Performance Analysis of Spoken Arabic Digits Recognition Techniques

Ali Ganoun and Ibrahim Almerhag

***Abstract*—A performance evaluation of sound recognition techniques in recognizing some spoken Arabic words, namely digits from zero to nine, is proposed. One of the main characteristics of all Arabic digits is polysyllabic words except for zero. The performance analysis is based on different features of phonetic isolated Arabic digits. The main aim of this paper is to compare, analyze, and discuss the outcomes of spoken Arabic digits recognition systems based on three recognition features: the Yule-Walker spectrum features, the Walsh spectrum features, and the Mel frequency Cepstral coefficients (MFCC) features. The MFCC based recognition system achieves the best average correct recognition. On the other hand, the Yule-Walker based recognition system achieves the worst average correct recognition.**

***Index Terms*—Arabic digits, spectrum analysis, speech recognition.**

1. 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, airline reservation, and automatic directory to retrieve or send information[1].

The main advantage of automatic recognition systems of spoken digits is the ease of speech inputting as it does not require any specialized skills. Another advantage is that the information could be recorded even if the user is involved in other activities.

Manuscript received June 1, 2012; revised June 11, 2012; presented at 2012 2nd International Conference on Signal, Image Processing and Applications, Hong Kong, August 3–4, 2012.

A. Ganoun is with the Faculty of Engineering, University of Triploi, Triploi 13275, Libya (e-mail: ali.ganoun@ee.edu.ly).

I. Almerhag is with the Faculty of Information Technology, University of Triploi, Triploi 13275, Libya (e-mail: almerhag@hotmail.com).

Digital Object Identifier: 10.3969/j.issn.1674-862X.2012.02.011

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 discrimination between spoken digits difficult, and 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 interest in this work is motivated by the minimum efforts in applying known speech recognition techniques on Arabic language recognition in comparison with other languages. In addition, we think that, the performance of recognition systems is language dependent. Therefore, conclusions drawn 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 a single speaker using three features used to represent sound signals: the Yule-Walker spectrum analysis, the Walsh spectrum, and the Mel frequency Cepstral coefficients (MFCC) analysis. The performance evaluation of the recognition system is based on the overall system performance and the individual digit accuracy using two parameters: the normalization of the sound feature vector and filtering of the sound feature vector[3].

The rest of the paper is organized as follows. Section 2 presents a description of the database used by the system. Section 3 presents a brief description of feature extraction processes. Section 4 discusses the experimental setup. Section 5 presents the results of comparisons obtained as a result of this work. The paper concludes with Section 6.

2. Database Preparation

In order to evaluate the selected recognition techniques, a database of the sounds of the Arabic digits (0 to 9) was created; where a male Arabic native speaker was asked to utter all digits; each time the speech was recorded in a single file which was approximately 12 s long. This process was repeated 13 times, so that 13 speech files were collected, and each file contained all the Arabic digits.

Every speech file contained 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. In this process, two measures were used in the segmentation of the sound signals: the zero crossing rate and the signal energy.

The set of recorded samples has been divided into two groups. One group, consisting of ten samples, was chosen to form the dataset, while the remaining three samples were used as a test set.

3. 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 large. Therefore, the most important features have to be extracted; this process is called feature extraction.

The main objective of this step is to transform the original data into a dataset with a reduced number of variables that contain the most discriminatory information and provide a relevant set of features for a classifier, resulting in improved recognition performance[1]. An example of the recorded speech file with the segmented spoken digits is shown in Fig. 1.

Another goal is to recover a new meaningful underlying variables or features; the data may easily be viewed with a reduced bandwidth compared with the input data.

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 three features were used in the comparison, specifically: Yule-Walker spectrum analysis, Walsh spectrum analysis, and MFCC.

