Spoken Arabic Dialect Identification Using Motif Discovery

Mohsen Moftah[*1], Mohamed Waleed Fakhr[**2], Salwa El Ramly[*3]

[*]Electronics &Communications Engineering Department, Faculty of Engineering, Ain Shams University
Cairo, Egypt
[1]mohsen.moftah@barmagyat.com
[3]salwahelramly@gmail.com

[**] College of Computing, Arab Academy for Science and Technology
Ahmed Ismail street, Heliopolis, Cairo, Egypt
[2]waleedf@aast.edu

Abstract: ***In traditional Dialect Identification (DID) approaches, regardless of the level and type of features used for identification, they use either predefined references such as phones, phonemes, or even acoustic sounds that characterize a language/dialect, or involve some sort of transcription of the input data. The transcription may be manual or automatic using tools such as ASRs, Tokenizers, or Phone Recognizers. In this paper, we introduce a new approach based on analyzing the speech signal directly and extracting the features that characterize the dialect without any predefined references and without any sort of transcription. The main idea is that we find the repeated sequences (motifs) of the dialect by treating the speech signal as a times series, so we can apply motif discovery techniques to extract the repeated sequences directly from the speech signal. For motif extraction, we adopted an extremely fast parameter-free Self-Join motif discovery algorithm called Scalable Time series Ordered-search Matrix Profile (STOMP). We implemented the new approach in two stages; in the first we built a base line system in which we extracted 12 Mel Frequency Cepstral Coefficients (MFCC) from each motif, in the second stage we built an improved system using 39 coefficients by adding 13 Delta coefficients, 13 Delta-Delta coefficients, and 1 Log Energy coefficient. In both systems, we used Gaussian Mixture Model-Universal Background Model (GMM-UBM) as a classifier. We applied our new approach on three different motif lengths 500ms, 1000ms, and 1500ms using 1gmm component up to 2048gmm components. We downloaded the data set from Qatar-Computing-Research- Institute domain. We carried out our experiments on different Arabic dialects: the Egyptian (EGY), Gulf (GLF), Levantine (LEV), and North African (NOR).The base line results were very competitive with the traditional, more sophisticated approaches, while the improved system showed very good result. The improvement was so significant that we can consider the new approach as competitive, simple, and dialect-independent approach.***

Key words: ***motif discovery, dialect identification, language identification, GMM-UBM, time series***

1 INTRODUCTION

The main Arabic dialects can be classified as: Egyptian, Gulf, Levantine, and North Africa. Automatic Dialect Identification (DID) is a special case of the more general task which is Automatic Language Identification (LID). LID became a mature technology and has various applications [1]. An Arabic DID system is required to automatically identify the dialect of the input speech; this is a challenging task since there are no solid boundaries between different Arabic dialects. As mentioned above, DID is a special case of LID, therefore, we can apply the same techniques used in LID to establish an Arabic DID system. Most LID systems, and therefore DID systems, operate in two phases, a training phase and recognition phase. In training phase, the system is trained using examples of every target dialect. This training data can be as simple as the digitized speech utterances mapped to the corresponding spoken language. More sophisticated system may require more data such as phonetic transcription in a form of sequence of symbols of the spoken sounds, and an orthographic transcription of the spoken words. From the training speech, fundamental characteristics of each language are analyzed to produce language-dependent models. The second phase, recognition, makes use of the language-dependent models produced in the training phase to identify new unknown utterances [2]. Based on the type of dialect features extraction and modeling, DID approaches can be divided into two main classes, a high level lexical and phonetic features approach such as Phone Recognition followed by Language Modeling (PRLM) and Parallel Phone Recognition followed by Language Modeling (PPRLM), and low level acoustic features concerned with spectral characteristics of speech such as Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) as acoustic front end and Gaussian Mixture Model (GMM), Universal Background Model-Gaussian Mixture Model (UBM-GMM) as acoustic backend [3] [4]. In this paper we are presenting a new approach based on the acoustic features of the spoken dialect. This approach is based on first discovering the repeated sequences/patterns i.e. motifs, of the speech signal directly, and then extract the MFCC features of the motifs. To examine the new approach we selected the well-known UBM-GMM method for modeling and classification. The reset of this paper is organized as follows: section 2 will present a brief description of the most popular DID/LID approaches; section 3 will discuss the motif discovery approach. Section 4 will be dedicated to explain our proposed approach; section 5 will show the experiments results, while the last section 6 will be a conclusion and future work.

2 DID/LID APPROACHES

A. High Level Lexical and Phonetic Features Approach

1) PPRLM Approach: In Phone Recognition followed by Language Modeling (PRLM) approach, a phone recognizer is used to tokenize the training dataset of the target dialects to produce phone sequences. The phone sequences are used to train a statistical language model to generate phonotactic language model for the dialects in question. These phonotactic language models are used to compute the dialect likelihood for the unknown utterances [5][6][7].

*2) PPRLM Approach:*In Parallel Phone Recognition followed by Language Modeling (PPRLM), phonotactic statistics of a language are extracted using multiple phone recognizers. Every phone recognizer is trained on different languages to capture acoustic characteristics of each language. The recognizers are combined to form a parallel recognizer PPR to characterize the spoken language [4].

