary words is performed after removing the redundancy from speech signals of the reference words. Also, in the recognition phase, the unknown word is compressed first after its feature extraction, then the appropriate sub-vocabulary is selected. figure 4 shows the steps of this approach.

The details of the time compression technique can be found in [3] and figure 5 shows its algorithm. The advantage of applying time compression is that, in addition to reducing V, it also reduces T (in equation 6). Reducing T is due to the time compression of the input unknown word. Thus, although there is an extra processing time due to compressing the input unknown word during recognition phase, faster system response has been obtained. Moreover, this approach has shown recognition accuracy compared with that obtained when manual separation is applied without time compression.

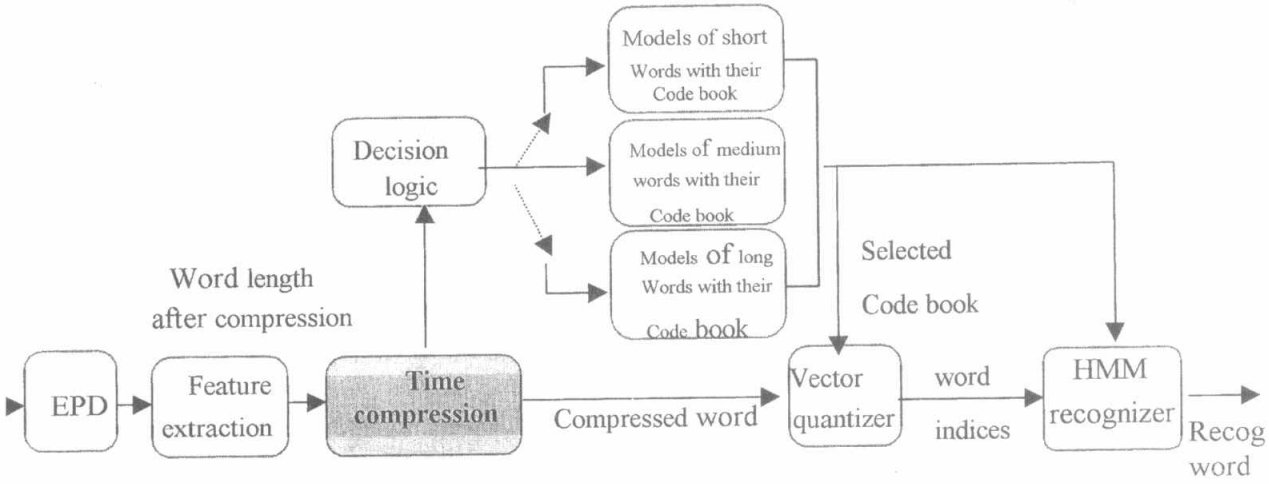



Figure 4. Separation of words according to their lengths after time compression

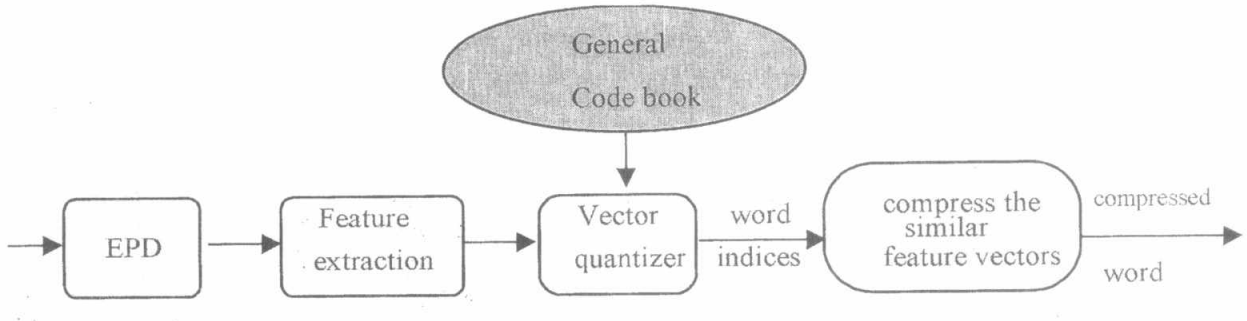


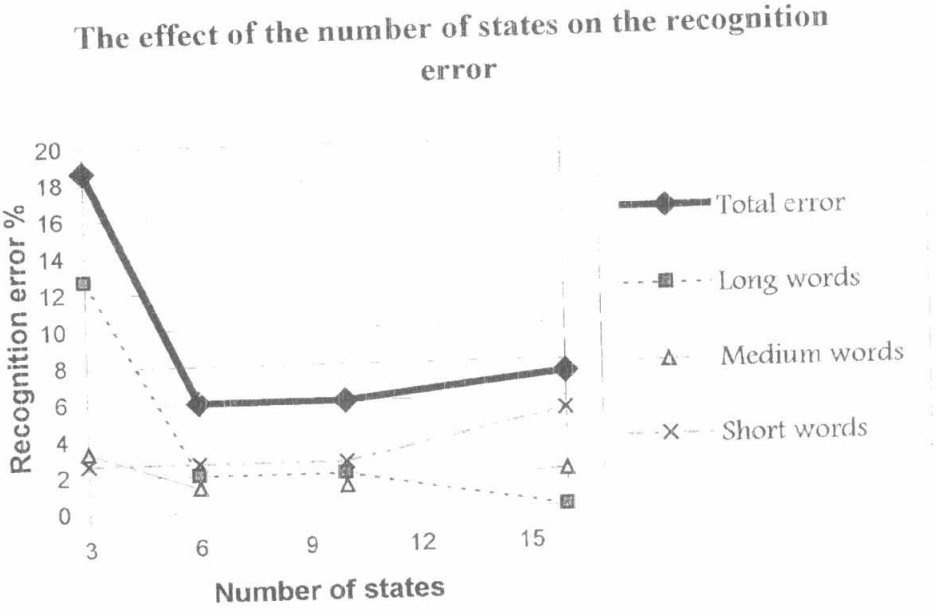

Figure 5. Details of the time compression technique [3]

## V. TEST RESULTS:

The IWR systems introduced above have been tested experimentally using 30 Arabic words vocabulary that consists of 7 short words (such as: في-إلى على-هل-كم ), 13 medium- length

words (such as: صفر-واحد...تسعة ) and 10 long words (such as: ...استفهام-المؤمنين-المستقيم ). Each word has been spoken 10 times by a single speaker. 5 repetitions have been used for training (HMM construction), while the other 5 repetitions have been used for test. The speech signals have been digitized using sound blaster data acquisition card with 10 kHz sampling rate and 8 bits resolution. The system that is shown in figure 1 has been implemented. The end points are detected using the EPD technique that is mentioned above. The speech signals have been analyzed using Perceptual Linear predictive coding (PLP).