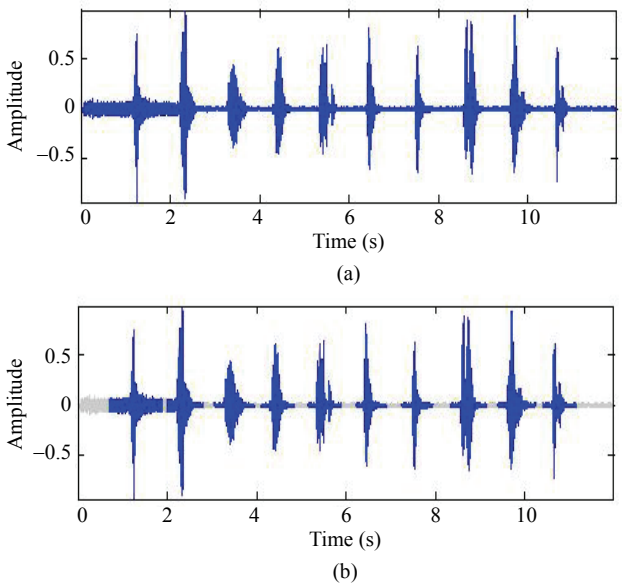

Fig. 1. Example of sound signals: (a) recoded sound signal of Arabic spoken digits and (b) segmentation of sound signal.

The Yule-Walker algorithm estimates the spectral content of the sound signal by fitting an auto-regressive linear prediction filter model of a given order to the signal. Cepstral based features, such as MFCC, typically represent the magnitude of frequency band power for each speech window, which are widely used in speech processing.

The comparison between the test signal and the signals stored in database is based on the Euclidean distance between the two features; the closer the distance, the better the matches. So, the minimum distance value corresponds to the best match.

Figs. 2 to Fig. 4 show the spectra of the selected features of two spoken Arabic digits, One and Nine. For more details on those audio features and their application on audio analysis, one can refer to [4]–[7].

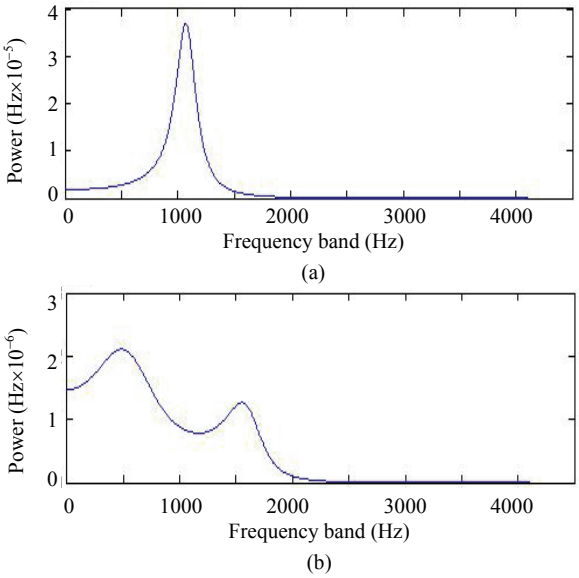

Fig. 2. Yule-Walker spectrum of the spoken digits: (a) One and (b) Nine.

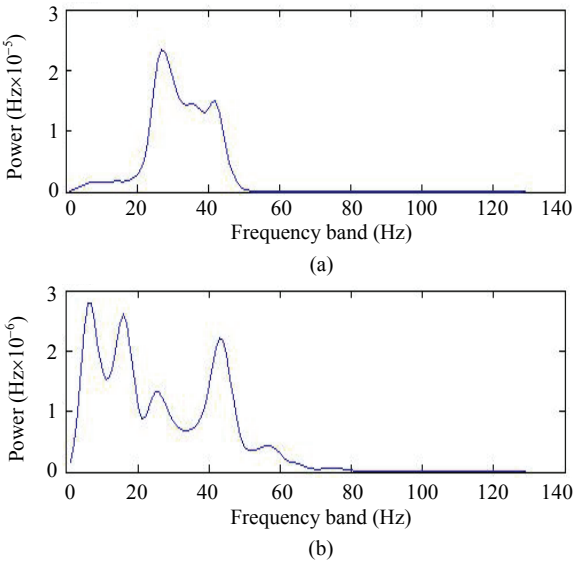

Fig. 3. Samples of Walsh spectrum of Arabic spoken digits: (a) One and (b) Nine.