B. Acoustic Approach

The implementation of the acoustic approach is comprised of two phases, a feature extraction phase, followed by a classification phase [8] [3][9]. The most popular features used in this phase are:

1) Mel Frequency Cepstral Coefficients (MFCC): Frequency domain features are characterized by their robustness and reliability to variations of speakers and recording conditions.

2) Shifted Delta Cepstral coefficients (SDC): SDC is a stack of delta spectra computed across multiple speech frames. SDC is an efficient method to model temporal features of languages, which is very important in language identification. It is based on Delta-Delta coefficients extracted in MFCC with the capability of modeling temporal features over multiple frames to accommodate the phonemes length which is at least 50ms.

3) Relative Spectra Filtering (RASTA): Filtering of cepstral trajectories is used to remove slowly varying, linear channel effects from raw feature vectors.

The second phase in acoustic based approach is the classification phase. The following are the most popular classifiers applied in DID/LID. These classifiers are used successfully in speaker recognition:

1) GMM-UBM: GMM is extensively used in speaker recognition. In GMM-UBM approach the first step is to create a Universal Background Model (UBM) by training the GMM with a large amount of data using iterative Expectation Maximization (EM) algorithm to maximize the likelihood of the GMM. To create a speaker specific model, GMM parameters; the mixture weight, mean vector, and covariance matrix are adapted to specific speaker using Maximum a Posteriori (MAP) scheme. During the adaptation process, parameters for the Gaussian mixtures which bear a high probabilistic resemblance to the language specific training data will tend towards the parameters of that training data whereas the parameters of the Gaussian mixtures bearing little resemblance to the language specific data will remain fairly close to their original UBM values[4].

2) GMM-SVM: Support Vector Machines (SVM) became as popular as GMM.It uses a linear kernel in a supervector space for rapid computation of language distance. The kernel computes the distance between two supervectors one represents the GMM model and the other represents the target language [4].

3) i-vector: Dehak [10] developed a new classifier by finding a low dimensional subspace from the GMM super-vector space based on Joint Factor Analysis (JFA) as a feature extractor. The low dimensional subspace is called total variability space since it includes both speaker and channel variations. The dimensionality of the low-dimensional space is reduced using Linear Discriminant Analysis (LDA). The vectors in the low-dimensional space are called i-vectors, which are of small size compared with those in GMM super-vector to reduce execution time while keeping the recognition rate acceptable.

3 MOTIF DISCOVERY

Motif discovery has been applied in many applications such as summarizing and visualizing massive time series databases, in addition to various data mining tasks, including the discovery of association rules. Figure 1, shows an example of motifs discovered in a time series [11].

One common approach of Motif discovery applies similarity search approach which depends on similarity threshold, a value that is difficult to determine [12]. Another approach called All-Pairs-Similarity-Search, Similarity Join, or Self Join approach. A brief explanation of this approach is introduced in the following paragraphs showing how to apply it on speech signals.

In speech, a speech audio signal can be easily considered a time series. As will be explained, a time series is defined as a sequence of real-valued numbers, in digital audio these valued numbers are the audio sample values. A motif in a speech time series can represent repeated words or sub words. The following is a background on motif discovery in speech and a brief explanation of the self-join algorithm used as the base of our approach in Arabic DID [13].

Definition 1: A time series T is a sequence of real-valued numbers t_i: $T = t_1, t_2, ..., t_n$ where n is the length of T.

A local region of time series is called a *subsequence*:

Definition 2: A *subsequence* $T_{i,m}$ of a time series T is a continuous subset of the values from T of length m starting from position i. Formally, $T_{i,m} = t_i, t_{i+1}, ..., t_{i+m-1}$, where $1 \leq i \leq n\text{-}m\text{+}1$.

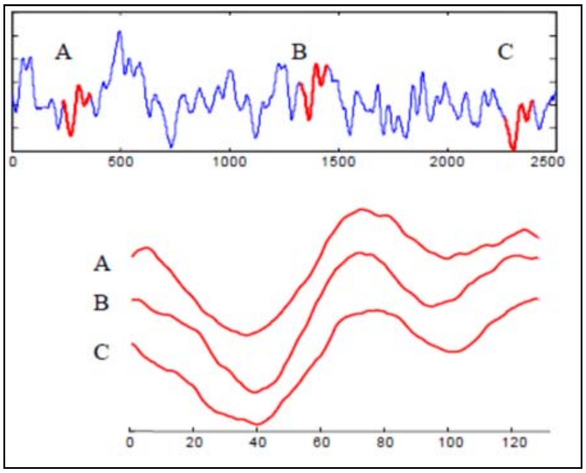


Figure 1: An astronomical time series (above) contains 3 near identical subsequences. A “zoom-in” (below) reveals just how similar to each other the 3 subsequences are.

