Shady Y. EL-Mashed – Mohammed I. Sharway – Hala H. Zayed
Department of Electrical Engineering, Shoubra Faculty of Engineering, Benha University, Cairo, Egypt
shadyehia@yahoo.com; mshaarawi@gmail.com; hala.zayed@gmail.com

SPEAKER INDEPENDENT ARABIC SPEECH RECOGNITION USING SUPPORT VECTOR MACHINE

Abstract

Though Arabic language is a widely spoken language, research done in the area of Arabic Speech Recognition is limited when compared to other similar languages. Also, while the accuracy of speaker dependent speech recognizers has nearly reached to 100%, the accuracy of speaker independent speech recognition systems is still relatively poor.

This paper concerns with the recognition of speaker independent Arabic speech using Support Vector Machine. The proposed model is applied on the connected Arabic digits (number) using Neural Networks as an example. Also we can apply the system to any other domain.

A spoken digit recognition process is needed in many applications that use numbers as input such as telephone dialing using speech, airline reservation, and automatic directory to retrieve or send information.

This has been realized by first building a corpus consisting of 1000 numbers composing 10000 digits recorded by 20 speakers different in gender, age, physical conditions…, in a noisy environment. Secondly, each recorded number has been digitized into 10 separate digits. Finally these digits have been used to extract their features using the Mel Frequency Cepstral Coefficients **(MFCC)** technique which are taken as input data to the Neural Networks for the recognition phase.

The performance of the system is nearly **94%** when we used the Support Vector Machine **(SVM)**.

Keywords: *Automatic Speech Recognition; Arabic Digits; Neural Networks; Support Vector Machine*

1. Introduction

Automatic Speech Recognition (ASR) is the process of converting captured speech signals into the corresponding sequence of words in text.

It can be used, for example, in a car environment to voice control non critical operations, such as dialing a phone number or on-board navigation by GPS, presenting the driving route to the driver.

It can also be used to facilitate for people with functional disability. With their voice they could operate their PC's, operate the light switch, turn off/on the coffee machine or operate some other domestic appliances.

It is used also in so many applications such as learning foreign language, learning the correct reading of the holly Quraan, speech interfaces (increasingly on mobile devices) and indexing of audio/video databases for search. Human machine interaction is switching from buttons and screens to speech. Speech recognition is an important element in this interaction [1].

The need for highly reliable ASR lies at the core of such rapidly growing application areas.

Speech recognition is, in its most general form, a conversion from an acoustic waveform to a written equivalent of the message information. **Figure (1)** shows a basic speech recognition system [2].

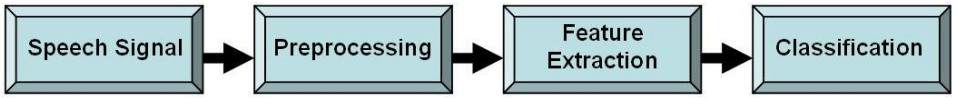


Figure 1: *The basic speech recognition system*

"Speech signal processing" refers to the operations performed on the speech signal (e.g., noise reduction, digitization, spectral analysis, etc.). "Feature extraction "is a pattern recognition term that refers to the characterizing measurements that are performed on a pattern or signal. These features form the input to the classifier that recognizes the pattern.

The difficulty of automatic speech recognition is coming from many aspects of these areas. The following are some of the difficulties that come from speaker variability and environmental interference [3]:

a. Variability from speakers:

A word may be uttered differently by the same speaker because of illness or emotion. Different speakers vary according to their gender, age, way of speaking, speech style, and dialect.

b.Variability from environments:

The acoustical environment where recognizers are used introduces another layer of corruption in speech signals. This is because of:

- Background noise.
- Room reverberation.
- Microphones with Different Characteristics.
- Transmission Channels – Telephone.

c. Continuity of Natural Speech

Natural Speech is continuous not isolated words. There are difficulties in separating words.

Figure (2) shows the typical sources of distortion in adverse environments [3].