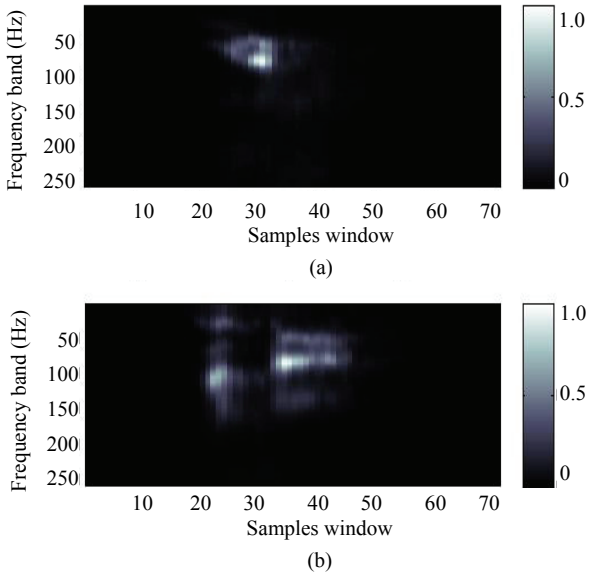

Fig. 4. MFCC features of the Arabic spoken digits: (a) One and (b) Nine.

From Fig. 2 to Fig. 4, 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. In general, for the selected three features, the correlation between the features of different spoken digits is very low. On the other hand, even for the same spoken digit we noted, there are variations in the features, as shown in Fig. 5.

Normalization and filtration of the sound feature vector are the parameters used for the sake of comparison between the selected features. Normalization will adjust the feature level from 0 to 1. We expect that this will increase the performance of comparison between the features of the same spoken digit with different volumes. The othe parameter is the filtration of the features in order to smoot the feature vectors. Fig. 6 shows the effect of these tw parameters on the Walsh feature vector of the Arabi spoken digit Zero.

4. Experimental Setup

For each test sequence every spoken sound i recognized independently. The performance of the selected techniques is evaluated based on the recognition of Arabic spoken digits by performing 12 distinct experiments. Every single experiment is concerned about a specific feature with certain parameters as shown in Table 1.

For all experiments we select the best five matches among the test signal and the signals stored in the database.

The main stages of the comparison steps are shown in 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 1: Comparison experiments

Experiment number	Recognition approaches	Normalize feature vector	Feature filtering
Exp 1	Yule-Walker spectrum analysis	0	0
Exp 2		1	0
Exp 3		0	1
Exp 4		1	1
Exp 5	Walsh spectrum	0	0
Exp 6		1	0
Exp 7		0	1
Exp 8		1	1
Exp 9	MFCC analysis	0	0
Exp 10		1	0
Exp 11		0	1
Exp 12		1	1

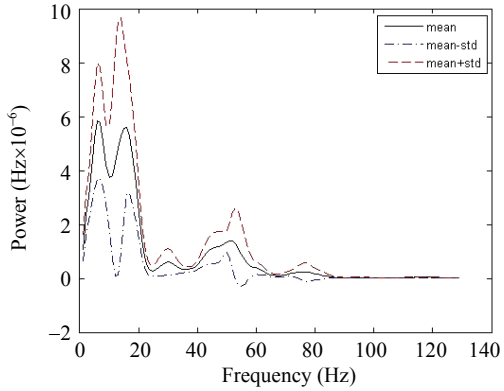

Fig. 5. Mean and the variance of the Walsh features of the Arabic spoken digit Zero.

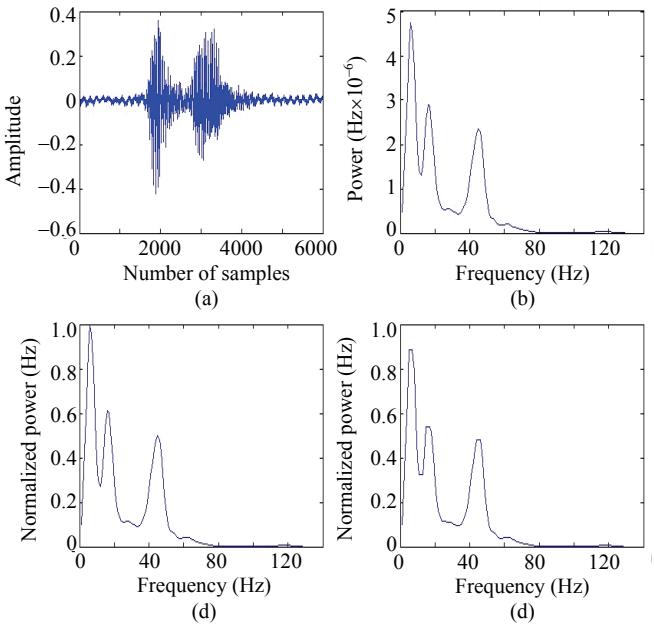

Fig. 6. Walsh features of the spoken digit zero and the effect of normalization and filtering of the feature vector: (a) sound signal (b) unnormalized feature, (c) normalized feature, and (d) filtered and normalized feature.

