Building a System for Arabic Dialects Identification based on Speech Recognition using Hidden Markov Models (HMMs)

ZAKARIA SULIMAN ZUBI[1], EMAN JIBRIL IDRIS[2]

[1,2]Department of Computer Science, Faculty of Science, Sirte University, Sirte, Libya

Abstract: - In fact, millions of people in the world speak many languages. In order to communicate with each other, it is necessary to know the language we use to perform this operation. The Arabic language has many different dialects and it must be recognized before using the automatic speech recognition (ASR). On the other hand, it is observed in all Arab countries that the standard Arabic language is widely written and used in an official speech, newspapers, public administration, and schools but it is not used in the daily conversations instead the dialect is widely spoken in daily life and rarely written.

In this paper, we examine the difficult task of properly identifying various Arabic dialects and propose a system developed to identify a set of four regional and modern standard Arabic speeches, based on speech recognition using Hidden Markov Models (HMMs) algorithms. HMMs have become a very popular way to build a speech recognition system. It is set as hidden states and possibilities of transition from one state to another. Due to the similarities and differences between the Arabic dialects, speeches collected from the ADI5 datasets were retrieved from the MGB-3 challenge source. We proposed an Arabic Dialect Identification System called "Building a System for Arabic Dialects Identification based on Speech Recognition using Hidden Markov Models (HMMs)" that takes Input as speech utterances and produces output as dialect being spoken. During the training phase, speech utterances from one or more dialects were analyzed to capture the important properties of audio signals in terms of time and frequency. During the testing phase, previously unseen test utterances were utilized to the system, and the system outputs the dialect associated with the model of dialect that most closely matches the test utterance. The proposed model of the system shows promising results of the model for each dialects match.

Key-Words: - Arabic Dialect Identification (ADID), Hidden Markov Models (HMMs), Automatic Speech Recognition (ASR).

1 Introduction

The principle of dialects in any language represents a challenge to Machine Learning (at Automatic Speech Recognition (ASR) systems) and many important Natural Language Processing (NLP) applications such as machine translation, social media analysis, etc. Since a great deal of work on the, automatic identification (AID) of languages from the speech signal alone were accomplished widely. Recently, dialect identification has begun to receive attention from the speech science and technology communities. Spoken dialect identification (DID) is the process of identifying the spoken dialect within speech. This task must be performed without knowing any information about spoken speech. The Arabic language has multiple variants, including Modern Standard Arabic (MSA), the formal written standard language of the media, culture, and education, and the informal spoken dialects that are the preferred method of communication in daily life. While there are commercially available Automatic Speech Recognition (ASR) systems for recognizing MSA with low error rates (typically trained on Broadcast News), these recognizers fail when a native Arabic speaker speaks in his/her regional dialect. Even in news broadcasts, speakers often mix between MSA and dialect, especially in conversational speech, such as that found in interviews and talk shows. Being able to identify dialect via MSA as well as to identify which dialect is spoken during the recognition process will enable ASR engines to adapt their acoustic, pronunciation, morphological, and language models appropriately and thus improve recognition accuracy [1].

The root of every current Automatic Speech Recognition (ASR) system basically consists of a set of statistical models that display the different sounds of the language to be identified. Hidden Markov models are one way to automatically recognize spoken speech. Speech has a temporal structure and can be disguised as a series of spectral vectors that cover a wide range of sound

frequencies. Hence the Hidden Markov Model (HMM) provides a natural framework for building such models [2].

In addition, the Hidden Markov Model (HMM) is one of the most important machine learning models used for the purpose of Automatic Speech Recognition (ASR) systems for the task of dialect identification. The Hidden Markov Model is the basis for a set of successful acoustic modeling techniques in speech recognition systems. The reasons for this success are due to the analytical ability of this model in the phenomenon of speech and its accuracy in practical speech recognition systems.

1.1 Identification of Arabic dialects

Dialect Identification (DID) problem is a special case of the more general problem of Language Identification (LID). LID refers to the process of automatically identifying the language class for a given speech segment or text document, while DID classifies between dialects within the same language class, making it a more challenging task than LID. A good DID system used as a front-end to an automatic speech recognition system can help improve the recognition performance by providing dialectal data for acoustic and language model adaptation to the specific dialect being spoken [3].

The Applications of speech based DID can be broadly categorized into two classes in relation to their end users: human operators or machines. As for human operators, speech based DID systems can be used in routing calls, provisioning assistance and more. On the other hand, for the machines, numerous domains use DID such as: detection and classification of spoken documents, document retrieval, enhancing the performance of automatic speech/speaker recognition [4].

Most dialects identification systems operate in two phases: training and recognition. During the training phase, the typical system is presented with examples of speech from a variety of dialects. Fundamental characteristics of the training speech then can be used during the second phase of dialect identification: recognition. During recognition, a new utterance is compared to each of the dialect dependent models. Each dialect has characteristics that are different from one dialect to another. We need to examine the sentence as a whole to determine the acoustic signature of the dialect, the unique characteristics that make one dialect sound distinct from another [5]. In figure 1, we illustrate the variations in dialects across the Arab world. The figure shows that dialects are a continuum that often transcends geographic regions and borders.

Acoustic features are obtained from the raw speech signal and these are extracted without knowledge of language. Dialect Identification DID also have three major phases having feature extraction, training and testing phase. The methodology of feature vectors extraction and kind of feature vector effects the performance of DID as these feature vectors are input for training and testing phase.

In training phase, generally reference models are created such that one for each language using statistical models like Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Neural Networks (NN) ..etc. These reference models are a compact representation of a huge speech corpus of particular language. In testing phase, the input test speech utterance is labelled with one of known language (represented by reference models) based on the decision criteria

Arabic dialects differ across several dimensions: mainly according to geography and social class. As for the geographical aspect of the language, the Arabic dialects can be divided in many different ways. The following is only one of many (and not all members of any particular dialect group should be considered completely linguistically homogeneous):

In this study, we will test our approach on the following four Arabic dialects with **Modern Standard Arabic (MSA)**.

