International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:6, No:11, 2012

Comparison of Parameterization Methods in Recognizing Spoken Arabic Digits

Ali Ganoun

***Abstract*—**This paper proposes evaluation of sound parameterization methods in recognizing some spoken Arabic words, namely digits from zero to nine. Each isolated spoken word is represented by a single template based on a specific recognition feature, and the recognition is based on the Euclidean distance from those templates.

The performance analysis of recognition is based on four parameterization features: the Burg Spectrum Analysis, the Walsh Spectrum Analysis, the Thomson Multitaper Spectrum Analysis and the Mel Frequency Cepstral Coefficients (MFCC) features. The main aim of this paper was to compare, analyze, and discuss the outcomes of spoken Arabic digits recognition systems based on the selected recognition features. The results acquired confirm that the use of MFCC features is a very promising method in recognizing Spoken Arabic digits.

***Keywords*—**Speech Recognition; Spectrum Analysis; Burg Spectrum; Walsh Spectrum Analysis; Thomson Multitaper Spectrum; MFCC.

I. Introduction

AUTOMATIC Speech Recognition (ASR) is a technology that allows an electronic platform such as a smart phone or a computer to identify spoken words. Automatic recognition of spoken digits is one of the challenging tasks in the field of ASR. There are many applications where recognition of spoken digits systems are used; such as recognizing telephone numbers, telephone dialing using speech, and automatic directory to retrieve or send information, etc. [1].

However, the automatic recognition of spoken digits process is not straightforward because it involves a number of problems. Such as: different duration of the same word sound, the redundancy in the speech signal that makes discriminating between spoken digits difficult, the presence of temporal and frequency variability in pronunciation of spoken digits and signal degradation due to different types of noise found with the signal.

The performance of recognition systems is language dependent. Therefore, conclusions drown as a result of evaluating recognition techniques based on other languages may not be applied to Arabic language [2]. The main aim of this paper is to compare, analyze, and evaluate the accuracy of spoken Arabic digit recognition system of using four parameterization features used to represent sound signals: the Burg Spectrum Analysis, the Walsh Spectrum Analysis, the Thomson Multitaper Spectrum Analysis and the Mel Frequency Cepstral Coefficients (MFCC) features [4] - [7]. The performance evaluation asses are both the overall system performance and the individual digit accuracy. Compared to the work in [3], this paper performs a more comparisons and analysis that based on other features and larger databases with a higher number of speakers.

The rest of the paper is organized as follows: section II presents a description of the database used by the system; section III presents a brief description of feature extraction processes. Section IV discusses the experimental setup. Section V presents the results of comparisons obtained as a result of this work. The paper concludes with section VI.

A. Ganoun is with the Electrical and Electronic Engineering Department, Faculty of Engineering, University of Tripoli, Libya (phone: +218-214625559; fax: +218-214625558; e-mail: ali.ganoun@ee.edu.ly).

II. Database Preparation

In order to evaluate the selected recognition techniques a database of the sounds of the Arabic digits (0 to 9) was created. Three male and three female Arabic native speakers were asked to utter all digits; each time the speech is recorded in a single file which is approximately 12 second long. Each file was played back to ensure that the entire digits were included in the recorded file. This process was repeated 13 times for each user. In general, 78 speech files were created; each file contains all the Arabic digits.

Every speech file contains both speech signals and non-speech signals. Then, each file was analyzed by a detection program in order to locate and segment each spoken digit accurately. Two measures were used in the segmentation process: the zero crossing rate and the signal energy. An example of the recorded speech file with the isolated spoken digits is shown in Fig. 1.

The set of recorded files for each user has been divided into two groups. One group, consisting of ten files, was chosen to form the dataset, while the remaining three files were used as a test set. Thus, the total tokens considered for training is 600 (6 speakers × 10 repetitions × 10 digits), and the total samples dedicated for testing phase is 180 tokens (6 speakers × 3 repetitions × 10 digits).

III. Feature Extraction

The speech is a signal consisting of a finite number of samples, yet a direct comparison between signals is impossible as the amount of information contained is high. Therefore, the most important features have to be extracted; this process is called feature extraction.

The main objective of this step is to recover a new meaningful underlying variables or features; that the data may easily be viewed with a reduced bandwidth compared to the input data resulting in improved recognition performance [1].