If we compute the distance of a subsequence to *all* subsequences in the same time series; we come up with a *distance profile*:

Definition 3: A *distance profile* $D_i$of time series T is a vector of the Euclidean distances between a given query subsequence $T_{i,m}$ and each subsequence in time series T. Formally, $D_i = [d_{i,1}, d_{i,2}, ..., d_{i,n-m+1}]$, where $d_{i,j}(1 \leq i, j \leq n\text{-}m\text{+}1)$ is the distance between $T_{i,m}$ and $T_{j,m}$ where the distance is measured by Euclidean distance between z-normalized subsequences. Equation 1 shows how to calculate distance between two z-normalized subsequences. A z-normalized subsequence has a mean value of zero and standard deviation value of one [14].

$$d_{i,j} = \sqrt{2m - \left(\frac{QT_{i,j} - m\mu_i m\mu_j}{m\sigma_i\sigma_j}\right)} \qquad (1)$$

where m is the subsequence length, μ_i is the mean of $T_{i,m}$, μ_j is the mean of $T_{j,\mathrm{m}}$, σ_i is the standard deviation of $T_{i,m}$, and $\sigma_j$is the standard deviation of $T_{j,\mathrm{m}}$, $QT_{i,j}$is the dot product of $T_{i,m}$ and $T_{j,\mathrm{m}}$.

The mean can be calculated by [14]

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad (2)$$

and the standard deviation can be calculated by [14]

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m} x_i^2 \ - \mu^2 \qquad (3)$$

We use a vector called matrix profile to represent the distances between all subsequences and their nearest neighbors:

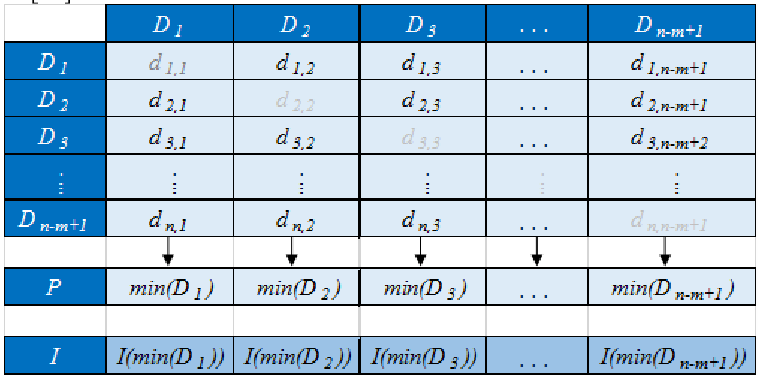
Definition 4: A *matrix profile P* of time series *T* is a vector of the Euclidean distances between each subsequence $T_{i,m}$ and its nearest neighbor (closest match) in time series *T*.
Formally, P = [min(D_1), min(D_2),..., min($_{Dn\text{-}m+1}$)], where $D_i(1 \leq i \leq n\text{-}m+1)$ is the distance profile $D_i$of time series *T* (Figure 2) .

The i^{th} element in the matrix profile *P* tells us the Euclidean distance from subsequence $T_{i,m}$to its nearest neighbor in time series *T*. However, it does not tell us *where* that neighbor is located. This information is recorded in a companion data structure called the *matrix profile index*.

Definition 5: A *matrix profile index I* of time series *T* is a vector of integers: I= [I_1, I_2, … $I_{n\text{-}m+1}$], where I_i=*j* if

$d_{i,j}$= min(D_i).

We can use the *matrix profile P* and the *distance profile D* to extend the notion of motifs to sets of subsequences that are very similar to each other [15]



	D_1	D_2	D_3	. . .	$D_{n\text{-}m+1}$
D_1	$d_{1,1}$	$d_{1,2}$	$d_{1,3}$	. . .	$d_{1,n\text{-}m+1}$
D_2	$d_{2,1}$	$d_{2,2}$	$d_{2,3}$	. . .	$d_{2,n\text{-}m+1}$
D_3	$d_{3,1}$	$d_{3,2}$	$d_{3,3}$	. . .	$d_{3,n\text{-}m+2}$
⋮	⋮	⋮	⋮	⋮	⋮
$D_{n\text{-}m+1}$	$d_{n,1}$	$d_{n,2}$	$d_{n,3}$	. . .	$d_{n,n\text{-}m+1}$
	↓	↓	↓	↓	↓
P	min(D_1)	min(D_2)	min(D_3)	. . .	min($D_{n\text{-}m+1}$)
I	I(min(D_1))	I(min(D_2))	I(min(D_3))	. . .	I(min($D_{n\text{-}m+1}$))

Figure 2: An illustration of the relationship between the Distance Profile D, the Matrix Profile P, and the Matrix Index Profile I along with the full distance matrix

Definition 6: The *Range motif* with range *r* is the maximal set of subsequences that have the property that the maximum distance between them is less than 2*r*. More formally *S* is a *range motif* with range *r* iff $\forall T_x, T_y \in S$, dist(T_x, T_y) ≤ 2*r* and $\forall T_d \in D\text{-}S$ dist(T_d, T_y) > 2*r.*

4 Proposed Approach

Our proposed approach introduces a new technique for LID/DID identification. Referring to Definition 1 in Section 3, a digital speech recording is typically a time series; accordingly, all techniquesapplied to time series can by directly applied to a digital speech signal. The idea is to extract language/dialect characteristics by extracting the repeated sequences of the speech signal without the need to transform the signal to text or symbols. These sequences (motifs) do not resemble any predefined entity such as words or phone etc. They are abstract repeated sequences, if uniquely repeated in speech signals of a dialect, can be considered a unique characteristic of the given dialect. The main issue in this approach is selecting the motif length, which is still a subject of trial and error. This is because selecting very short motif length will result in a very large number of non-informative, mostly non-lexical motifs such as breath intakes. On the other hand selecting a very long motif will not yield any motifs. Therefore, we applied our approach experiments using three different motif lengths, 500ms, 1000ms, and 1500ms for the base line system. We applied the improved system on the motif length that gave the best results in the base line implementation.