- **Gulf Arabic (GLF):** includes the dialects of Kuwait, Saudi Arabia, Bahrain, Qatar, United Arab Emirates, and Oman.
- **Levantine Arabic (LEV):** includes the dialects of Lebanon, Syria, Jordan, Palestine, and Israel.
- **Egyptian Arabic (EGY):** covers the dialects of the Nile valley: Egypt and Sudan.
- **North Africa (NOR):** covers the dialects of Morocco, Algeria, Tunisia, and Mauritania.

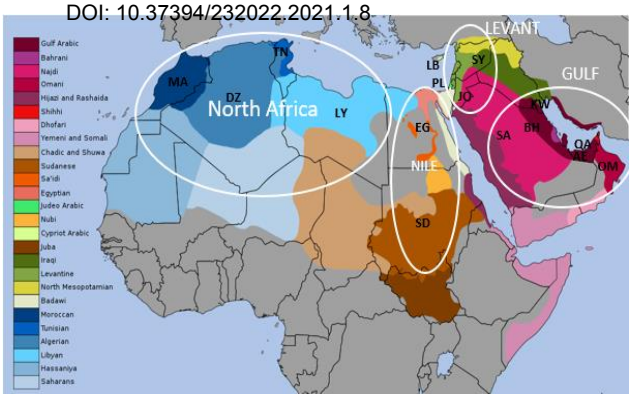


Figure 1. Geographical distribution of Arabic dialects.(Source:https://en.wikipedia.org/wiki/Varieties_of_Arabic, country codes and regions are added).

1.2 Related Work

There are many related works that exam the dialect and identified it in many ways. In reviewing these works we focus on two main factors:

- Type of recognition system used.
- Results have been obtained.

The Spoken Arabic dialects identification: The case of Egyptian and Jordanian dialects defined in [M. Al-Ayyoub, 2014][6] designed an acoustic model using fixed-size segmentation for which they extracted the features using Wavelet transform with a significant feature reduction. They deal with two dialects Jordanian and Egyptian. They achieved 97% precision barrier.

The Speech Recognition of Moroccan Dialect Using Hidden Markov Models mentioned, in [B. Mouaz, 2019][7]**,** The purpose of this work are to verify the ability of HMM Speech Recognition System to distinguish the vocal print of speakers, and identify them by giving each of them a specific class. This is done through creating a speech recognition system, and applying it to a Moroccan Dialect speech. By investigating the extracted features of the unknown speech and then comparing them to the stored extracted feature vectors for each different speaker, in order, to identify the unknown speaker. The model utilized in this work was the Hidden Markov Model. The MFCC + Delta + Delta-Delta features performed best reaching an identification score. The accuracy of HMMSRS is about 90%.

Automatic Identification of Arabic Dialects were showed in [M. Belgacem, 2010]**[8],** A new model has been presented in this work based upon the features of Arabic dialects, nine dialects (Tunisia, Morocco, Algeria, Egypt, Syria, Lebanon, Yemen, Golf's Countries, and Iraq); namely, a model that recognizes the similarities and differences between each dialect. The model utilized in this work was the Gaussian Mixture Models (GMM). Therefore, this new initialization process is used and yields a better system performance of 73.33%.

Swedish dialect classification using Artificial Neural Networks and Gaussian Mixture Models, which are indicated in [V. Blomqvist and D. Lidberg, 2017][9], is a thesis which investigated the classification of seven Swedish dialects based on the SweDia2000 database. The classification was done using Gaussian mixture models, which are a widely used technique in speech processing. Inspired by recent progress in deep learning techniques for speech recognition, convolutional neural networks, and multi-layered perceptron's were also implemented. The Gaussian mixture models reached the highest accuracy of 61.3% on a test set, based on single-word classification. Performance is greatly improved by including multiple words, achieving around 80% classification accuracy using 12 words.

Multi-Dialect Arabic Broadcast Speech Recognition were mentioned in [A. M. A. M. Ali, 2018][10], is also a thesis which investigated Multi-Dialect Arabic Automatic Speech Recognition (ASR) with no prior knowledge about the spoken dialect. In this study, they proposed Arabic as a five-class dialect challenge comprising of the previously mentioned four dialects as well as MSA. They also investigated the different approaches for ADI in a broadcast speech. Since, they studied both generative and discriminative classifiers, and combined these features using a Multi-Class Support Vector Machine (SVM), Deep Neural Network (DNN), and Convolutional Neural Network (CNN). They validated their results on an Arabic/English language identification task, with an accuracy of 100%. As well as they evaluated these features in a binary classifier to discriminate between MSA and DA, with an accuracy of 100%.

Arabic Speech Recognition System Based on MFCC and HMMs illustrated in [H. A. Elharati, M. Alshaari, 2020][5], the primary contribution of this work was to design an Arabic ASR system and find the performance of the selected Arabic words that is successfully verified and examined. For this purpose, 24 Arabic words were recorded from native speakers, all the experiments are conducted, and the recognition results of the ASR system were investigated and evaluated. The system is designed

by MATLAB based on MFCC and discrete-observation multivariate HMM. The best recognition rate reaches 92.92% (51 total error counts from 1368 total words count).

An Automatic identification of Arabic dialects using Hidden Markov Models declared in [F. S. Alorifi, 2008][11], used an ergodic HMM to model phonetic differences between two Arabic dialects (Gulf and Egyptian Arabic) employing standard MFCC (Mel Frequency Cepstral Coefficients) and delta features. The best parameter setting of this system achieves high accuracy of 96.67% on these two dialects.

Our proposed system addresses the problem of Arabic dialect identification. On the other hand, we used a dataset of audio examples for Four Arabic dialects (Gulf, Levantine, North Africa and Egyptian Arabic) with Modern Standard Arabic (MSA). The proposed system will use the Hidden Markov Model methods to build the dialects models for the dialect identification task.