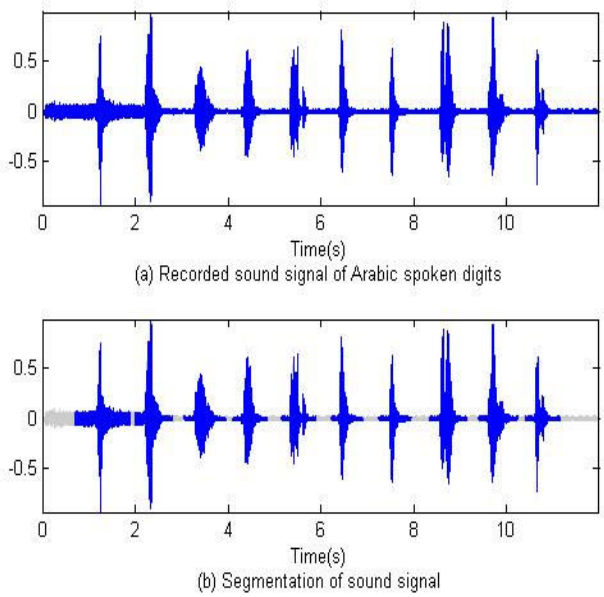

Fig. 1 An example of sound signals (a) and the segmentation result (b)

Most feature extraction methods use spectral analysis to extract meaningful components from the speech signal. Choosing effective features is important to achieve a high recognition performance. In this paper four features were considered to represent the sound template, specifically: Burg feature, Walsh feature, Thomson Multitaper feature and the MFCC. The recognition is based on the Euclidean distance from those templates, the closer the distance the better the match. So, the minimum distance value corresponds to the best match.

The Burg algorithm is a parametric spectral estimation method used to estimates the spectral content of signals by fitting an auto-regressive (AR) linear prediction filter model of a given order to the signal. As an example, Fig. 2 shows the Burg spectra of two spoken Arabic digits, four and six.

The Walsh spectrum analysis is an orthogonal transformation technique that decomposes a signal into a set of basis functions and estimates the power spectral density of the input signal [4].

The Multitaper method is a technique used to estimate the power spectrum of signals by utilizing several different data tapers (windows in the frequency domain) which are orthogonal to each other. It overcomes some of the limitations of conventional Fourier analysis [5]. Fig. 4 shows the Multitaper spectra of two spoken Arabic digits, four and six.

Cepstral based features such as MFCC are typically representing the magnitude of frequency band power for each speech window; they are widely used in speech processing. The MFCC maintained its dominance since its introduction in 1980 and because of its effectiveness, and even in noisy conditions it retains its strength. [6]. Fig. 5 shows the MFCC spectra of two spoken Arabic digits, four and six respectively. For more details on those audio features and their application on audio analysis one can refer to [1], [2] and [6].

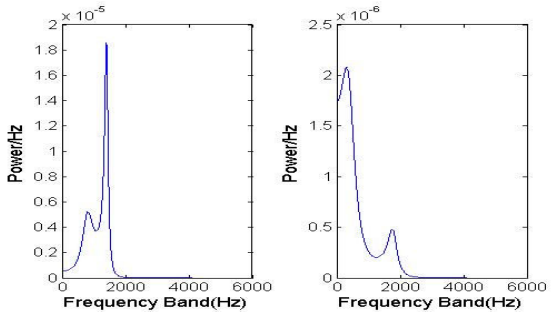

Fig. 2 Burg spectrum of spoken Arabic digits Four (a) and Six (b)

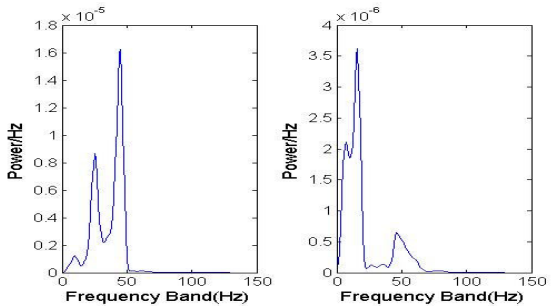

Fig. 3 Walsh spectrum of spoken Arabic digits Four (a) and Six (b)

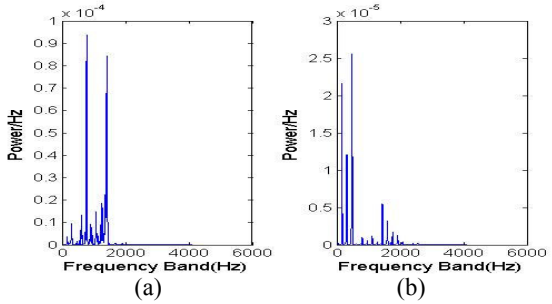

Fig. 4 Thomson Multitaper spectrum of spoken Arabic digits Four (a) and Six (b)