We selected the GMM-UBM approach to carry out our experiments because it is a well-established approach in the field of speech processing, in addition to its fast and simple implementation. In our new approach, we first extract motifs from speech utterances, then the MFCC features are extracted from the motifs, model training, and classification are carried out in the same way as in traditional GMM-UBM. Our work is based on the STOMP algorithm [13], which has many advantages over other algorithms. The time complexity of STOMP is $O(n^2)$ much better compared with the $O(n^4)$ proposed by Patel et al [16], and the $O(mn^2)$ of the classical brute force algorithms, where *n* is the length of the time series and *m* is the length of the subsequence (motif). Chiu et al [16], introduced a method based on the random projection algorithm which transforms the data to symbolic sequences using the Symbolic Aggregate Approximation (SAX) method. The time complexity of this method is quadratic and it depends on the chosen SAX word length. Another advantage of STOMP is that it is a self-join algorithm; hence, it does not need a similarity threshold. In addition, STOMP is a parameter-free algorithm its only input is the motif length *m*[10], [13]. The implementation of the proposed approach was carried out as follows: the training corpus was downloaded from Qatar-Computing-Research-

Institute domain. The test data is a high quality audio from Aljazeera channel covering the period of July 2014 until January 2015. The recordings were domain independent that is free talks of different guests in different domains, which added another challenge. We applied our new approach to the most common Arabic dialects; the Egyptian (EGY), Gulf (GLF), Levantine (LEV), and North African (NOR). We used Microsoft MSR Identity Toolbox v1.0: A MATLAB Toolbox for training and scoring GMM-UBM system and VOICEBOX: Speech Processing Toolbox for MATLAB for computing the MFCC coefficients as features. Figure3 shows a block diagram of the proposed approach. The following is a description of each phase of the proposed approach.

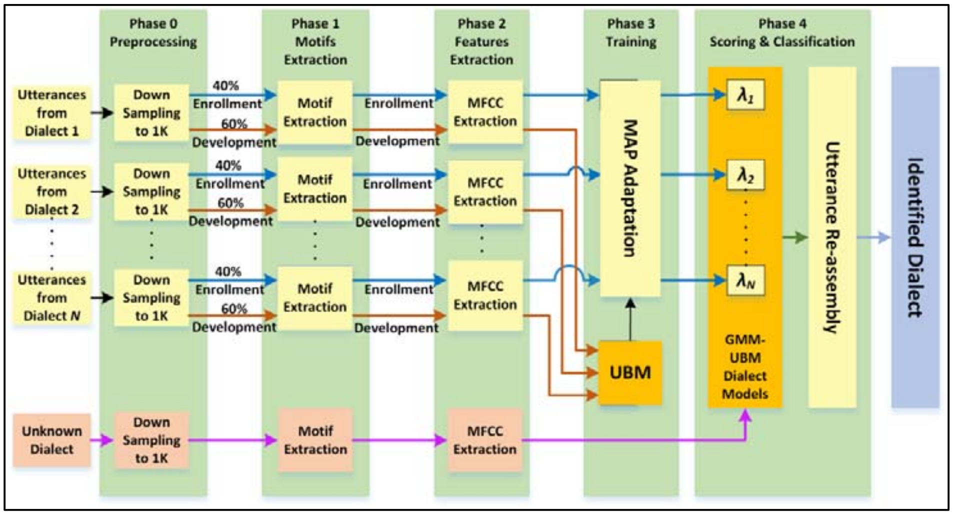


Figure 3: Proposed Approach Block Diagram

A. Preprocessing phase

For the implementation, we selected to use training and test utterances of up to 15 seconds long. The first step in this phase is down sampling of both training and testing utterances to 1K samples/Sec. The purpose of this operation is to reduce the dimensionality of the speech signal, while preserving its acoustic features, to reduce the needed computational resources.

The second step is motif extraction. TABLE I shows "Motif Discovery", our motif discovery algorithm. The algorithm computes the distance of each subsequence $T_{i,m}$ from all subsequences in the time series T and creates the Distance Matrix as described in Section 3. For a subsequence $T_{i,m}$ to be a motif, two conditions should be fulfilled (lines 9-13 of the algorithm shown in TABLE I):

1) The number of motif neighbors should be greater than one.
2) The difference between the index of the current candidate motif $T_{i,m}$ and the index of the previous motif should be greater than the length m of the subsequence. $I(T_{i,m}) - I(T_{i-1,m}) > m$. This will guarantee that there is no overlap between motifs.

In this step, we created three sets of motifs for each dialect. The first set with motif length of 500ms, the second with motif length of 1000ms, and the third set with motif length of 1500ms.

We split the training utterances into two parts; the first part is 60% of the utterances count of each dialect and is used for development. The remaining 40% of the training utterances are used as enrollment data. For testing, 100 utterances from each dialect are selected for classification and testing. TABLE II and TABLE III show the statistics of training and testing data respectively in terms of duration, utterance count, and motif counts for different motif lengths 500ms, 1000ms, and 1500ms.