2 Motivations and the Statement of the Problem

The motive, through which we chose the subject of this work, is that there is only one standard Arabic language, and all other Arabic languages are Arabic dialects that are considered as derived from it. In most cases all dialects express one thing in the original Arabic language, but in different ways. This motivated us to prompt and define the research problem first in general in the following points: -

(1) Arabic Language research is growing very slowly compared to English Language research. Mainly the reason for this slow growth is due to the lack of recent studies on the phonetic nature of the Arabic language and the difficulties in speech recognition.
(2) Most Automatic Speech Recognition (ASR) systems for Arabic are based on the Modern Standard Arabic (MSA) Language, and in fact, most people speak regional dialects. Therefore, determining the Arabic dialect from the input speech will help ASR in the Arabic language for optimal performance.
(3) The identifying of dialects is still one of the facing problems.
(4) Few recent studies examining Arabic dialects for speech recognition purposes.

Secondly, in particular, we may define the problem of this study as follows:

The issues of speech recognition systems in identifying Arabic dialects can be represented by finding the most suitable sequence of utterance based on the segment of the Arabic dialect sound. Suppose that *O* stands for an acoustic observation sequence, obtained by the sequence of word for each Arabic dialect [in our work the Dialects were Egyptian (EGY), Levantine (LAV), Gulf (GLF), North African (NOR), and Modern Standard Arabic (MSA)].

On the other hand, using the hidden Markov model for recognizing the Arabic dialect identification problem based on speech recognition aims to search for a sequence of words that is translated into a sequence of the Hidden Markov modelλ. Thus, a model reference is created for each dialect λ_D to perform the comparison with unknown words to determine the specific dialect.

The main problem is to determining the Arabic dialect based on the speech recognition conditional maximization holds in the equation 1 as following:

$$D^* = \arg\max P(\lambda_D|O) \qquad (1)$$

Where $D^* = D_1, D_2, \ldots, D_n$, and D is a number of possible Dialects and $O = o_1, o_2 \ldots o_T$, are denoted as a sequence of acoustic observations. Therefore, by applying Bayesian rule to find $P(\lambda_D|O)$ the probability which was computed by using the equation 2 as follows:

$$P(\lambda_D|O) = \frac{P(O|\lambda_D)P(\lambda_D)}{P(O)} \qquad (2)$$

3 The Objectives of the Research

The main objective of this study is to build a system that automatically identifies the Arabic dialects from the input speech model (file) using HMM model. This objective can be achieved in the following means:

(1) We will use the five classes Arabic Dialect Identification ADI-5 dataset from MGB-3 challenge data source to exam our proposed model.
(2) To segment and label Arabic corpora that is suitable for implementing our aim.

(3) To analyze the extracted features using the Mel-frequency Cepstral Coefficients (MFCC) algorithm.
(4) To improve the accuracy of the Arabic dialect identification system to classify and identify Arabic dialects based on Automatic Speech Recognition (ASR) using Hidden Markov Model (HMM).
(5) Testing and evaluating our implemented approach.

4 Methodologies

In this section, the methods and techniques that are used to achieve the objectives of this paper are presented in the following sections.

4.1 Automatic Speech Recognition (ASR) for Dialect Identification (DID)

Automatic spoken language identification is defined as the process that determines the identity of the language spoken in a speech audio sample. The importance of DID can be gauged from the growing interest in automatic speech recognition. A good language recognition system can facilitate labelling the language of a speech segment for many tasks like multilingual speech processing, such as spoken language translation, spoken document retrieval, metadata labelling and multilingual speech recognition [12]. The same principle can be applied on automatic spoken dialect identification that can help reduce the ASR word error rate for dialectal data by training ASR systems for each dialect, or by adapting the ASR models to a specific dialect. Automatic speech recognition (ASR) is a process that converts an acoustic signal, captured by the device microphone or over a telephone line, to a set of textual words. Over the years, ASR systems have been developed for many via-voice applications. Examples include: speech to speech translation, dictation, Computer aided language learning, and voiced based information retrieval etc. [10].

Accurate acoustic models (AM) are a significant requirement of automatic speech recognizers. Acoustic modeling of speech describes the relation between the observed feature vector sequence, derived from the sound wave, and the non-observable sequence of phonetic units uttered by speakers. The major concerns of the automatic speech recognition are determining a set of classification features and finding a suitable recognition model for these features.

We demonstrated HMMs to be a special case of regular Markov models. It became a more powerful model for representing time varying signals as a parametric random process. It works perfectly when some input occurs as a new state will be generated. The "hidden" model in Hidden Markov means that changes from the old state to the new state are not directly observable and that the transition probability depends on how the model is trained with the training sets [13].

Based on that, making the hidden Markov model works, we need to train the model using training sets. The training we used will hold all classes to be graded since the model will only learn from what has been trained. The Training sets will be trained more likely than the test set since the more data we get, the more information the model will be able to learn.

Typically, modern ASR system represents the speech signal using state-of-the-art Mel Frequency Cepstral Coefficients (MFCCs). The Hidden Markov Models (HMMs) are then used to model the MFCCs observation sequence. These features are computed every 10 ms with an overlapping analysis window of 25 ms.

Automatic speech recognition consists from a numerous of component Figure 2, below illustrates the key tasks of ASR and their components [14].

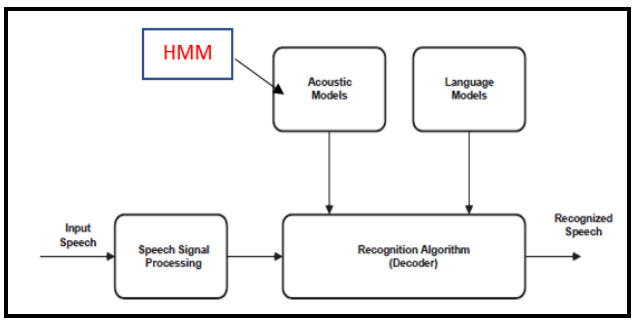


Figure 2. The components and the key tasks of ASR

(1) **Speech Signal Processing**: In this process, the speech signal is converted to a set of feature vectors.