The first experiment has been performed to test the effect of the number of states (FNS-HMM for all the vocabulary words) of the HMM of the reference words on the recognition accuracy. Figure 6 shows these results. From this plot, it is clear that when the number of states satisfies the word length (e.g., 3 satisfies the short words while > 10 satisfies the long words), the recognition accuracy is better. From the figure also, it is clear that there is a range of number of states (6 – 10) gives the best average recognition accuracy, which is 94%.




To test the effect of using VNS-HMM, the second experiment has been performed. In this experiment, minimum and maximum numbers of states ($N_{min}$ and $N_{max}$) have been prescribed for each word in the vocabulary according to its phonemic structure and its length. In fact, the vocabulary has been categorized to 3 parts, as mentioned before, which are short, medium, and long. A common number of states has been selected for all the words that belong to each of these three categories. Many experiments have been conducted to test the effect of using different combinations of the numbers of states for the three categories of the words on the recognition accuracy. Table 1 shows these results. From these results, it can be noticed that lower accuracies are obtained when there is a big difference between the numbers of states of the three categories. The best accuracy has been obtained when the numbers of states of the three categories are close to each other (6 for short words, 8 for medium words, and 10 for long words).

Table 1. Recognition accuracy for VNS-HMM

| Max. number of states of the three categories Short / Medium / Long | Recognition error Short / Medium / Long | Total error |
|---|---|---|
| 3 /5 /16 | 3.33% / 1.33% / 0.67% | 5.33% |
| 4 / 6 /12 | 2% / 0.67% / 0.67% | 3.34% |
| 5 / 7 / 10 | 1.33% / 0.67% / 0.67% | 2.67% |
| 6 / 8 / 10 | 0.67% / 0.67% / 0.67% | 2.1% |
| 6 / 6 / 6 | 2.67% / 1.33% / 2 % | 6% |

In the third experiment, the reference models of the vocabulary have been separated manually to three sub-vocabularies according to their lengths (short words, medium words, and long words). The IWR system has been applied and tested for each sub-vocabulary. Different numbers of states as well as different codebook sizes have been tested to obtain the maximum accuracy. Table 2 shows the results of these tests.