TABLE I

MOTIF DISCOVERY ALGORITHM

```
Procedure Motif Discovery(T,m,R)
Input: Time Series $T$, Subsequence length $m$, Radius $R$
Output: Motifs Indices List
$n \leftarrow Length(T)$, $l \leftarrow n-m+1$
Compute mean $\mu$ and Standard deviation $\sigma$ from (2) and (3)
Compute the dot product (QT) between every subsequence and all subsequences in $T$
For $i=1$ to $l$
    Compute the distance between subsequence $i$ of length $m$ and every subsequence in $T$ using (1),
    Update the Distance Matrix $Di$
    Compute the minimum distance $d_{min}=min(D_i)$
    Find the neighbors vector $M_i$. neighbors are subsequences in $D_i$ whose distances from subsequence $T_i$<= $d_{min}$* $R$(Definition 6)
    If neighbors count >1
        If $I(T_{i,m})$ - $I(T_{i-1,m})$ >m
            Update Motifs List
        End if
    End if
End for
```

TABLE II
TRAINING DATA STATISTICS

Dialect	Training Data				
	Utterances		Motifs Count by Length		
	Hours	Count	500ms	1000ms	1500ms
EGY	7.37	2794	5460	4401	3802
GLF	6.81	2577	4949	4096	3478
LEV	7.17	3060	5455	4499	3755
NOR	7.61	2913	5767	4636	3961

TABLE III
TESTING DATA STATISTICS

Dialect	Testing Data				
	Utterances		Motifs Count by Length		
	Hours	Count	500ms	1000ms	1500ms
EGY	0.281	100	212	185	152
GLF	0.275	100	206	177	149
LEV	0.281	100	206	168	139
NOR	0.284	100	202	176	154

TABLE IV
60% OF TRAINING DATA USED AS DEVELOPMENT DATA FOR UBM CREATION

Dialect	Training Data			
	60% for UBM Generation			
	Utterance Count	Motif Count		
		500ms	1000ms	1500ms
EGY	1676	3263	2669	2314
GLF	1546	2967	2411	2044
LEV	1836	3563	2695	2276
NOR	1748	3458	2815	2399

TABLE IV and TABLE V show the statistics of splitting the training data into 60% UBM part and 40% MAP part respectively. The UBM part was used to create the UBM model from all dialects, then dialect specific GMM model is created by adapting the UBM model to each dialect model using features in the MAP portion of training dataset.

B. Features Extraction Phase

In this phase, the MFCC features are extracted from the motifs for both training and testing data. For the base line system, we used MFCC coefficients with the following configuration: 12 coefficients, 20ms window width, 10ms window overlap, and 24 filter bank. For the improved system, we used the same configuration but we added 13 Delta coefficients, 13 Delta-Delta coefficients, and 1 Log Energy coefficient to have a total of 39 features coefficients vector.

C. Training Phase

Training the UBM model: The MFCC features of the development training motifs from both dialects are combined and used to train the UBM model through Expectation Maximization (EM) algorithm.

Training the Dialects models: Dialect specific GMM models were trained by adapting the UBM model to each dialect using the MFCC features of its enrollment data. The adaptation is done using Maximum A Posteriori (MAP) algorithm.

In our experiments, we created UBM and dialect specific GMMs with Gaussian Mixtures from 1gmm component up to 2048gmm components.

TABLE V

40% OF TRAINING UTTERANCES USED AS ENROLLMENT DATA FOR DIALECT SPECIFIC GMM MODELS CREATION

Dialect	Training Data			
	40% For Dialect Specific GMM			
	Utterance Count	Motif Count		
		500ms	1000ms	1500ms
EGY	1118	2197	1732	1488
GLF	1031	1982	1685	1434
LEV	1224	1892	1804	1479
NOR	1165	2309	1821	1562

D. Classification Phase

In this phase, the test data is used to evaluate the system. The test data is MFCC files of the motifs extracted from the test utterances. The test utterances of each dialect are passed through the GMM models of all target dialects. The scores are computed for every motif as the log-likelihood ratio between the given dialects models and the UBM given the test observations using MSR Identity Toolbox. The utterance score is computed as the sum of its individual motifs scores. The classification of the utterance is determined according to its maximum score. For example, an EGY utterance is passed through all the GMM models, if the maximum score came from the EGY GMM model the utterance is considered correctly identified, otherwise it is considered wrongly identified. The accuracy *Acc* is calculated as follows:

$$Acc = \frac{no\ of\ correctly\ identified\ utterances}{total\ no\ of\ test\ utterances} \times 100 \tag{4}$$

5 EXPERIMENTS RESULTS

The results for the base line system are shown in TABLE VI, TABLE VII, and TABLE VIII. The results show that the best average accuracy is obtained using 1gmms for all motif lengths (results in red). The best of all is 45.75% with 1gmm using 500ms motif length. In all motif lengths, the GLF dialect has the best identification score, while NOR has the worst.