(2) **Acoustic Models:** The representation of knowledge about acoustic, phonetics, and the speaker variability are included in the models. Hidden Markov Models are the foundation for acoustic phonetics models. The acoustic models are modified during training to ensure that system performance is optimized.

(3) **Language Models:** The knowledge of the system about what words are likely to appear together, in what sequence, and what the possible words are.

(4) **The Recognition Algorithm (Decoder):** The most important component on the ASR systems and it was represented as the reason behind the ASR system. For each audio frame, there is a process of pattern matching. Hence, the decoder evaluates the received feature against all other patterns. The best match can be achieved when more frames are processed or when the language model is considered.

Acoustic features are obtained from the raw speech signal and these are extracted without knowledge of language. Dialect Identification DID also have three major phases having feature extraction, training and testing phase. The methodology of feature vectors extraction and kind of feature vector effects the performance of DID as these feature vectors are input for training and testing phase.

In training phase, generally reference models are created such that one for each language using statistical models like Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Neural Networks (NN)etc. These reference models are a compact representation of a huge speech corpus of particular language. In testing phase, the input test speech utterance is labelled with one of known language (represented by reference models) based on the decision criteria.

4.2 Hidden Markov Models (HMMs)

HMM is a probabilistic model for machine learning and language processing. It is mostly used in speech recognition [21], to some extent; it is also applied for the classification task. HMM provides solutions of three problems: evaluation ,decoding, and learning to find the most likelihood classification. The core idea in using HMM for speech recognition applications is to create a stochastic model as shown in figure 3, from known utterances and compares it with the unknown utterances was generated by the speaker. An HMM λ is defined by a set of states N the individual states are denoted by S = {S1; S2; …; SN}, and the state at time t is qt. that have O observation symbols as well as, three possibility metrics for each state which are in (Equation 3) [15].

$$\boldsymbol{\lambda = (A, B, \pi)} \quad (3)$$

Where:

A: a set of state transition probabilities $A = a_{ij}$

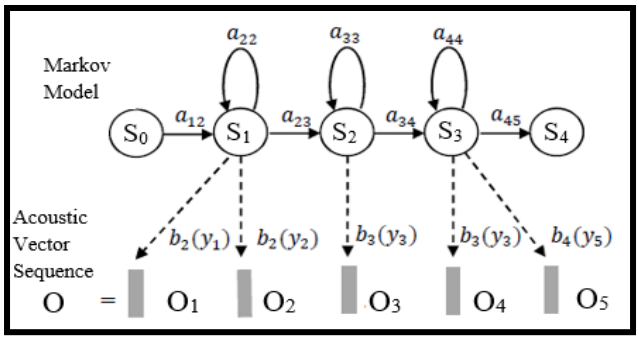
$a_{ij} = P\,[q_t+1 = S_j \mid q_t = S_i\,]$ for $1 \le i\,; j \le N$.

B: A probability distribution in each of the states $B = bj(k)$ in which

$bj(k) = P[Ok \text{ at } t \mid qt = Sj\,]$ where $1 \le k \le M \;\; 1 \le j \le N$.

$\boldsymbol{\pi}$: The initial state distribution $\pi = \pi_i$ in which $\pi_i = P\,[q1 = Si]$ where $1 \le i \le N$.

Figure 3. The stochastic model using HMM for speech recognition



HMMs are designed and analyzed with three associated problems. These problems are:

(1) Evaluation problem: deals with evaluation of probability/likelihood value of observation sequence against given an HMM. in another meaning, computing the likelihood P(O/λ), the probability of model λ emitting observation sequence O= O1; …; OT. With this problem, testing is performed with Forward and Backward algorithms [16].

A. Forward Algorithm

Let αt(i) be the probability of the partial observation circuit O_t = {o(1), o(2), ..., o(t)} to produce all possible state sequences at the i-th state.

$$\alpha t(i) = P(o\,(1), o(2), \ldots, o(t) \mid q(t) = qi) \quad (4)$$

The probability of the partial observation sequence is the sum of αt(i) for all N states.

B. Backward Algorithm

In a similar manner, the backward variable βt(i) as the probability in the partial observation

sequence of o(t+1) to the end that will be generated by all state sequences, starting at state i-th. Backward algorithm counts backward variables back and forth along the observation sequences.

$$\beta t(i) = P(o(t+1), o(t+2), \dots, o(T) \mid q(t) = qi) \quad (5)$$

(2) Learning problem: for training purposes the model is responsible to store data collected for a specific dialect class (i.e., in our work the dialects were (EGY, GLF, LAV, MSA, and NOR). We will adjust the model parameter by $\lambda = (A, B, \pi)$ to maximize $P(O|\lambda)$ [16]? The most difficult thing is to adjust the model parameters (A,B,π) to maximize the probability of a given sequence of observations with this problem, the testing process is performed with Baum-Welch algorithm.

Let ξ t (i, j), the combined probabilities are in qi state at time t and state qj at time t + 1, given the model and sequence observed:

$$\xi_t(i,j) = P(q(t) = q_i, q(t+1) = q_j | O, \lambda) \quad (6)$$

Where it will be obtained by;

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o(t+1)) \beta_{i+1}(j)}{P(O|\Delta)} \quad (7)$$

The output sequence probability can be expressed as follows;

$$P(O|\Delta) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o(t+1)) \beta_{i+1}(j) = \sum_{i=1}^{N} \alpha_t(i) \beta_t(i) \quad (8)$$

The probability will be in state qj at time t;

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) = \frac{\alpha_t(i) \beta_t(i)}{P(O|\Delta)} \quad (9)$$

(3) Decision problem: given observations O = O1, O2, O3 …OT, and model $\lambda = (A, B, \pi)$, is to choose the corresponding state sequence Q = q1 q2 · · · qT which is optimal in some meaningful sense [16]. To solve this problem we will use Viterbi algorithm to compare between the training and the testing data and find otu the optimal scoring path of state sequence by selecting the high probabilities between the model and the testing data.
The Viterbi algorithm chooses the best hidden state sequence that maximizes the likelihood of the state sequence for the given observation sequence. Let δt(i) be the maximum probability of the state sequence, the length t ends with state i and yields the first observation for the given model.

$$\delta t(i) = max\{P(q(1),q(2),\dots,q(t-1); o(1),o(2),\dots,o(t)|q(t)=qi)\} \quad (10)$$

With the advantages of HMM, we use HMM to create a reference model for each dialect included in our paper and design HMMs with different number of states such as three or four or five.