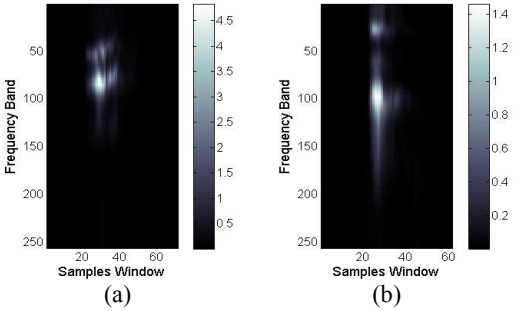

Fig. 5 MFCC features of spoken Arabic digits Four (a) and Six (b)

From Figures (2-5) we can see that there is a difference between the features of the chosen Arabic spoken digits. In fact, the same conclusion is true for all Arabic spoken digits. On the other hand, even for the same spoken digit we noted that there are variations in the features, as shown in Fig. 6.

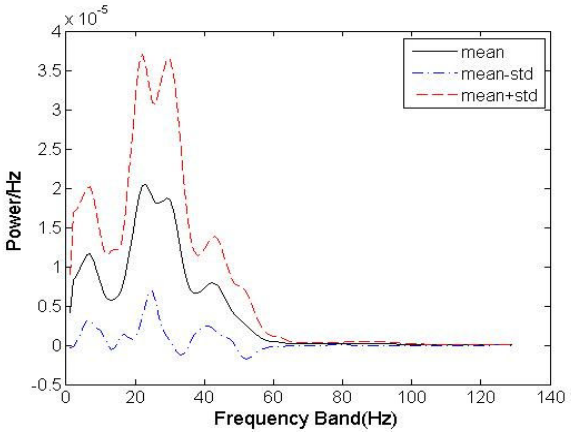

Fig. 6 The mean and the variance of the Burg features of the Arabic spoken digit Four

IV. Experimental Setup

The recognition of spoken Arabic digits was evaluated by performing 4 distinct experiments. Every experiment is concerned with a specific feature as shown in Table I.

The main stages of comparison steps are shown in the flowchart of Fig. 7. The Dynamic Time Wrapping (DTW) step is the nonlinear process that expands or contracts the time axis to match the same landmark positions between the input speech signal and the reference signal in the Database.

TABLE I
Comparison Experiments

Experiment Number	Recognition Feature
Exp 1	Burg Spectrum
Exp 2	Walsh Spectrum
Exp 3	Thomson Multitaper Spectrum
Exp 4	MFCC Analysis

V. Experimental Results

In order to assess the performance of recognition approaches the recognition of the Arabic spoken digits were evaluated for each experiment for six users. The obtained results are summarized in the following Tables and Figures.

Tables II-V show the system performance (recognition success rate (%) and confusion matrix that was generated for each experiment. The last two rows in each table show, respectively, the missed tokens per digit, and the individual system accuracies for each spoken digit.

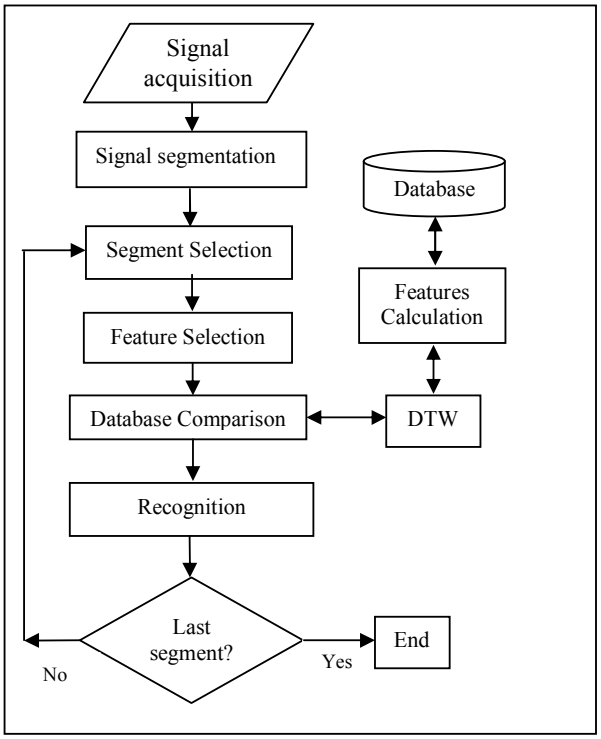

Fig. 7 Flowchart of the comparison tests

Table II shows the accuracy of the first experiment for each digit in addition to the system overall accuracy. The overall recognition accuracy is about 37% with 112 miss-recognized tokens out of 180. The worst performance was found in the case of digit 9 (with accuracy of about 22%); and the best performance was encountered in the case of digit 2 (with accuracy of about 77%). Two of the digits achieved over 70% accuracy but the remaining eight did not.