TABLE VI

BASE LINE: AVERAGE ACCURACY FOR 500MS MOTIFS FOR ALL DIALECTS AND ALL GMMS

GMM Models	Motif Length 500ms One Dialect All Models MFCC				
	Accuracy				Average Accuracy
	EGY	GLF	LEV	NOR	
1gmm	**46**	**51**	**50**	**36**	**45.75**
2gmm	44	49	41	39	43.25
4gmm	40	32	45	22	34.75
8gmm	44	40	42	40	41.5
16gmm	26	27	30	19	25.5
32gmm	26	32	32	26	29
64gmm	20	27	38	32	29.25
128gmm	25	28	37	30	30
256gmm	24	30	28	24	26.5
512gmm	24	29	33	26	28
1024gmm	28	19	40	20	26.75
2048gmm	23	22	32	34	27.75

TABLE VII

BASE LINE: AVERAGE ACCURACY FOR 1000MS MOTIFS FOR ALL DIALECTS AND ALL GMMS

GMM Models	Motif Length 1000ms One Dialect All Models MFCC				
	Accuracy				Average Accuracy
	EGY	GLF	LEV	NOR	
1gmm	**41**	**45**	**47**	**32**	**41.25**
2gmm	37	45	37	34	38.25
4gmm	37	38	43	27	36.25
8gmm	34	44	38	34	37.5
16gmm	27	29	31	25	28
32gmm	27	28	37	26	29.5
64gmm	25	31	30	20	26.5
128gmm	27	34	31	25	29.25
256gmm	23	30	24	33	27.5
512gmm	27	31	31	20	27.25
1024gmm	30	19	30	20	24.75
2048gmm	27	25	17	23	23

Table VIII

Base Line: Average Accuracy for 1500ms Motifs for all Dialects and all GMMs

GMM Models	Motif Length 1500ms One Dialect All Models MFCC				
	Accuracy				Average Accuracy
	EGY	GLF	LEV	NOR	
1gmm	**43**	**48**	**47**	**31**	**42.25**
2gmm	37	46	35	27	36.25
4gmm	44	43	42	26	38.75
8gmm	40	45	35	32	38
16gmm	24	33	18	20	23.75
32gmm	26	24	17	19	21.5
64gmm	20	33	32	26	27.75
128gmm	26	30	25	24	26.25
256gmm	30	23	32	19	26
512gmm	26	30	33	20	27.25
1024gmm	32	40	28	25	31.25
2048gmm	29	32	28	22	27.75

Table IX

Base Line: Confusion Matrix for 1gmm 500ms Motif Length

Confusion Matrix of best results (1gmm) 500ms					
	EGY	GLF	LEV	NOR	
EGY	**46**	12	6	36	100
GLF	15	**51**	2	32	100
LEV	14	6	**51**	29	100
NOR	15	**4**	45	**36**	**100**

For the best result (1gmm) using 500ms Motif length, we computed the Confusion Matrix as shown in Table IX. The confusion matrix shows that the discrimination between any two dialects is very good. NOR, caused the highest confusion with all other dialects. The results of the improved system applied to the 500ms motif length is shown in Table X.

Table X

Improved System: AVERAGE ACCURACY FOR 500MS MOTIFS FOR ALL DIALECTS AND ALL GMMS

GMM Models	Motif Length 500ms One Dialect All Models Delta+Delta-Delta+Energy				
	Accuracy				Average Accuracy
	EGY	GLF	LEV	NOR	
1gmm	7	10	8	2	6.75
2gmm	11	12	8	9	10
4gmm	**62**	**66**	**68**	**55**	**62.75**
8gmm	36	38	37	33	36
16gmm	32	32	32	31	31.75
32gmm	37	40	38	25	35
64gmm	25	40	31	29	31.25
128gmm	26	26	37	30	29.75
256gmm	20	31	36	21	27
512gmm	19	32	29	25	26.25
1024gmm	32	27	30	26	28.75
2048gmm	28	27	34	32	30.25

The results of the improved system show improvement in all dialects resulting in an average accuracy of 62.75% for 4gmm model. TABLE XI shows the improvement achieved by the improved system over the base line system for every individual dialect and for the average accuracy.

TABLE XI

IMPROVEMENT ACHIEVED OVER THE BASE LINE

	EGY	GLF	LEV	NOR	Average Accuracy
Base Line	46	51	50	36	45.75
Improved	62	66	68	55	62.75
Improvement %	34.78	29.41	36.00	52.78	37.16

TABLE XII

IMPROVED SYSTEM: CONFUSION MATRIX FOR 4GMM 500MS MOTIF LENGTH

Confusion Matrix for (4gmm) 500ms Improved system					
	EGY	GLF	LEV	NOR	
EGY	**62**	10	12	16	100
GLF	11	**66**	8	15	100
LEV	**9**	**4**	**68**	**19**	**100**
NOR	7	**10**	28	**55**	100

TABLE XII shows the confusion matrix for 4gmm model applied to 500ms motif length with improved system. NOR still causes the highest confusion with other dialects; however, the percentage decreased significantly compared with the base line system.