5 The Proposed System Architecture

The aim of this proposed system is to assign an audio signal to each appropriate Arabic dialect entry. The main idea is to compare the phoneme model representing the input audio signal with the reference models for the different Arabic dialects. According to the results of this comparison, we assign the audio input signal to the class that reduces cosine similarity. Figure 4 describes the operation of the proposed system. The diagram below is an abstract view using standard flowchart notation to illustrate the processes and their links.

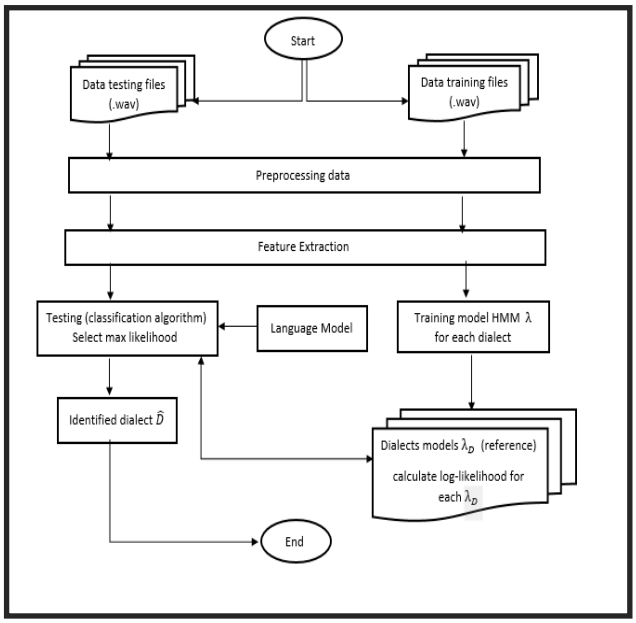


Figure 4. Proposed System Flowchart

5.1 Dataset

In our proposed system, we will use the sample five classes Arabic Dialect Identification ADI-5 dataset obtained from Multi-Genre Broadcast (MGB) competition, to implement the practical part

in the proposal work. This dataset will be used in training and testing the system for each dialect.

The MGB challenge is a core evaluation of speech recognition, speaker diarization, lightly supervised alignment, and dialect identification using TV recordings from the BBC and Aljazeera, as well as YouTube videos [17].

MGB-3 Challenge: The third edition of the MGB challenge is the MGB-3 for ASRU-2017 [18]. MGB-3 focuses on dialectal Arabic (DA) using a multi-genre collection of Egyptian YouTube videos. Seven genres were used for the data collection. The MGB-3 is using 16 hours of multi-genre data collected from different YouTube channels [19]. In 2017, the challenge featured two new Arabic tracks based on TV data from Aljazeera as well as YouTube recordings.

The dataset for the ADI supplied with more than 50 hours labeled for each dialect. This will be divided across the five major Arabic dialects; Egyptian (EGY), Levantine (LAV), Gulf (GLF), North African (NOR), and Modern Standard Arabic (MSA). Table 1 presents some statistics about the training and test datasets [18].

Dialect	Train data		Test data	
	Hours	Utterances	Hours	Utterances
EGY	12.4	3,093	2.0	302
GLF	10.0	2,744	2.1	250
LAV	10.3	2,851	2.0	334
NOR	10.5	2,954	2.1	344
MSA	10.4	2183	1.9	262
Total	53.6	13,825	10.1	1,492

Table 1. The number of hours and utterances of data available for each dialect for training and testing.

5.2 Pre-processing

The functionality of pre-processing stage is to prepare the input signal to the feature extraction stage. The main goal of this phase is to get the speech signal of each word that had been spoken. Thus, this phase handles with any snags or loopholes that might affect feature extraction. It basically tries to remove noises and silence gaps. This is important because noises and silence gaps in the inputs have very inconsistent properties that can cause misidentification as well as segmenting the speech audio file by detecting endpoints. In the Pre-processing stage, a speech waveform transforms into a sequence of parameter vectors. This process will be performed in both the training data and testing data.

5.3 Feature Extraction

Feature extraction is a fundamental part of any speech recognition or identification system like language, dialect, speaker from speech utterances from speech signal. The performance of proposed system depends on feature vectors, the selection of feature vectors and some other parameters of features are very important to get good and significant results.

In case of dialect identification, the selection of feature vectors must discriminate the content of speech i.e., phoneme, sequence of phoneme and frequency of phoneme.

Features can be extracted from speech signal in the frequency or time domains as it's indicated in figure 5, which represents an acoustic features vector. We will use in this phase the Mel Frequency Cepstral Coefficients (MFCC) technique, that which is a popular speech feature representation. Mel Frequency Cepstral Coefficients (MFCC) is an important feature of a speech signal that reveals phonemic differences between dialects. In our work, we extract the MFCC from short term duration of windowed speech signal. This process will be performed on both the training data and testing data.

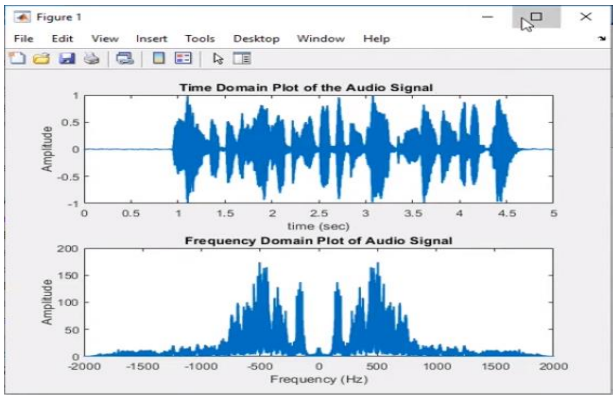


Figure 5. Frequency and Time Domains of Audio Signal.