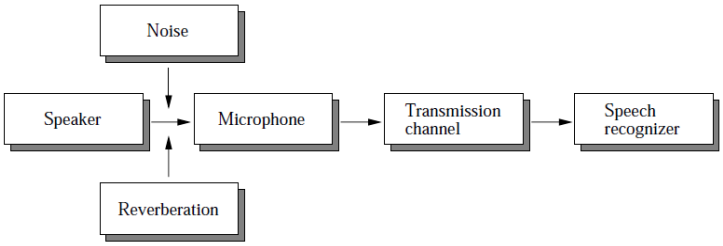


Figure 2: *Source of Distortion in adverse Environment*

2. Related Work

Though Arabic language is a widely spoken language, research done in the area of Arabic Speech Recognition (ASR) is limited when compared to other similar languages. Also while the accuracy of speaker dependent speech recognizers has achieved best performance, the performance of speaker independent speech recognition system is still relatively poor.

Some researches have been done in the area of Arabic speech recognition; we can mention some of relevant researches:

In 2006 Abderrahmane Amrouche and Jean Michel Rouvaen [4] have been presented an efficient system for independent speaker speech recognition based on neural network approach. Its architecture comprises two phases: a preprocessing phase which consists in segmental normalization and features extraction and a classification phase which uses neural networks based on nonparametric density estimation namely the general regression neural network (GRNN). The proposed model has several advantageous characteristics such as fast learning capability, flexibility network size, and robustness to speaker variability which means ability to recognize the same words pronounced in various manners.

In 2008 E. M. Essa, A. S. Tolba and S. Elmougy [5] designed a system for recognition of isolated Arabic words by using a combined classifier. A combined classifier is based on a number of Back-Propagation/LVQ neural networks with different parameters and architectures. And The MFCC features used. For the unseen test set, the recognition rate of the Back-Propagation combined classifier was 96% and that of the LVQ combined classifier was 86.6%. The best individual classifiers resulted in 93% correct classification.

Also in **2008** Akram M. Othman and May H. Riadh [2] designed a system using the scaly type architecture neural network for the recognition of isolated words for small vocabularies as it gave (79.5-88) % success. The scaly type needs (426) iterations to reach acceptable error of (0.01), while the fully connected type needs (2394) iterations. Recognition of the words was carried out in speaker dependent mode. In this mode the tested data is presented to the network are different from the trained data. The Linear Prediction Coefficient (LPC) with 12 parameters from each frame has been used and has improved a good feature extraction method for the spoken words.

Also in **2008**, Yousef Ajami Alotaibi, Mansour Alghamdi, Fahad Alotaiby [6] designed a spoken Arabic digits recognizer system to investigate the process of automatic

digits recognition. This system is based on HMM and by using Saudi accented and noisy corpus called SAAVB. This system is based on HMM strategy carried out by HTK tools. This system consists of training module, HMM modules storage, and recognition module. The overall system performance was 93.72%.

In [7], Moaz et al. has proposed a system to recognize the Arabic Alphabet letters spoken by any speaker using artificial neural networks. The system was based on analyzing phonetic isolated Arabic alphabet letters. He used the Principal Component Analysis (PCA) technique for features extraction. PCA coefficients corresponding to each alphabet are used to train multilayer perceptron & feed-forward back propagation neural networks to produce recognized binary codes corresponding to each letter. He showed a 96% detection rate over large dataset.

Arabic ASR Challenges

Arabic language has two main forms: Standard Arabic and Dialectal Arabic. Standard Arabic includes Classical Arabic and Modern Standard Arabic (MSA) while dialectal Arabic includes all forms of currently spoken Arabic in day life and it varies among countries and deviate from standard Arabic to some extent and even within the same country we can find different dialects. While there are many forms of Arabic, there still many common features on the acoustic level and the language level.

MSA phonetics inventory consists of 38 phonemes [6]. Those phonemes include 29 original consonants, 3 foreign consonants, and 6 vowels. Standard Arabic has 34 phonemes, of which six are vowels, and 28 are consonants. Vowels are produced without obstructing air flow through the vocal tract, while consonants involve significant obstruction, creating a nosier sound with weaker amplitude. A phoneme is the smallest unit of sound that indicates a difference in meaning, word, or sentence. Arabic phonemes contain two distinctive classes, which are named pharyngeal and emphatic phonemes. The allowed syllables in Arabic language are: CV, CVC, and CVCC where V indicates a (long or short) vowel while C indicates a consonant. Arabic utterances can only start with a consonant. **Table (1)** shows the ten Arabic digits along with the way to pronounce them and the number and types of syllables in every spoken digit. [8]