Table 2 Recognition accuracies for manually separating and selecting the sub-vocabulary

| Word length | Num. Of states | Codebook size | Rec. accuracy |
|---|---|---|---|
| Short | 3 | 64,128 | 97.14% |
| | 4,5 | 128 | 100% |
| Medium | 5,6 | 128,256 | 98.46% |
| Long | 6,7,8,9 | 128,256 | 98% |
| | 10 | 128,256 | 100% |
| The total and best recognition accuracy | | | 99.3% |

The third experiment has been to test the automation process of selecting the sub-vocabulary with which the input unknown word must be matched. In this test, the input unknown word is directed to the appropriate sub-vocabulary according to its length. Table 3 shows the results of this experiment in which the parameters that showed better recognition accuracies in the previous experiments have been used.

Table 3 Recognition accuracy for direct selection of the sub-vocabulary according to the word length

| Word length | Num. Of states | Codebook size | Recognition accuracy |
|---|---|---|---|
| Short | 4 | 128 | 94.29 |
| Medium | 5 | 128 | 92.3 |
| Long | 10 | 128 | 96 |
| Total recognition accuracy | | | 94% |

To test the effect of overlapping the sub-vocabularies, the reference words that has shown confusion during recognition are added to two sub-vocabularies. This trend has been tested through conducting the fourth experiment whose results are shown in Table 4.

Table 4 Recognition accuracy in the case of overlapping the sub-vocabularies

| Word length | Recognition accuracy |
|---|---|
| Short | 100% |
| Medium | 98.46% |
| Long | 100% |
| Total recognition accuracy | 99.3% |

The last experiment has been to test the effect of applying the time compression technique before the distribution to the appropriate sub-vocabulary. Table 5 shows the results of this experiment. Although the total recognition accuracy is little bit less than that in the cases of manual selection and overlapping the sub-vocabularies, the recognition time becomes shorter.

Table 5 Recognition accuracy in the case of selecting the sub-vocabulary after time compression

| Word length | Recognition accuracy |
|---|---|
| Short | 100% |
| Medium | 98.46% |
| Long | 98% |
| Total recognition accuracy | 98.67% |

## VI. CONCLUSIONS:

In this work, trends has been tested in order to solve the problem of recognition accuracy degradation of an IWR in the case when the vocabulary comprises words with considerable difference in length. These trends have been tested experimentally. According to the obtained results, it can be concluded that classification of such vocabularies into sub-vocabularies according to the words lengths is the best solution. Switching between these sub-vocabularies can be performed automatically by finding an accurate measure to the length of the input unknown word. One of these methods that has been applied in this paper is to remove the unnecessary elongations of the words via time compression. By this way, the length of the input unknown word is decreased and it is directed to the appropriate sub-vocabulary. In addition to that, this process results in reducing the recognition time.

Separation of the vocabulary into sub-vocabularies together with the time compression reduces the recognition time by reducing the two factors that affect the number of computations required for recognition (V and T in equation 6). In addition to that, separation of vocabulary paves the way to increase the total vocabulary size.

## REFERENCES:

[1] L. R. Rabiner and B. Juang, "Fundamentals of Speech Recognition" Prentice-Hall Inc., Englewood Cliffs, NJ 07632, 1993.

[2] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition", Bell system tech j., pp. 1035-1074, April, 1982.

[3] Mahmoud E. Gadallah, "Data Compression For isolated and Connected Word Recognition", Ph.D. thesis, Cranfield inst. Of Tech., U.K., 1991

[4] Mahmoud E. Gadallah, E. Soleit and A. Mahran, ""Noise Immune Speech Recognition System", 16th National Radio Science Conf., Ain Shams Univ., Feb., 23-25, 1999 Cairo, Egypt.

[5] H. Hermanski, B. A. Hanson, and H. Wakita, "Perceptually based linear predictive coding analysis of speech", IEEE Proc. ICASSP'85, vol. 1, pp. 509-512, 1985.

[6] Carnegie Mellon University CMU: Robust Speech Recognition Group (Subset of ARPA speech recognition community). http://www.cs.cmu.edu.