5.3.1 Mel Frequency Cepstral Coefficients (MFCC)

MFCC are popular acoustic features and these have significant results in speech processing tasks. These features mainly extracted from preprocessed speech signal. The steps to extract MFCC from speech signal are described in Figure 6.

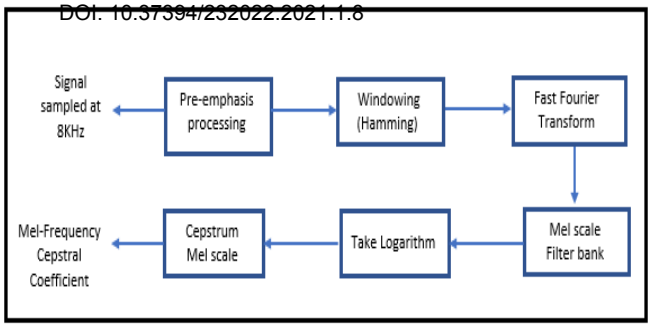


Figure 6. Extraction of MFCC features vectors.

The extraction of Mel Frequency Cepstral Coefficients features vectors from speech signal have several steps:

(1) The signal is smoothed by removing noise with a digital filter (pre-emphasis filter) in order to improve the system's efficiency performance. Here is the pre-emphasis filter equation:

$$y(t) = x(t) - \alpha x\,(t-1) \qquad (11)$$

$y(t)$ is the result of pre-emphasis signal, $x(t)$ is the initial signal prior to pre-emphasis, the constant value for the filter coefficient αx is 0.95.[16]

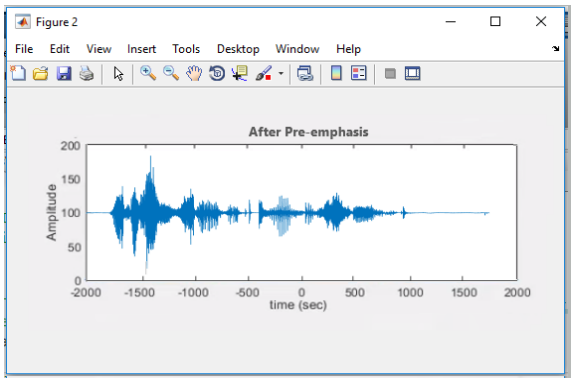


Figure 6. Result of Pre-emphasis.

(2) Frames and windows, instead of analyzing the entire speech signal at once, are divided into overlapping frames with short time duration, and the frame size is generally 10ms-30ms. The information at the beginning and end of the frame is very important. To avoid lose the information; we overlap frames to preserve the information. Windowing technology is implemented as well to avoid stopping in the signal so that a windowing function is applied to each frame length using a hamming window. A mathematical equation hamming window is represented as follows:

$$w[n] = 0.54 - 0.46\cos(2\pi n/N-1) \qquad (12)$$

(3) Apply a fast Fourier transform to get the scale frequency of each frame of windowed speech signal.

Where, N is usually 256 or 512. As well as to calculate the power spectrum (periodgram) will obtained by using the following equation:

$$P = \frac{|FFT(x_i)|^2}{N} \qquad (13)$$

Where, x_i is the i_{th} frame of signal x.

(4) Using Mel scale filter bank, smooth the spectrum of the speech signal that gets the spectrum data values from more significant parts. The Mel frequency scale is a linear frequency below 1000 Hz and a logarithmic space above 1000 Hz. The bank filter will be applied in the frequency domain shown in figure 7. The converting method between Hertz (f) and Mel (m) will be achieved by using the following equation: The formula for converting from frequency to Mel scale; Mel scale is defined as equation (1).

$$m = 2595\log_{10}(1 + \frac{f}{700}) \qquad (14)$$

Where, f is a frequency in Hz.

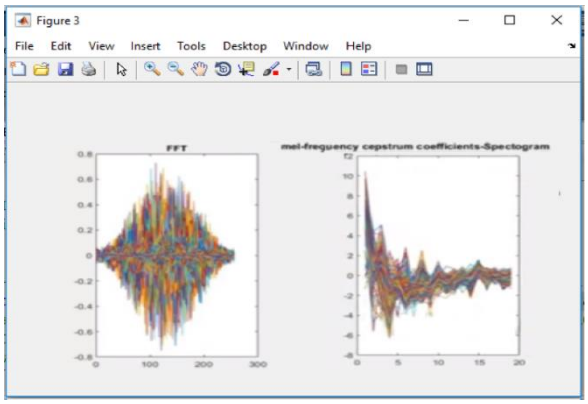


Figure 7. Apply FFT and Mel scale filter bank.

(5) The logarithm is applied to the Mel spectrum which converts the Mel cepstrum to a time domain, i.e. Mel cepstrum, and then a discrete cosine transform (DCT) is applied to the Mel cepstrum to obtain the coefficient. Discrete Cosine Transform (DCT) is the last stage in forming Mel Frequency Cepstrum. The equation used to calculate DCT is indicated as following:

$$Ci = \sqrt{\frac{2}{N}}\sum_{K=1}^{K} mk\cos\left[\frac{\pi i}{N}(K-0.5)\right] \qquad (15)$$

(6) Reduce the interrelations between compressed information and coefficients to coefficients of lower order.

5.4 Training Phase

During training, spectral features vectors are extracted from the digitized of training speech utterances. Given O acoustic features vectors for each Dialect. Then, by using the forward-backward algorithm HMMs are designed as one reference model for each dialect to capture the characteristics of each Dialect spoken within the speech data by initializing randomly the parameters (initial probabilities, transition probabilities, and output densities) of an HMM for each Dialect D, the result is A model set, $\boldsymbol{\lambda_D}$ where D is a number of possible Dialects, then using a forward algorithm to calculate log-likelihood $\mathbf{P(O/\lambda_D)}$ for each $\boldsymbol{\lambda_D}$.