TABLE II
Confusion Matrix of Experiment 1

Num	0	1	2	3	4	5	6	7	8	9	
0	5	0	0	1	2	2	5	1	1	1	
1	0	13	0	1	3	2	2	3	1	3	
2	0	0	14	0	0	0	0	0	1	2	
3	0	0	0	6	0	2	1	0	1	2	
4	0	2	0	1	5	2	0	4	1	1	
5	1	0	0	0	0	5	0	2	1	1	
6	4	1	2	4	0	0	5	0	2	2	
7	2	1	0	1	4	2	2	5	1	2	
8	1	1	0	2	2	2	0	1	6	0	
9	5	0	2	2	2	1	3	2	3	4	
MR	**13**	**5**	**4**	**12**	**13**	**13**	**13**	**13**	**12**	**14**	**112**
Acc	**27**	**72**	**77**	**33**	**27**	**27**	**27**	**27**	**33**	**22**	**37**

MR: Number of Miss-recognized tokens
Acc: Percentage of recognition Accuracy (%).

It can be seen from Table III, that the totals of missed tokens of the second experiment were 78 out of 180. The overall recognition accuracy is about 56%. The best recognition accuracy was encountered with digit 2. The accuracy for this digit is about 94% and only one token of digit 2 was missed. On the other hand, the worst accuracy was encountered for the case of digit 3. The system accuracy for this digit was about 22% with a total of 14 missed tokens.

The overall recognition accuracy of the third experiment shown in Table IV is about 45% with a total of 98 miss-recognized tokens. The worst performance was found in the case of digits 4, 7 and 8 (with accuracy of about 33%); and the best performance was encountered in the case of digit 2 (with accuracy of about 77%).

TABLE III
CONFUSION MATRIX OF EXPERIMENT 2

Num	0	1	2	3	4	5	6	7	8	9	
0	11	0	0	0	0	1	0	0	2	1	
1	0	12	0	0	5	0	1	1	0	0	
2	1	0	17	0	0	0	2	0	0	1	
3	1	0	0	4	0	1	1	3	0	0	
4	1	1	0	1	9	1	1	2	1	2	
5	0	0	0	2	1	10	1	2	2	1	
6	3	0	0	4	0	1	12	1	1	2	
7	1	4	0	2	0	3	0	9	3	0	
8	0	0	0	3	1	0	0	0	7	0	
9	0	1	1	2	2	1	0	0	2	11	
MR	**7**	**6**	**1**	**14**	**9**	**8**	**6**	**9**	**11**	**7**	**78**
Acc	**61**	**66**	**94**	**22**	**50**	**55**	**66**	**50**	**38**	**61**	**56**

MR: Number of Miss-recognized tokens
Acc: Percentage of recognition Accuracy (%).

TABLE IV
CONFUSION MATRIX OF EXPERIMENT 3

Num	0	1	2	3	4	5	6	7	8	9	
0	8	0	2	0	2	0	0	0	3	1	
1	0	9	0	0	2	1	1	1	0	1	
2	3	1	14	2	0	5	4	2	4	3	
3	0	1	0	9	2	1	3	2	2	0	
4	0	1	0	2	6	1	0	3	1	0	
5	0	3	1	2	3	7	1	2	2	1	
6	5	1	0	1	1	0	9	1	0	2	
7	0	1	0	1	1	2	0	6	0	0	
8	0	1	1	1	1	1	0	0	6	2	
9	2	0	0	0	0	0	0	1	0	8	
MR	**10**	**9**	**4**	**9**	**12**	**11**	**9**	**12**	**12**	**10**	**98**
Acc	**44**	**50**	**77**	**50**	**33**	**38**	**50**	**33**	**33**	**44**	**45**

MR: Number of Miss-recognized tokens
Acc: Percentage of recognition Accuracy (%).

Analyzing the confusion matrix of last experiment shown in Table V, we can notice that the overall recognition accuracy is about 94%. The system failed in recognizing only 9 tokens out of the 180 total tokens. The worst performance was found in the case of digits 0 and 9 (with accuracy of about 83%); and the best performance was encountered in the case of digits 1, 2, 3, 4, 5 and 7 (with accuracy equal to 100%). Thus in this case six of the digits achieved 100% accuracy and the remaining four digits achieved over 80% accuracy.

Figures 8 & 9 depicted extra information about the performance of recognition experiments. The conclusion is that the recognition of spoken Arabic digits based on MFCC features (Experiment 4) was better than the other approaches. This conclusion is true for all users as shown in Fig. 8 and for all digits as shown in Fig. 9.