Table 1: *Arabic digits*

Dig it	Arabic writing	Pronunciation	Syllables	No. of syllables
1	واحد	wâ-hĕd	CV-CVC	2
2	اثنين	ʿaâth-nāyn	CVC-CVC	2
3	ثلاثه	thâ-lā- thâh	CV-CV-CVC	3
4	أربعه	ʿaâr-bâ-ʿaâh	CVC-CV-	3
5	خمسه	khâm-sâh	CVC-CVC	2
6	سته	sĕt-tâh	CVC-CVC	2
7	سبعه	sûb-ʿaâh	CVC-CVC	2
8	ثمانيه	thâ-mā-nyĕh	CV-CV-CVC	3
9	تسعه	tĕs-âh	CVC-CVC	2
0	صفر	sĕfr	CVCC	1

Some of the difficulties encountered by a speech recognition system that are related to the Arabic language are:

I. Word Knowledge

Word meaning is needed in order to recognize exactly the intended speech.

II. Patterns Variability Caused By Dialectical Differences:

Variability in dialect between Arab countries and even dialectical difference in the same country between different regions causes the word to be pronounced in a different way. This variability in word pronunciation might cause a great difficulty in recognition.

III. Co articulation effects:

The acoustic realization of a phoneme may heavily depend on the acoustic context in which it occurs. This effect is usually called Co articulation. Thus, the acoustic feature of a phoneme is affected by the neighboring phonemes, the position of a phoneme in a word and the position of this word in a sentence. Such acoustic features are very different from those of isolated phonemes, since the articulatory organs do not move as much in continuous speech as in isolated utterances.

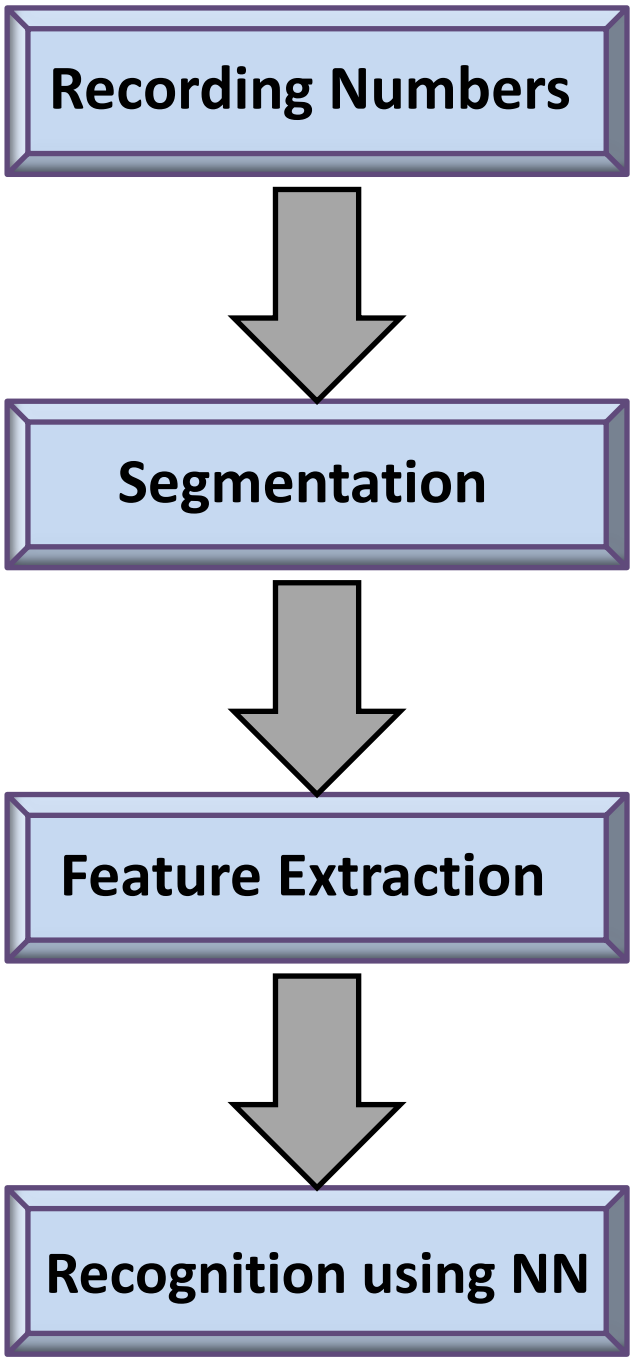
IV. Diacritization:

Diacritics play an important part in written Arabic material. The absence of diacritics in most Arabic texts causes many ambiguities in the pronunciation of words [9], [10], [11], [12].