5.5 Language Model (LM)

The Dialect speech is recognized by classification phase by using extracted features and a dialect template where the dialect template contains syntax and semantics related to the responsible dialect which help the classifiers to identify the input utterance. The language model is an N-gram model trained separately which the probability of each word $P(\lambda_D)$ is conditionally obtained on its N-1 predecessors.

5.6 Testing Phase

During testing, we will study the results of the acoustic features vector from the feature extraction phase. That will be extracted from the test sampled data. Using the Viterbi (Decoding) algorithm of the feature vector sequences which will be performed against each of the HMMs, producing a likelihood score that the given test utterance was produced by the λ_D models. The final step is to select the most likely model according to:

$$D^* = \arg\max_{1 < D < Dn} P(\lambda_D|O) \qquad (16)$$

The Dialect of the model most likely to have produced the test utterance observations is hypothesized as the Dialect of the test utterance one of (EGY, GLF, LAV, and NOR) Dialect.

6 Experiments and Results

In this work, the proposed system will be implemented to identify the Arabic audio, which could classify and recognize the dialect of input speech audio. This section will present the results of the identification process, as well as the design and implementation of the proposed identification system. It focuses only on five major dialects such as: NOR, EGY, GLF, and LAV with MSA.

In this work also we build and development an Automatic Speech Recognition (ASR) system using Hidden Markov Model (HMM). The proposed system was implementation and satisfactory performance was developed using MATLAB platform in term to make the system more interactive and faster.

Our system follows a standard recipe. corpus-suggested timings are used to segment the audio data, 13 Mel Frequency Cepstral Coefficients (MFCCs) as well as their 1st and 2nd derivatives were extracted with 10 ms using a 25 ms framed speech signal, and each conversation side was normalized using mean cepstral and variance normalization. All the models were trained with context-dependence triphones by using the maximum likelihood function. Whereas, each phone was modelled using left-to-right HMM and the outcomes are represented in three states.

In this work we investigate the similarity between each pair dialects of Arabic through acoustic models (HMMs), that refer to different type's dialects of Arabic. The performance of this purpose will be illustrated by using cosine similarity. Table 2, presents the results of the classification of dialectical speech and describes the confusion.

The test set of the experiment was defined by 200 utterances of 10-sec and 200 utterances of 45-sec from each target dialect. The score of each token sequence was obtained by summing all the log bigram probabilities given each bigram language model. For dialect identification purpose, a maximum likelihood classifier was finally used to hypothesize the language being spoken in each utterance.

For classifying the dialect in the testing speech, the forward score of the speech utterance must be computed. The five different scores from the different dialect models were processed by the maximum likelihood classifier and the one with the highest log likelihood was taken to be the hypothesized dialect.

True Dialect	Predict Dialect						
	EGY	GLF	LAV	NOR	MSA	Accuracy	Recall
EGY	150	11	8	19	11	75.0	75.3
GLF	5	146	34	7	8	73.0	73.0
LAV	7	23	144	3	6	72.0	78.6
NOR	13	11	9	160	7	80.0	80.0
MSA	25	9	5	10	151	75.5	75.5
Precision	75.0	73.0	72.0	80.40	82.51		

Table 2. The matrix of confusion given by the acoustic model.

In table 2, we found some confusion between Arabic dialects. It is clearly shown that the highest confusion rates are those between GLF and LAV dialects. This confusion is justified by the closeness between these pairs of dialects; e.g., GLF and LAV dialects share significant vocabulary. Figure 8, illustrated the performance details of our proposed system for each Arabic dialect.

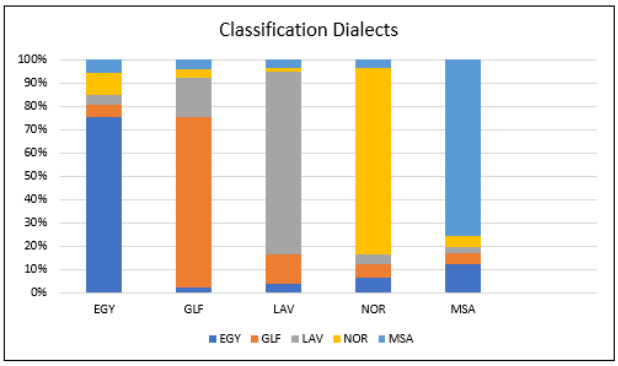


Figure 8, Classification rate for each Arabic dialect

7 Evaluations

Since the dialect identification task is a standard statistical classification problem, Throughout the experiments, the performance metric for ADI test datasets includes Accuracy (ACC), Recall (RCL) (false negative value), and Precision (PRC) (positive predictive value).

In this study, we evaluate the testing results by organizing a report which holds a given dialect d, which will be calculated by the dialect identification measurements defined as:

$$ACC_d = 100 * \frac{S_{correct}}{S_d} \quad (17)$$

Where $\mathbf{S_{correct}}$ is the amount of correctly identified test sequences of all test sequences $\mathbf{S_d}$ voiced in dialect **d**.

$$RCL_d = \frac{S_{TruePositives}}{S_{TruePositives} + S_{FalseNegatives}} \quad (18)$$

$$PRC_d = \frac{S_{TruePositives}}{S_{TruePositives} + S_{FalsePositives}} \quad (19)$$

In order to ensure the reliability of our results, we use a k-fold cross-validation technique with k = 10.

Figure 9, shows the Accuracy, Precision, and Recall of the system.

