Arabic voice system to help illiterate or blind for using computer

Shaker K. Ali¹ and Zahraa M. Mahdi²

¹ Computer sciences and mathematics college, University of Thi_Qar, Thi_Qar, Iraq.
² College of Education for Pure, University of Thi_Qar, Thi_Qar, Iraq.

shaker@utq.edu.iq and 2 zahraamahda23@gmail.com

Abstract

Speech recognition is one type of technology, which make a computer to recognize the voice of words that an individual speaks through a microphone and convert it into the written text. In this paper, the proposed system for helping illiterate and blind peoples to open applications with them voices.

The proposed system includes two parts; the first part is the training part while the second part is used for the testing. The system contains seven-steps; the first step is the recording of voices and the second step is voice pre-processing. The third step is the feature extraction using MFCC (Mel Frequency Cepstrum Coefficient) method that involves seven steps. The fourth step is for classification voices, there are 1400 voices samples used in training by using a Naïve Bayesian method as a classifier. The fifth step is the matching step using the Correlation Coefficient, there are 200 voices samples in testing. The sixth step was to convert voice into text and the seventh step for execution one of 20 commands.

The accuracy of results from using the Naïve Bayesian algorithm in the training phase gives (100 %) while the accuracy of results in the testing phase using Correlation Coefficient gives (98%).

Keywords: Speech recognition, MFCC(Mel Frequency Cepstrum Coefficient), Naïve Bayesian, Correlation Coefficient.

1 Introduction

Human beings interact with each other in many ways, including speaking words, hand gestures, facial expressions, etc. But speech is considered to be the most important means of human use, because it promotes contact and is the most commonly used by speakers[1]. Speech is a useful term that has a specific meaning and is composed of several words which, in effect, include several letters accompanied by voices[2]. The researchers adapted this
concept and developed it to become a main branch of human communication with the computer where the sound has helped encourage the user's use of the computer and to communicate naturally with them. Automatic speech recognition has significantly contributed to the advancement of artificial intelligence, which helps to create more versatile methods of managing the computer, which enables the user to communicate and share information without using existing input / output modules like the keyboard [1].

Most of the worldwide voice recognition systems are not available in the Arabic language due to its current challenges. The research effort on Arab Automatic Speech Recognition (ASR) is sadly still insufficient, despite its significance. Arabic is one of the most widely spoken languages. About 420 million people speak Arabic worldwide making it one of the six United Nations (UN) spoken language. One may describe three different forms of Arabic language: Classical Arabic, Modern Standard Arabic (MSA), and Colloquial Arabic[3].

2 Speech Recognition

Speech recognition is the mechanism where the human speech is inserted into the device in analog form and the computer translates it into digital form to make it comprehensible. It is the method of translating a speech signal through an algorithm implemented as a computer program to a series of words (i.e. spoken words to text). The job is to get a machine to understand the language spoken. Through "understand" we mean properly responding and translating the input speech into another form [4]. This is too known as "Automatic Speech Recognition" (ASR), "machine speech recognition", "or simply "speech to text" (STT)[1]. The following figure(1) display the block diagram of a speech recognition system [5].
Figure (1): Shows the block diagram of a speech recognition system

3 Related Work

Prerana Das, et al 2015 [6] They are developing a system that allows the computer to use MFCC and VQ techniques to convert voice requests and dictations into text. Using MFCC and Vector Quantization method, feature extraction, and feature matching will be performed. The extracted feature is stored in a file (.mat). A measure of distortion that is based on minimizing the Euclidean distance will be used while matching the unknown speech with the speech signal database.

Jaffar Alkhier 2017 [7] in the process of feature extraction, he proposed three systems for speech recognition, which vary from each other in the methods used. While the MFCC algorithm used by the first system, the second system used the LPCC algorithm, and the PLP algorithm was used by the third system. HMM has been used as a classifier by all these three systems.

Dr.Ch. Raja 2020 [8] he has using standard Mel-Frequency Cepstral Coefficients (MFCCs), and pitch-related features derived from a speech segment, automatic detection of emotions has been tested. Such acquired characteristics are then categorized using the Naïve Bayes classifier. The accuracy of recognition for these characteristics is taken into consideration as it mimics the perception of the human ear.