5. Results

In order to investigate the performance of recognition approaches the recognition of the Arabic spoken digits was evaluated for each experiment with three test sequences. The obtained results are summarized in Fig. 8, Fig. 9 and Table 2. Fig. 8 shows the best match of the three test sequences with each experiment. In general it can be noted that the comparison based on MFCC features gives the best recognition results.

Another way to represent the recognition results is by calculating the percentages of the exact (best) match in the first five matches. Fig. 9 shows the percentages of the correct match in the first five matches of the three test sequences with each experiment. Table 2 shows both the average score per experiment and the average score for each recognized digit.

The results show that the spoken digit 2 achieved the highest recognition rate (with accuracy equal to 83%); then the spoken digit 1 (with accuracy equal to 76%). Again, MFCC analysis gives the best recognition results for the percentages of the first five correct recognition matches. Experiments 9 and 11 in the MFCC analysis without normalization of the feature vectors can be considered here as the best approaches for the recognition of Arabic spoken digits (with accuracy equal to 87% for both cases). From the result shown in Table 2 we remark also that the recognition of spoken digits 9 and 7 was the worst compared with other spoken digits (with accuracy equal to 45% for both cases).

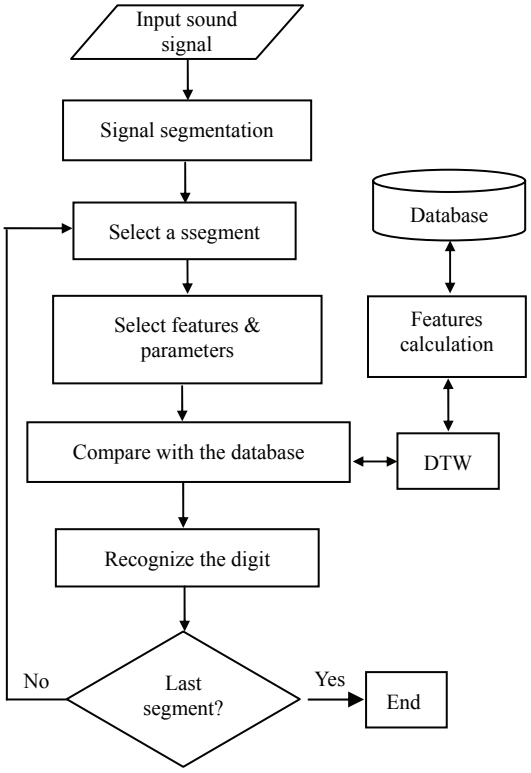

Fig. 7. Flowchart of the comparison tests.

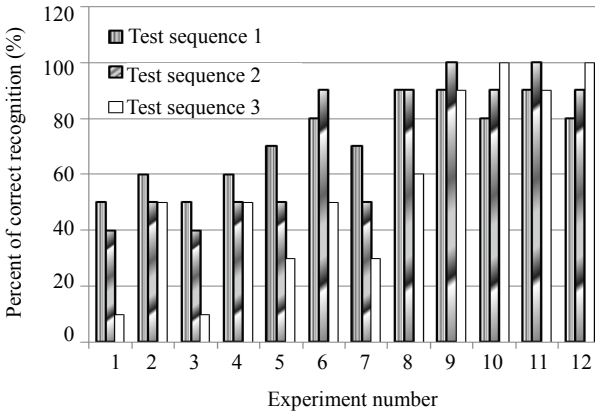

Fig. 8. Percentages of the best correct match of the three test sequences with each experiment.