3. The Proposed System

3.1 System Overview

The system starts by recording Arabic numbers by twenty volunteers (10 males and 10 females) then the segmentation techniques (semi-automatic and fully-automatic) are used to segment these numbers into their digits, then a feature extraction technique is applied to extract the features of the digits, in our system the **(MFCC)** is used to extract these features, finally the Neural Networks **(NN)** is used for the training and testing to recognize the speech which is the target. In this study mainly we use the Support Vector Machine **(SVM)** for the recognition as we can see in the block diagram in **figure (3)**.


Recording Numbers
Segmentation
Feature Extraction
Recognition using NN

3.1.1 Recording System

The Recording System used consists of two tabs as shown in **figure (4)**:

(1) Admin tab:

(a) Audio Settings

Audio Settings tab is used to adjust the parameters used in the digitization process as shown in **Table (2)**.

(b) Data Settings

Data Settings tab contains the volunteer's names to select from them when recording, and the data which is 100 mobile numbers, each one consisting of 10 continuous digits.

(2) Volunteer tab: allows choosing a volunteer name from the stored list and starting recording the displayed number then we stop recording when the speaker finished, after that we can play the sound again to be checked, finally the file is saved automatically in the same location and is given the name of the volunteer followed by the recorded number and the date.

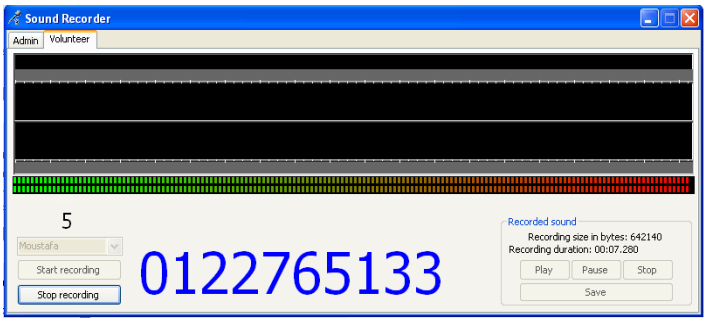


Figure 4: *The recording system*

Table 2: *System Parameters*

Parameter	Value
Database	10 Arabic digits
Speakers	10 males & 10 females
Condition of Noise	Normal life
Sampling rate	8KHz, 16 bits
Accent	Colloquial Egyptian dialect

3.2 Data Set

We construct a database contains of speech waves of twenty speakers (ten males and ten females) varying in region, age, gender, personality, mode… every speaker record about 50 numbers, each number contains ten continuous digits, resulting of 1000 files, Then we automatically segment every file to ten separate digits resulting of the 10000 files (digits).

The system designed to train and test automatic speech recognition engines and to be used in speaker, gender, accent, and language identification systems.

We partitioned these parts of corpus into two separate sets, the first one for the training which is 75% of the database and contain about 7500 recorded digits. On the other hand, the second part is for testing which contains 25% and contains nearly 2500 recorded digits.

Figure (5), shows the information of data set speakers and how many numbers they recorded in the data set.

Figure (6), shows the variety in gender (both male and female) in the data set, The males recorded 5200 digits which are 52 percent; on the other hand the females recorded 4800 digits which are 48 percent of the data set.