4 Proposed System

The proposed system includes two parts as shown in Figure (2), the first part is for training while the second part is used for the testing. The system contains seven-phase; the first phase is recording the voices, the second phase is voice preprocessing, the third phase is the feature extraction using MFCC method that involves seven steps, the fourth phase is the classification phase, using 1400 samples using a Naïve Bayesian method for training. The fifth phase is the matching phase using the Correlation Coefficient, using 200 samples for testing. The sixth phase was converted Voice into text and the seventh phase has executed the command for one from 20 command.
4.1 Recording Commands Voice

We collected our database by recording voice samples from different persons (males and females). There are 1600 samples recorded for 20 commands used for training and testing. All the voice's signals were obtained under different conditions, such as the length of record time and the level of sound amplitude. Each command was recorded in the format (.wav) of 16000Hz.

4.2 Speech Preprocessing

Preprocessing is done before feature extraction to remove the noise in the voice signal and to improve recognition accuracy. Preprocessing operations such as pre-emphasis, framing, windowing, and endpoint recognition must be conducted prior to processing and analyzing the voice signal. The extraction of speech features is affected by speech preprocessing.
4.3 Feature Extraction Using MFCC

At this phase, MFCC technology was used to help extract the characteristics of a voice signal that helps to distinguish the content of this one signal from others, which is one of the best methods and most widely used to extracted features for speech recognition. In the training and testing stage of the classification algorithm and matching, these characteristics were use. In the input signal, MFCC applies those steps shown in Figure (3) [12]:

![Figure (3): Shows steps involved in MFCC features extraction](image)

4.4 Naïve Bayesian Classification

The Naïve Bayesian algorithm is a classical classification algorithm that in different applications has proven its simplicity and efficiency [15]. The Naïve Bayes classification algorithm is a simple and very efficient method of probabilistic classification and shows that good performance can be comparable to neural networks and decision tree learning in some areas. In the large data set, Naive Bayes showed acceptable speed and accuracy, but in the case of the small data set, the effect is noticeably poor. However, the classification performance of this system is still strong in the limited sample set[16].

Naive Bayesian classifiers presume that the effect on a given class of an attribute value is independent of the other attribute values. Class conditional independence is called this assumption. It is designed to simplify the computations involved and is called "naïve" in this context[15]. In classification, the Bayes methods are common because they are optimal. It's important that prior probabilities and pattern distribution to class must be known in order to apply Bayes methods. The pattern is assigned to the upper class of posterior likelihood [17]. It is possible to represent the Bayesian theorem as:
\[ P(H|X) = \frac{P(X|H) P(H)}{P(X)} \]  \hspace{1cm} (1)

'X' is a tuple from a database 'D' and the probability of a certain tuple 'X' falling under a particular class is H. "P(X|H)" and "P(H|X)" represent the posterior probabilities indicating the conditional evaluations of 'X' on 'H' and 'H' on 'X', respectively. "P(H)" and "P(X)" is prior probabilities and are unconditional, which means that they are not dependent on others[18].

4.5 Matching using Correlation Coefficient

The coefficient of correlation is a statistical calculation representing the degree of linear correlation between the two variables, with a value range of [-1,1]. The higher the absolute value, the better the correlation is. Otherwise, the weaker it is[19]. In statistical analysis, pattern recognition, and image processing, the approach of correlation is commonly used. The Correlation principle is used to recognize the spoken word perfectly[20].

Cross-correlation is a measure of the similarity between two waveforms, in communication and signal processing as a function of a time-layer related to one of them. These are typically used to search for a shorter, recognized feature for a long-signal. Which also has applications in pattern recognition[21]. This is described in the following equation (2.15):

\[ R1 = \frac{\sum (x_i - x_m)(y_i - y_m)}{\sqrt{\sum (x_i - x_m)^2} \sqrt{\sum (y_i - y_m)^2}} \]  \hspace{1cm} (2)

Their \( x_i \) is the intensity of the i pixel in voice 1, \( y_i \) is the intensity of the i pixel in voice 2, \( x_m \) is the mean intensity of voice 1, and \( y_m \) is the mean intensity of voice 2.