TABLE V
CONFUSION MATRIX OF EXPERIMENT 4

Num	0	1	2	3	4	5	6	7	8	9	
0	15	0	0	0	0	0	1	0	0	0	
1	1	18	0	0	0	0	0	0	0	1	
2	0	0	18	0	0	0	0	0	0	0	
3	0	0	0	18	0	0	0	0	2	0	
4	0	0	0	0	18	0	0	0	0	0	
5	0	0	0	0	0	18	0	0	0	1	
6	0	0	0	0	0	0	17	0	0	1	
7	0	0	0	0	0	0	0	18	0	0	
8	0	0	0	0	0	0	0	0	16	0	
9	2	0	0	0	0	0	0	0	0	15	
MR	**3**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**2**	**3**	**9**
Acc	**83**	**100**	**100**	**100**	**100**	**100**	**94**	**100**	**88**	**83**	**94**

MR: Number of Miss-recognized tokens
Acc: Percentage of recognition Accuracy (%).

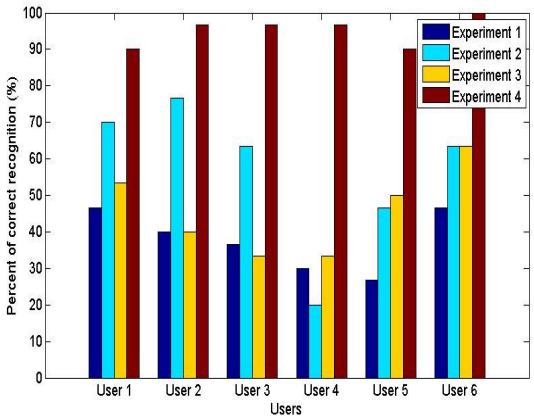

Fig. 8 Recognition result for each user

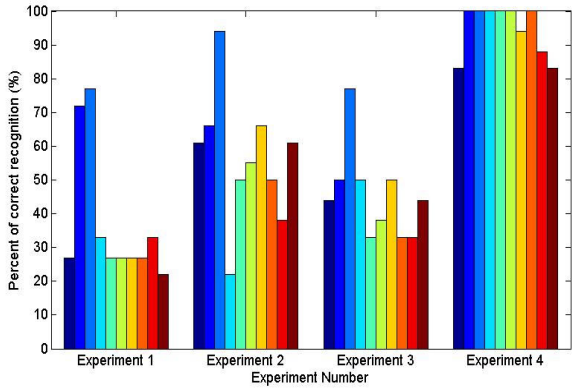

Fig. 9 Recognition accuracy rate for individual Arabic digits are for each experiment. For each experiment, the first bar from the left corresponds to the accuracy of digit 0, and the right one corresponds to the accuracy of digit 9

VI. Conclusion

In this work four parameterization features (the Burg spectrum features, Walsh spectrum features, the Thomson Multitaper spectrum features and MFCC features) were tested and compared for recognizing spoken Arabic digits. It has been noticed that the overall performance of spoken Arabic digits recognition based on MFCC features outperform the recognition based on other features for all users and for all digits.

References

[1] S. Theodoridis and K. Koutroumbas, Pattern Recognition, 3rd ed. Academic Press, Inc., 2008.
[2] J. Holmes, W. Holmes, Speech Synthesis and Recognition, 2nd ed., CRC Press, 2001.
[3] A. Ganoun and I. Almerhag, Performance Analysis of Spoken Arabic Digits Recognition Techniques, Journal of Electronic Science and Technology, vol. 10, no. 2, pp 153-157, June 2012.
[4] Beauchamp, K.G., Applications of Walsh and Related Functions, Academic Press, 1984.
[5] Percival, D. B., and A. T. Walden. Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques, Cambridge University Press, 1993.
[6] Stoica, P., and R.L. Moses, Introduction to Spectral Analysis, 1st ed., Prentice-Hall, 1997.
[7] K. Saeed and M. Nammous, A Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-Signal Image, IEEE Transactions On Industrial Electronics, vol. 54, no. 2, April 2007.
[8] Z. Hachkar et al., Comparison of MFCC and PLP Parameterization in pattern recognition of Arabic Alphabet Speech, Canadian Journal on Artificial Intelligence, Machine Learning & Pattern Recognition vol. 2, no. 3, April 2011.
[9] M. Abushariah et al., 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.
[10] T. Ganchev, M. Siafarikas and N. Fakotakis, Evaluation of speech parameterization methods for speaker recognition, Proc. of the Acoustics, vol. 18-19, pp. 105-110, 2006.