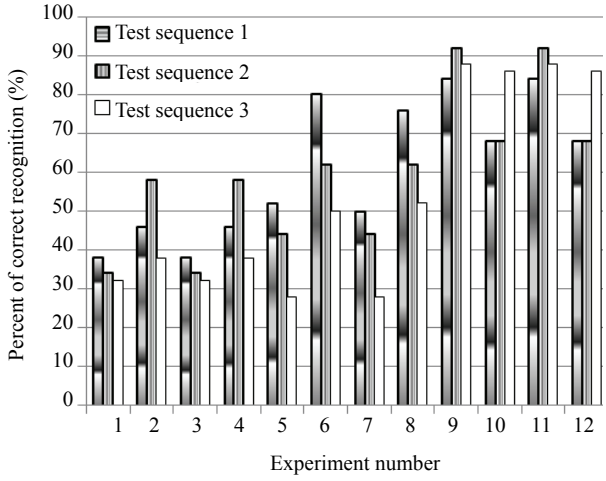

Fig. 9. Percentages of the correct match in the first five matches of the three test sequences for each experiment.

6. Conclusions

In this paper a comparison of three approaches for the recognition of Arabic spoken digits has been presented. As expected, it has been shown that the recognition of Arabic spoken digits based on MFCC features outperform the recognition based on both Yule-Walker features and Walsh spectrum features.

Further research will attempt to produce more comparisons based on other features and larger databases with more than one speaker.

Table 2: Recognition rate of Arabic spoken digits

Num.	0	1	2	3	4	5	6	7	8	9	Avg.
Exp. 1	60	53	60	40	26	20	13	20	33	20	34
Exp. 2	73	86	86	40	40	46	26	33	20	20	47
Exp. 3	60	53	60	40	26	20	13	20	33	20	34
Exp. 4	73	86	86	40	40	46	26	33	20	20	47
Exp. 5	66	40	60	53	33	33	20	20	60	26	41
Exp. 6	73	93	100	73	46	46	60	46	53	46	63
Exp. 7	66	40	60	46	33	33	20	20	60	26	40
Exp. 8	73	93	93	73	46	46	53	46	53	53	62
Exp. 9	100	100	100	80	86	73	80	93	80	86	87
Exp. 10	46	86	100	80	73	66	86	60	66	73	73
Exp. 11	100	100	100	80	86	73	80	93	80	86	87
Exp. 12	46	86	100	80	73	66	86	60	66	73	73
Avg.	69	76	83	60	50	47	46	45	52	45	

References

[1] S. Theodoridis and K. Koutroumbas, *Pattern Recognition,* 3rd ed. San Diego: Academic Press, Inc., 2006.

[2] J. Holmes and W. Holmes, *Speech Synthesis and Recognition*, London: Taylor & Francis, 2001.

[3] K. Saeed and M. Nammous, "A speech-and-speaker identification system: feature extraction, description, and classification of speech-signal image," *IEEE Trans. on Industrial Electronics*, vol. 54, no. 2, pp. 887–897, 2007.

[4] Z. Hachkar, B. Mounir, A. Farchi, *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, pp. 56–60, 2011.

[5] M. Abushariah, R. Ainon, R. Zainuddin, *et al*., "Arabic speaker-independent continuous automatic speech recognition based on a phonetically rich and balanced speech corpus," *The Int. Arab Journal of Information Technology*, vol. 9, no. 1, pp. 84–93, 2012.

[6] M. Abdulfattah and R. El Awady, "Phonetic recognition of Arabic alphabet letters using neural networks," *Int. Journal of Electric & Computer Sciences*, vol. 11, no. 1, pp. 52–58, 2011.

[7] 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, Sep. 2006.

Ali Ganoun was born in Libya, in 1966. He received the B.S. degree from the University of Benghazi in 1988, the M.Sc. degree from the University of Tripoli in 1995, both in electrical engineering, and the Ph.D. degree from Orleans University, France, in 2007. He is currently a lecturer with the Electrical Engineering Department of University of Triploi, Faculty of Engineering, Libya. His research interests include signal and image processing and computer vision.

Ibrahim Almerhag was born in Libya, in 1963. He received his Ph.D. degree in computing in 2006 and the MBA in 2002 from Bradford University. He also holds the M.Sc. degree in electronics and computer engineering from the Technical University of Warsaw in 1995. Currently, he is holding the post of assistant professor with the Faculty of Information Technology, University of Tripoli-Libya. His research interests include networking, information security, and signal and image processing.