5 Database Recording

The recorded voices contain 20 commands which recorded in the Arabic language. The data recorded different persons (males and females). The total dataset consists of 1600-recorded samples, these data divided into 1400 voices samples for the training stage and 200 voices samples for testing. Where the voice is recorded using a laptop microphone with a sampling frequency of 16000Hz, in format (.wav). The list of 20 commands shown in Table 1.1.

Table (1): List of commands that used in the system.
6 Results

The results will start from preprocessing into commands execution as explain in following steps:

6.1 Speech Pre-processing

The voices or speech firstly pre-processing step to remove noise and deal by filter voices also segmentation used to divide voice into frames. The results of preprocessing for command (افتتح المستندات) as shown in Figure (4).

![Figure (4): pre-processing for command(افتتح المستندات)](image)

6.2 Features Extraction

At this phase, we extracted features of 20 commands for 1400 samples using the MFCC algorithm in the training stage and store the results in file with format (.mat) as shown in Figure(5).
Figure (5): MFCC for command (افتتح آلة الحاسبة)

6.3 Training by Naïve Bayesian Classification

In the training stage, Naïve Bayesian was used for training the classes of 20 commands and store the results in a database to be compared with new voices in the testing stage. The result of applying Naïve Bayesian algorithm using the measurement accuracy rate that compares the features between voices with a database of the number of 1400 as shown in Table (4.2).

Table (2): shows the accuracy rate using Naïve byes algorithm

| Type           | No Sample | Accuracy Rate | Execution time |
|----------------|-----------|---------------|----------------|
| Commands voice | 1400      | 100 %         | 2 ms           |

6.4 Testing by Correlation Coefficient

In the testing stage, the Correlation Coefficient was used for testing the classes of 20 commands and compare the new voices with the database. The result of applying Correlation Coefficient using the measurement accuracy rate in the testing stage by using equation (2) which compares the new features of input voice with the features of database voices as shown in Table (3).

Table (3): shows the accuracy rate using Correlation Coefficient

| Type           | No Sample | Accuracy Rate | Execution time |
|----------------|-----------|---------------|----------------|
| Commands voice | 200       | 98 %          | 2 ms           |

6.5 Convert voice into text
In this stage, the voice will be convert into text after the matching phase has been complete, and this text will be the commands that will be send to the Python program where the required command of the person to be execute as shown in Figure(6).

Figure (6): shows the execution for command (afaq لحة التحكم).

6.6 Execute Voice Commands

In this stage, the last result (text) from the system will appear, which will implement the voice commands required to open the (asked) applications in computer screen as shown in Figure(7).

Figure (7): shows the result for command (afaq لحة التحكم).

7 Conclusions
The aim of this paper to build system may help the illiterate or blinds persons to use the computer using them voice to open the most important orders in the computer without using them hands. We find that the signal pre-processing stage is most important stage before feature extraction and pattern matching which is the main elements of a speech recognition system. The feature extraction stage was gave high accuracy when we use the MFCC algorithm where there are 13 important features extracted for each voice signal. In the training phase, we found that the Naïve Bayesian Classification has proven its simplicity and efficiency method of probabilistic classification and shows good performance. In the testing phase, we found that the Correlation Coefficient is a measure of the similarity between two waveforms, and the reason of using this method as used to search for a shorter, recognized feature for a long-signal.

References

[1] V.Vijayakumari, “A Study on Automatic Speech Recognition”, Journal of Information Technology Review Volume 10, Number 3, August, 2019.

[2] Pahini A. Trivedi, “Introduction to Various Algorithms of Speech Recognition: Hidden Markov Model, Dynamic Time Warping and Artificial Neural Networks”, International Journal of Engineering Development and Research (IJEDR), Volume 2, Issue 4 | ISSN: 2321-9939, 2014.

[3] Mohammad Abushariah, Raja Ainon, Roziati Zainuddin, Moustafa Elshafei, and Othman Khalifa, “Arabic Speaker-Independent Continuous Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus”, The International Arab Journal of Information Technology, Vol. 9, No. 1, January 2012.