Comparing the results with known works used the same data set is [3] and [17].The results in [3] achieved a total accuracy of 60.2%. The results were achieved using sophisticated features extraction approach which involved human intervention, in addition of fusing the scores of a senone-based system and the SVM based i-vector system. Moreover, the work is based on phonetic and lexical features obtained from a speech recognizer system not direct to the speech signal. In [16] the system used the transcribed version of the same data set i.e. the system is text based. The average accuracy achieved is 52%. Compared to our base line system, we applied a very straightforward approach using a well-known GMM-UBM method and a simple feature extraction method using 12 MFCC coefficients to achieve a competitive accuracy compared to traditional more sophisticated techniques. Our improved system outperformed [3] where we achieved 62.75% against their 60.2%. Moreover, our approach works on the speech signal without the need to transform it to text or sequence of symbols, which nominates it as a dialect/language independent approach.

6 CONCLUSIONS

The overall results shows that short motifs 500ms give the best results. The results and comparison with other approaches indicates that our new approach for dialect identification is a promising new technique. The main advantage of this approach is its simple implementation in addition to being dialect independent, it does not need any prior experience of the target dialect/language, in addition no need to neither human intervention for labeling nor transforming it into text or symbol sequences.

REFERENCES

[1] Santhi.S , Raja Sekar, "An Automatic Language Identification Using Audio Features", *International Journal of Emerging Technology and Advanced Engineering*, Volume 3, Special Issue 1, pp 358-364, January 2013

[2] Marc A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", IEEE *Transactions On Speech And Audio Processing,* VOL. 4, NO. 1, January 1996.

[3] Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell5, Steve Renals, "Automatic Dialect Detection in Arabic Broadcast Speech" in *INTERSPEECH,* pp 2943-2938, San Francisco, USA, September 8–12, 2016,

[4] Fadi Biadsy, Julia Hirschberg, and Nizar Habash, "Spoken Arabic Dialect Identification Using Phonotactic Modeling" in *Workshop on Computational Approaches to Semitic Languages*, pp 53–61, Athens, Greece, 31 March, 2009.

[5] Eliathamby Ambikairajah, Haizhou Li, Liang Wang, Bo Yin, and Vidhyasaharan Sethu, "Language Identification A Tutorial", *IEEE Circuits And Systems Magazine* Second Quarter 2011, pp 82-108.

[6] Soumia Bougrine, Hadda Cherroun, Djelloul Ziadi, "Prosody-based Spoken Algerian Arabic Dialect Identification" in *International Conference on Natural Language and Speech Processing*, Algiers, Algeria 2015.

[7] M. Akbacak, D. Vergyri, A. Stolcke, N. Scheffer and A. Mandal, "Effective Arabic dialect classification using diverse phonotactic models", in Proc. Interspeech, 2011, pp. 141-144.

[8] Kshirod Sarmah and Utpal Bhattacharjee, "GMM based Language Identification using MFCC and SDC Features", *International Journal of Computer Applications (0975 – 8887)* Volume 85 – No 5, pp 36-42, January 2014
[9] Rania R Ziedan · Michael Nasief, · Abdulwahab K. Alsammak · Mona F. M. Mursi · Adel S. Elmaghraby , "A Unified Approach for Arabic Language Dialect Detection". *29th International Conference on Computer Applications in Industry and Engineering (CAINE 2016)*, pp 165-170, Denver, USA, September 26-28, 2016
[10] Najim Dehak, Reda Dehak, Patrick Kenny, Niko Br¨ummer, Pierre Ouellet, and Pierre Dumouchel. "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification". *In Interspeech*, volume 9, pages 1559– 1562, 2009.Place?, date?
[11] Jessica Lin, Eamonn Keogh, Stefano Lonardi, Pranav Patel, "Finding Motifs in Time Series" *Proceedings of the Second Workshop on Temporal Data Mining at the 8th SIGKDD, pp 53-68*, Edmonton, Alberta, Canada — July 23 - 26, 2002
[12] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh, "Matrix Profile I: All Pairs Similarity Joins for Time Series:A Unifying View that Includes Motifs, Discords and Shapelets". *IEEE International Conference on Data Mining IEEE ICDM 2016*, Pp 1317-1326, Barcelona, Spain 2016.
[13]- Yan Zhu, Zachary Zimmerman, Nader Shakibay Senobari, Chin-Chia Michael Yeh, Gareth Funning, Abdullah Mueen, Philip Brisk, and Eamonn Keogh, "Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins', *IEEE International Conference on Data Mining IEEE ICDM 2016*, pp 739-748, Barcelona, Spain 2016.
[14] Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, Eamonn Keogh, "Searching and Mining Trillions of Time Series Subsequences under Dynamic Time Warping" *Conference: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining,* pp 262-270, Beijing, China — August 12 - 16, 2012
[15] Abdullah Mueen , Eamonn Keogh , Qiang Zhu , Sydney Cash, "Exact Discovery of Time Series Motifs", *Conference: Proceedings of the SIAM International Conference on Data Mining, SDM 2009*, pp 473-484 Sparks, Nevada, USA, April 30 - May 2, 2009
[16] S. Torkamani and V. Lohweg, "Survey on time series motif discovery," WIREs Data Mining and Knowledge Discovery, vol. 7, pp. 1-8, 2017.
[17] S. Malmasi and M. Zampieri, "Arabic Dialect Identification in Speech Transcripts," in Proceedings of the Third Workshop on NLP for Similar, Osaka, Japan, 2016.