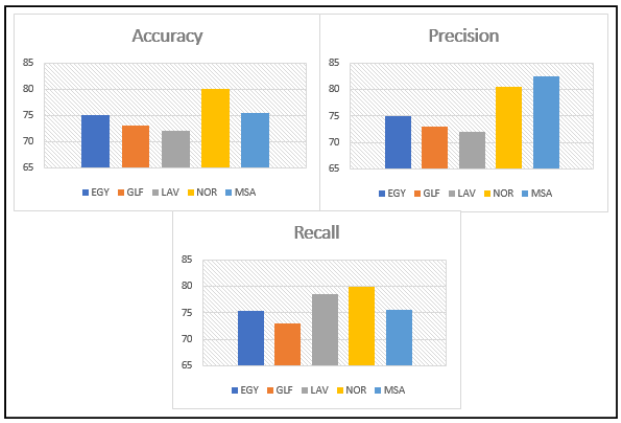


Figure 9. System Performances

11 Conclusions

We conclude in this paper an automatic identifying Arabic dialects system which had being proposed. The proposed system called " Building a System for Arabic Dialects Identification based on Speech Recognition using Hidden Markov Models (HMMs) ("ADIDSHMM") ", In the Identification Arabic Dialects system the difficult task of properly is to identify a various Arabic dialect and examining it. We applied the classification technique algorithm called Hidden Markov Models (HMM) to learn the results of the speech recognition based on the acoustic features vector from the feature extraction phase. The classification process in terms of HMM algorithm via using identification of the dialect of wav audio one of five dialects. The feature extraction process was implemented in which speech features are extracted for all the speech samples. All these features are given to the pattern trainer for training and are trained by HMM to create HMM model for each dialect. Afterward we will use the Viterbi algorithm of HMM to select the one with the maximum likelihood in which it recognized the dialect.

The dataset we used is widely known as ADI5. The ADI5 dataset became our experimental dataset that's created and collected by the MGB-3 challenge includes a multi-dialectal speech from various programs recorded from the Al-Jazeera TV channel. It includes also audio files in Egyptian (EGY), Levantine (LEV), Gulf (GLF), North African (NOR), and Modern Standard Arabic (MSA).

Finally, in our experimental results, we illustrated the overall system performance via four indices: overall accuracy, average precision and average recall for the five dialects.

References:

[1] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic dialect identification using phonotactic modeling," in Proceedings of the eacl 2009 workshop on computational approaches to semitic languages, 2009, pp. 53-61.

[2] A. M. J. E. J. o. E. Deshmukh and T. Research, "Comparison of hidden markov model and recurrent neural network in automatic speech recognition," vol. 5, no. 8, pp. 958-965, 2020.

[3] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," Columbia University, 2011.

[4] H. C. S. Bougrine and A. Abdelali, "Spoken arabic algerian dialect identification," in 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), 2018, pp. 1-6: IEEE.

[5] H. A. Elharati, M. Alshaari, V. Z. J. J. o. C. Këpuska, and Communications, "Arabic Speech Recognition System Based on MFCC and HMMs," vol. 8, no. 03, p. 28, 2020.

[6] M. Al-Ayyoub, M. K. Rihani, N. I. Dalgamoni, and N. A. Abdulla, "Spoken Arabic dialects identification: The case of Egyptian and Jordanian dialects," in 2014 5th International Conference on Information and Communication Systems (ICICS), 2014, pp. 1-6: IEEE.

[7] B. Mouaz, B. H. Abderrahim, and E. J. P. C. S. Abdelmajid, "Speech Recognition of Moroccan Dialect Using Hidden Markov Models," vol. 151, pp. 985-991, 2019.

[8] M. Belgacem, G. Antoniadis, and L. Besacier, "Automatic Identification of Arabic Dialects," in LREC, 2010.

[9] V. Blomqvist and D. Lidberg, "Swedish Dialect Classification using Artificial Neural Networks and Guassian Mixture Models," 2017.

[10] A. M. A. M. Ali, "Multi-dialect Arabic broadcast speech recognition," 2018.

[11] F. S. Alorifi, "Automatic identification of arabic dialects using hidden markov models," University of Pittsburgh, 2008.

[12] P. Heracleous, A. Yoneyama, K. Takai, and K. Yasuda, "Automatic Spoken Language Identification Using Emotional Speech," in International Conference on Human-Computer Interaction, 2020, pp. 650-654: Springer.

[13] N. Thiracitta, H. Gunawan, and G. Witjaksono, "The comparison of some hidden markov models for sign language recognition," in 2018 Indonesian Association for Pattern Recognition International Conference (INAPR), 2018, pp. 6-10: IEEE.

[14] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, "Dialect identification using Gaussian mixture models," in ODYSSEY04-The Speaker and Language Recognition Workshop, 2004.

[15] M. J. A. t. D. o. I. I. o. A. I. S. L. Heck, "Automatic Language Identification for Natural Speech Processing Systems," 2011.

[16] H. Z. Muhammad, M. Nasrun, C. Setianingsih, and M. A. Murti, "Speech recognition for English to Indonesian translator using hidden Markov model," in 2018 International Conference on Signals and Systems (ICSigSys), 2018, pp. 255-260: IEEE.

[17] P. Bell et al., "The MGB challenge: Evaluating multi-genre broadcast media recognition," in 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015, pp. 687-693: IEEE.

[18] A. Ali, S. Vogel, and S. Renals, "Speech recognition challenge in the wild: Arabic MGB-3," in 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2017, pp. 316-322: IEEE.

[19] Arabicspeech.org. 2022. MGB3_ADI – ArabicSpeech. [online] Available at: <https://arabicspeech.org/mgb3-adi/> [Accessed 1 March 2022].

[20] Mgb-challenge.org. 2022. MGB Challenge - MGB-3. [online] Available at: <http://www.mgb-challenge.org/MGB-3.html> [Accessed 6 March 2022].

[21] En-Naimani, Z. A. K. A. R. I. A. E., M. O. H. A. M. E. D. Lazaar, and M. O. H. A. M. E. D. Ettaouil. "Hybrid system of optimal self organizing maps and hidden Markov model for Arabic digits recognition." WSEAS Transactions on Systems 13.60 (2014): 606-616.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Zakaria Suliman Zubi, carried out the optimization as well as the statistics of the article.

Eman Jibril Idris, carried out the idea and implemented the algorithms with statistical used of Hidden Markov Model (HMM) in the ASR system as well as the code.

Sources of funding for research presented in a scientific article or scientific article itself

The research work was supported by Department of Computer Science, Faculty of Science, Sirte University, Sirte, Libya.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)