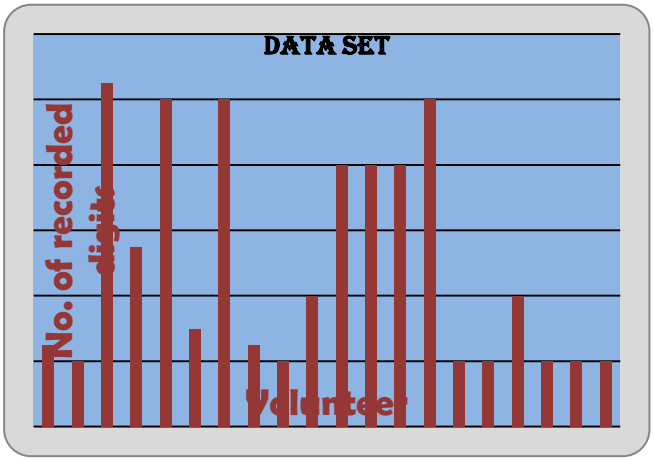


Figure 5: *Data set speakers Information versus recorded number*

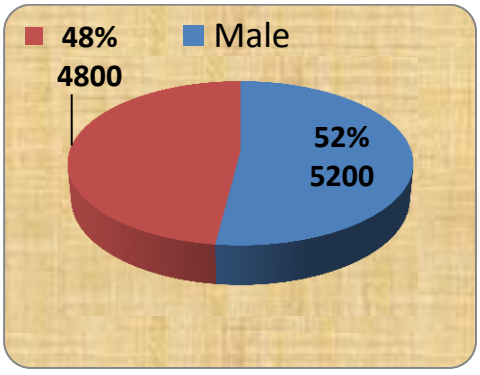


Figure 6: *Gender percent on the data set*

Figure (7), shows the variety in volunteer's ages, the volunteers whose ages are less than twenty recorded 600 digits which is 6 percent, the volunteers whose ages are between twenty and thirty recorded 6800 digits which is 68 percent, the volunteers whose ages are between thirty and forty recorded 1900 digits which is 19 percent, the volunteers whose ages are greater than forty recorded 700 digits which is 7 percent of the data set.

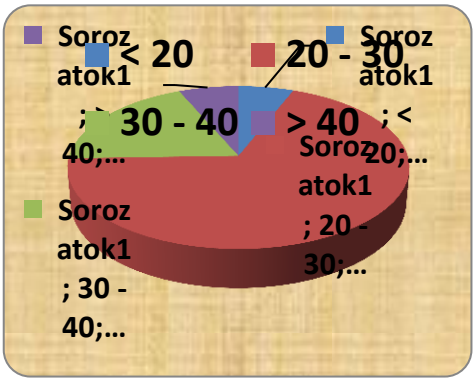


Figure 7: *Age percent on the data set*

3.3 Segmentation System

Speech segmentation plays an important role in speech recognition in reducing the requirement for large memory and in minimizing the computation complexity in large vocabulary continuous speech recognition systems.

With reference to the block diagram shown in **figure (8)**, we apply Fast Fourier Transform (FFT) to the wave file that represent the recorded number by adopting appropriate window size then a band pass filter (from 300HZ to 3400 Hz) is applied on the resulting file to remove the noise from the signal.

Then an Inverse FFT (IFFT) is applied on the resulting file to get the original wave after modification.

After that the segmentation process is implemented by two techniques;

semi-automatic and fully-automatic.

In the semi-automatic technique we adopt the segmentation parameters which are window size, minimum amplitude, minimum frequency, maximum frequency, minimum silence, minimum speech, and minimum word manually by trial &error.

On the other hand with the fully-automatic techniques, these parameters are set automatically to get better performance by using the K-Mean clustering.

"Clustering is the process of partitioning or grouping a given set of patterns into disjoint clusters."[13].

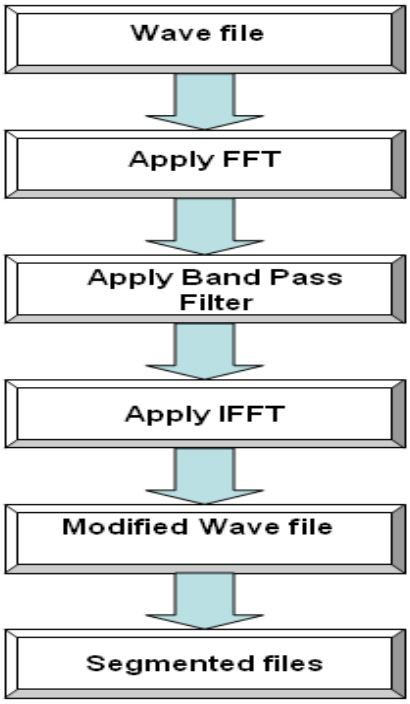


Figure 8: *block diagram of the segmentation process*

The K-Means Algorithm process is as follows: [14]