[4] Saloni and Dr. Williamjeet Singh, “Multilingual Speech to Text Conversion – A Review, Technologies, And Challenges”, Springer, 2020.

[5] Amer M. Elkourd, “Arabic Isolated Word Speaker Dependent Recognition System”, Islamic University, Gaza, Palestine, Deenery of Higher Studies , Faculty of Engineering, Computer Engineering Department, 2014.

[6] Prerana Das, Kakali Acharjee, Pranab Das and Vijay Prasad, “VOICE RECOGNITION SYSTEM: SPEECH-TO-TEXT”, Journal of Applied and Fundamental Sciences, JAFS|ISSN 2395-5554|ISSN 2395-5562, Vol 1(2), November 2015.

[7] Dr. Jaffar Alkhier, “Improvement of Speech Recognition by Merging Two Features Extraction Algorithms”, Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series Vol. (39), No. (1), 2017.

[8] Dr. Ch. Raja, “RECOGNITION OF SPEECH EMOTION USING NAIVE BAYES CLASSIFIER”, AEGAEUM JOURNAL, Volume 8, Issue 6, ISSN NO: 0776-3808, 2020.
[9] Sabreen Khalid Saud, “SOCIAL COMMUNICATION BETWEEN THE BLIND AND THE DUMB”, M.Sc. Thesis, Thi-Qar University, College of Education for pure science, Department of Computer Science, 2020.

[10] ANJALI BALA, ABHIJEET KUMAR and NIDHIKA BIRLA, “VOICE COMMAND RECOGNITION SYSTEM BASED ON MFCC AND DTW”, International Journal of Engineering Science and Technology, Vol. 2 (12), 7335-7342, 2010.

[11] Shivanker Dev Dhingra, Geeta Nijhawan and Poonam Pandit, “ISOLATED SPEECH RECOGNITION USING MFCC AND DTW”, International Journal of Advanced Research in Electrical, Vol. 2, Issue 8, August 2013.

[12] Ayisha Zubain Bhandari and Prof. B.C. Melinamath, “A Survey on Automatic Recognition of Speech via Voice Commands”, International Journal of New Innovations in Engineering and Technology, Volume 6 Issue 4, ISSN : 2319-6319, January 2017.

[13] Pooja Prajapati and Miral Patel, “Feature Extraction of Isolated Gujarati Digits with Mel Frequency Cepstral Coefficients (MFCCs)”, International Journal of Computer Applications (0975 – 8887), Volume 163 – No 6, April 2017.

[14] Parwinder Pal Singh and Pushpa Rani, “An Approach to Extract Feature using MFCC”, IOSR Journal of Engineering (IOSRJEN), ISSN (e): 2250-3021, ISSN (p): 2278-8719, Vol. 04, Issue 08, August 2014.

[15] S. Karthika and N. Sairam, “A Naïve Bayesian Classifier for Educational Qualification”, Indian Journal of Science and Technology, Vol 8(16), July 2015.

[16] Yuguang Huang, Lei Li, “NAIVE BAYES CLASSIFICATION ALGORITHM BASED ON SMALL SAMPLE SET”, Proceedings of IEEE CCIS, 2011.

[17] Harshad M. Kubade, “The Overview of Bayes Classification Methods”, International Journal of Trend in Scientific Research and Development (IJTSRD), Volume 2, Issue 4, 2018.

[18] Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques”, Book, Elsevier Inc, 2006.

[19] Yanping Zhang and Limin Liu, “Using Computer Speech Recognition Technology to Evaluate Spoken English”, Educational Sciences: Theory & Practice, vol. 18(5), October 2018.

[20] Noor Ali Abed Al-Helaly, “A New Approach for Arabic Text Extraction, Recognition, and Speaking from Video Film”, M.Sc. Thesis, Thi-Qar University, College of Education for pure science, Department of Computer Science, 2016.

[21] Arnab Pramanik and Rajorshee Raha, “Automatic Speech Recognition using correlation Analysis”, IEEE, 2012.