BIOGRAPHY

Mohsen Moftah holds a M.Sc. degree in Communications and Electronics Engineering, now perusing PhD in the same discipline. In the academic side, his research interest is Arabic language engineering, and participated in local as well as international conferences covering this field. He also has papers published in those conferences. In the industrial side, ha has more than 30 years of experience in the IT arena. His experience covers Application Development, Technical Support, and Operations Management, Projects Delivery involving multi-party in addition to deep exposure to many technologies such as ERP, Hospital Management Systems, and Content Management Systems. And other technologies such as wired/wireless networking, Access Control and Time Attendance using RFID technology. In addition to security applications, using Video Analytics based surveillance.

Prof. Fakhr finished his Ph.D. at the University of Waterloo, Canada, 1993, in the field of neural networks and machine learning; he then joined the speech research lab at NORTEL, Montreal, Canada, for 5 years where he was a researcher investigating and implementing different speech processing, speech recognition, language modeling, and statistical error analysis techniques and has 2 patents with NORTEL. Since 1999 he has been a professor with the Arab academy for science and technology (Cairo, Egypt) with 3 years sabbatical at the University of Bahrain. He has been doing research in the areas of Multimedia processing, Arabic Language Processing, Printed and handwritten character recognition, Statistical machine translation, Language modeling, Neural networks and Sparse coding.

Prof. Salwa Elramly, BSc. Degree 1967, MSc. Degree 1972 from Faculty of Engineering, Ain Shams University, Egypt &PhD degree 1976 from Nancy University, France. She is now professor Emeritus with the Electronics and Communications Engineering Department, Faculty of Engineering, Ain Shams University; where she was the Head of the Department (2004-2006). Her research field of interest is Wireless Communication Systems and Signal Processing, Language Engineering, Coding, Encryption, and Radars. She is a Senior Member of IEEE and Signal Processing Chapter chair in Egypt. She was awarded Ain Shams Award of Appreciation in Engineering Sciences (2010), Award of Excellence from the Society of Communications Engineers (2009) &Award of Excellence from the Egyptian Society of Language Engineering.

التعرف على اللهجات العربية المنطوقة بإستخدام العناصر المتكررة

[1]*محسن مفتاح، [2]**محمد وليد فخر، [3]*سلوى الرملي
*قسم هندسة الالكترونيات و الاتصالات، كلية الهندسة جامعة عين شمس. القاهرة، مصر
**كلية الحاسبات، الاكاديمية العربية للعلوم و التكنولوجيا و النقل البحري. القاهرة، مصر
[1]mohsen.moftah@barmagyat.com
[2]waleedf@aast.edu
[3]salwahelramly@gmail.com

الخلاصة - في الطرق التقليدية للتعرف على اللهجة/اللغة يتم الرجوع الى خصائص محددة مسبقا تميز اللهجة/اللغة مثل الفونيمات أو صوتيات اللغة أو حتى الخصائص الصوتية للغة فإن عملية التعرف تتضمن نوعا من تحويل الصوت الى نص سواء الى كلمات او تمثيل الصوت برموز صوتية. و تتم هذه العملية اما يدويا أو باستخدام برمجيات مثل التعرف الآلي على الكلام ASR او التعرف الآلي على الأصوات Phone Recognizer. في هذا البحث نقدم أسلوب جديد يقوم على التحليل المباشر للإشارات الصوتية للكلام و استخراج العناصر المتكررةMotifs لكل لهجة دون الحاجة الى تحويلها الى شكل نصي. الفكرة الأساسية هي اعتبار الاشارة الصوتية للكلام متسلسلة زمنية و عليه يمكن تطبيق جميع التقنيات الخاصة باستخلاص العناصر المتكررة من المتسلسلات الزمنية مباشرة على الإشارة الصوتية. لاستخراج العناصر المكررة تم استخدام خوارزم متناهي السرعة يسمى STOMP. وقد تم تطبيق الطريقة الجديدة على مرحلتين. في الأولى تم بناء نموذج مرجعي بإستخدام 12 MFCC. و في المرحلة الثانية تم بناء النظام المطور بإستخدام 39 معامل موزعة كالتالي: 12 MFCC, 13 Delta, 13 Delta-Delta, 1 Log Energy. و لكلا النموذجين تم بناء نموذج GMM-UBM لأطوال مختلفة للعناصر المتكررة هي ms500 و ms1000 و ms1500 في حالة النموذج المرجعي . بالنسبة للنموذج المطور فقد تم بناؤه لطول العنصر الذي حقق أعلى النتائج. و قد تم بناء GMM-UBM لنماذج تحتوي على gmm1 و حتى gmm2048. و قد تمت تجربة الأسلوب الجديد على التسجيلات الصوتية المتاحة على موقع معهد قطر لبحوث الحوسبة لكل من اللهجة المصرية (EGY)و الشامية (LEV)و الخليجية (GLF) و المغاربية (NOR). وقد جاءت نتائج النموذج المرجعي منافسة لتقنيات أخرى اكثر تعقيدا. و حقق النموذج المطور نتائج تفوق الطرق التقليدية مما يؤهل الطريقة المطروحة لتكون إحدى الطرق المنافسة في مجال التعرف على اللغات و اللهجات.