- The dataset is partitioned into K cluster and the data points are randomly assigned to the clusters.
- For each data point:
  - Calculate the distance from the data point to each cluster.
  - If the data point is closest to its own cluster, we leave it, and if not move it into the closest cluster.
- Repeating the above step until a complete pass through all the data points resulting in no data point moving from one cluster to another.

The K-means method is effective in producing good clustering results in many applications such as our system, here we choose the number of cluster(k) equal to ten, as the our segmentation target is to divide the wave file(number) into ten separate files (digits).

We started the segmentation process by the semi-automatic technique which enables us to realize the sensitivity of the segmentation process to the different parameters values. However the results obtained by it were not satisfactory because in our system the segmentation target is to obtain ten file, but which was not always realized as a result of incapability of dealing with corrupted wave.

3.4 Feature Extraction System

3.4.1 Introduction

After speech is segmented, we need to extract features from the signal. This must be done before we can jump to the next step, which is learning with Neural Networks **(NN)**.

Feature extraction is useful because when the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant (much data, but not much information) then the input data will be transformed into a reduced representation set of features (also named features vector). If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input.

In our system we used the Mel Frequency Cepstral Coefficients **(MFCC)** which we will describe it in the next section.

3.4.2 Mel Frequency Cepstral Coefficients (MFCC)

In the area of speech recognition, the Mel Frequency Cepstral **(MFC)** is a representation of the short term power spectrum of a speech. This representation is based on a linear cosine transform of a log power spectrum on nonlinear Mel Scale of frequency.

Mel Frequency Cepstral Coefficients **(MFCC)** are the coefficients of the **(MFC)**.

There is a difference between the cepstrum and the Mel-frequency cepstrum, which is the frequency bands are equally spaced on the Mel scale, but in case of the Mel-frequency cepstrum, the frequency bands are linearly spaced.

Hence the Mel- frequency cepstrum is better because it approximates the human auditory system's response more closely than the normal cepstrum.

This also allow for better representation of the speech.

Stages of the MFCC are shown in **figure (9)**: [15]

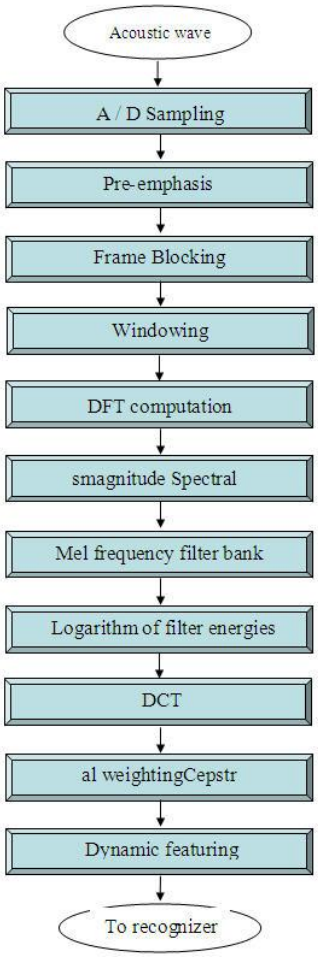


Figure 9: *Block diagram of MFCC*

3.5 A Neural Network Classifier

The Neural Network Techniques were used in our system, there are many Neural Models, Each model has advantages and disadvantages depending on the application. According to our application we choose the Support Vector Machine, first we give a brief note about each it.

Support Vector Machine (SVM):

With reference to **figure (10)**, Support vector machines (SVMs) are a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis.

More formally, a Support Vector Machine (SVM) is implemented using the **kernel Adatron algorithm** which constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. [16]

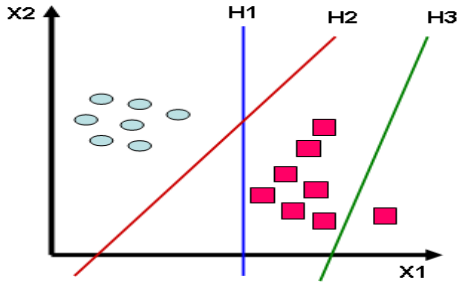


Figure 10: *Support Vector Machine (SVM)*

We build a SVM network contains no hidden layers. The output layer has 10 neurons.

We have 100 inputs and 10 outputs in each case. And we train with maximum epochs of 100 as shown in **figure (11)**.

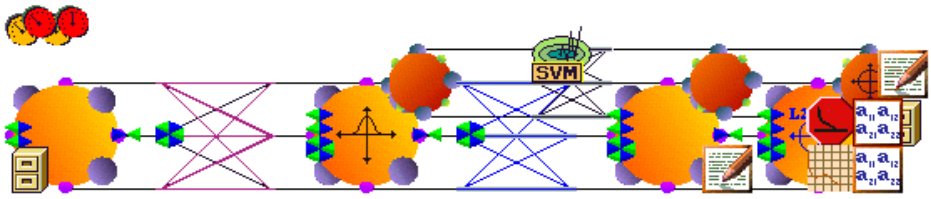


Figure 11: *Architecture of the SVM*

4. Results

A confusion matrix is a simple methodology for displaying the classification results of a network. The confusion matrix is defined by labeling the desired classification on the rows and the predicted classifications on the columns. Since we want the predicted classification to be the same as the desired classification, the ideal situation is to have all the exemplars end up on the diagonal cells of the matrix (the diagonal that connects the upper-left corner to the lower right). [17]

Training

Table (3), shows the Active Confusion Matrix of SVM, and **table (4),** shows the cross validation confusion matrix of the SVM.

Table 3: *Active Confusion Matrix of SVM*

	Col101	Col102	Col103	Col104	Col105	Col106	Col107	Col108	Col109	Col110
Col101	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Col102	0.00	**100.00**	0.00	0.00	0.00	0.00	0.37	0.74	0.00	0.00
Col103	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Col104	0.59	0.00	0.00	**98.79**	0.00	0.00	0.00	0.46	0.00	0.00
Col105	0.00	0.00	0.00	0.00	**99.57**	0.00	0.00	0.43	0.00	0.00
Col106	0.00	0.00	0.00	0.00	0.00	**99.00**	0.00	1.00	0.00	0.00
Col107	0.60	0.00	0.00	0.900	0.00	0.00	**96.70**	1.80	0.00	0.00
Col108	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00
Col109	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00
Col110	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.00	**99.20**

Table 4: *Cross Validation Confusion Matrix of SVM*

	Col101	Col102	Col103	Col104	Col105	Col106	Col107	Col108	Col109	Col110
Col101	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Col102	0.00	**100.00**	0.00	0.00	0.00	0.00	0.37	0.74	0.00	0.00
Col103	0.00	0.00	**98.50**	0.00	0.00	0.00	0.00	0.00	1.50	0.00
Col104	0.60	0.00	0.00	**96.80**	0.00	0.00	0.00	2.60	0.00	0.00
Col105	0.00	0.00	0.00	0.00	**98.50**	0.00	0.00	1.50	0.00	0.00
Col106	0.00	0.00	0.00	0.00	0.00	**99.00**	0.00	1.00	0.00	0.00
Col107	0.00	0.00	0.00	0.00	0.00	0.00	**92.90**	7.10	0.00	0.00
Col108	0.00	0.00	0.00	2.00	0.00	0.00	0.00	**98.00**	0.00	0.00
Col109	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00
Col110	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.00	**98.20**

4.2. Testing

Table (5), shows the results of the testing (which is 30% of the data set).

Table 5: *the testing Confusion Matrix of the SVM*

Output/ Desired	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
0	**255**	5	0	4	0	0	6	0	0	3
1	0	**313**	0	0	1	0	0	0	6	0
2	2	0	**122**	0	0	0	0	0	0	0
3	2	4	2	**105**	0	0	1	0	2	0
4	0	0	0	0	**135**	0	0	6	0	1
5	0	0	0	0	0	**90**	3	0	1	1
6	4	0	0	1	0	4	**83**	0	0	0
7	0	6	0	2	1	0	0	**119**	2	2
8	0	5	0	5	0	0	0	0	**95**	0
9	0	0	0	0	0	0	6	0	0	**95**

Performance =
(255+313+122+105+135+90+83+119+95+95) / 1500
= 1412 / 1500 = **94.13 %**

5. Conclusions

We can jump to the conclusion that, a spoken Arabic digits recognizer is designed to investigate the process of automatic digits recognition. This system is based on a NN and by using Colloquial Egyptian dialect within a noisy environment.

This system is based on NN and carried out by neuro solution tools. The performance of the system is nearly 94% when we use (SVM).

References

[1]H. Satori, M. Harti, and N. Chenfour, "Introduction to Arabic Speech Recognition Using CMUSphix System", http://arxiv.org/abs/0704.2083

[2] Akram M. Othman, and May H. Riadh, "Speech Recognition Using Scaly Neural Networks", World Academy of Science, Engineering and Technology 38, **2008**.

[3] Dongsuk Yuk, "Robust Speech Recognition Using Neural Networks And Hidden Markov Models- Adaptations Using Non-Linear Transformations", Ph. D, Graduate School—New Brunswick Rutgers, The State University of New Jersey, October **1999**.

[4] Abderrahmane Amorite, and Jean Michel Rouvaen, "Efficient System for Speech Recognition using General Regression Neural Network", International Journal of Intelligent Systems and Technologies 1;2 © www.waset.org Spring **2006**.

[5] E. M. Essa, A. S. Tolba and S. Elmougy, "Combined Classifier Based Arabic Speech Recognition", INFOS2008, March 27-29, **2008** Cairo-Egypt

[6] Yousef Ajami Alotaibi, Mansour Alghamdi and Fahad Alotaiby, "Speech Recognition System of Arabic Digits based on A Telephony Arabic Corpus", http://www.mghamdi.com/IPCV08.pdf

[7] Moaz Abdulfattah Ahmad and Rasheed M. El Awady; "Phonetic Recognition of Arabic Alphabet letters using Neural Networks"; International Journal of Electric & Computer Sciences IJECS-IJENS, Vol.: 11, No: 01; February-**2011**.

[8] Yousef Ajami Alotaibi, "A Simple Time Alignment Algorithm for Spoken Arabic Digit Recognition", *JKAU: Eng. Sci.*, Vol. 20 No.1, pp: 29-43 (**2009** A.D. / 1430 A.H.)

[9] Moustafa Elshafei, Husni Al-Muhtaseb, and Mansour Al-Ghamdi, "SPEAKER-INDEPENDENT NATURAL ARABIC SPEECH RECOGNITION SYSTEM", http://www.mghamdi.com/ICIS2008.pdf

[10] N.Uma Maheswari, A.P.Kabilan, R.Venkatesh, "SPEAKER INDEPENDENT SPEECH RECOGNITION SYSTEM USING NEURAL NETWORKS", Received July 4, **2009**, http://jre.cplire.ru/mac/jul09/1/text.pdf

[11] Mohamad Adnan Al-Alaoui, Lina Al-Kanj, Jimmy Azar1 and Elias Yaacoub, "Speech Recognition using Artificial Neural Networks and Hidden Markov Models", The 3rd International Conference on Mobile and Computer Aided Learning, IMCL**2008**, www.imcl-conference.org.

[12] Khalid Saeed, Mohammad K. Nammous, "A New Step in Arabic Speech Identification Spoken Digit Recognition", http://aragorn.pb.bialystok.pl/~zspinfo/arts/KSaeedSpringer.pdf

[13] Khaled Alsabti, Sanjay Ranka, Vineet Singh, "An Efficient K-Means Clustering Algorithm",

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.110.903&rep=rep1&type=pdf
[14]http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/K-Means_Clustering_Overview.htm
[15] YUAN Meng; "Speech Recognition on DSP: Algorithm Optimization and Performance Analysis"; A Master of Philosophy in Electronic Engineering, the Chinese University of Hong Kong; **July-2004**.
[16] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin; "A Practical Guide to Support Vector Classification"; National Taiwan University, Taipei 106, Taiwan; **April-2010**.
[17] http://www.neurosolutions.com