Hindawi Publishing Corporation
Conference Papers in Engineering
Volume 2013, Article ID 130473, 6 pages
http://dx.doi.org/10.1155/2013/130473

Conference Paper

Speech Recognition of Arabic Spoken Digits

Anas Allosh, Nura Zlitni, and Ali Ganoun

University of Tripoli, Faculty of Engineering, P.O. Box 13589, Tripoli, Libya

Correspondence should be addressed to Ali Ganoun; ali.ganoun@gmail.com

Received 27 February 2013; Accepted 9 May 2013

Academic Editors: A. Gaouda and H. Koivo

This Conference Paper is based on a presentation given by Anas Allosh at "International Conference on Electrical and Computer Engineering" held from 26 March 2013 to 28 March 2013 in Benghazi, Libya.

With the widespread growth in the use of digital computers, there has been an increasing need to be able to communicate with machines in a simple manner. One of the main tasks that simplify the communication with machines is the speech recognition. Speech recognition is the translation of spoken words into text. However, speech recognition is a very complex problem. This paper is related to the recognition of spoken Arabic digits. Two recognition techniques have been implemented and tested: Pitch Detection Algorithm (PDA) and Cepstrum Correlation Algorithm (CCA). In order to analyze the recognition accuracy of the selected techniques, a database of spoken Arabic digits has been created. The performance of the two techniques has been analyzed based on the created database.

1. Introduction

Speech can be classified into two general categories: voiced and unvoiced speech. A voiced speech is one in which the vocal cords of the speaker vibrate as the sound is made, and unvoiced speech is one where the vocal cords do not vibrate [1, 2].

There are many techniques used in recognizing voiced speech, and in this paper, we consider two techniques: Pitch Detection Algorithm (PDA) [3–6] and Cepstrum Correlation Algorithm (CCA) [7].

The PDA is one of the most robust and reliable techniques; it is known to have a very high accuracy for voiced pitch estimation. It is a commonly used method to estimate pitch level and is based on detecting the highest value of the autocorrelation function in the region of interest.

The CCA method is mostly used to obtain Mel Frequency Coefficients (MFCC) as vectors are used as a pattern recognition technique. But here we used a very simplified way with almost the same results of recognition instead of the MFCC recognition technique.

The main objective of this paper is to study and implement the two sound recognition techniques. Another objective is to create a database of spoken Arabic digits with different users and then use it to evaluate the performance of the selected techniques. Another objective is to design a Graphical User Interface (GUI) which will be used in the analysis and performance comparison of recognition techniques.

2. Speech Recognition

The frequency of human voice ranges from 20 Hz to 14,000 Hz (typically from 300 Hz to 4,000 Hz). The frequency of a sound wave determines the human tone and pitch. In general, the frequencies, which have the most significant part of speech, lie between about 100 Hz and 4,000 Hz.

2.1. Preprocessing Techniques. There are many preprocessing techniques needed to perform speech recognition. The main pre-processing techniques are explained as follows.

2.1.1. Normalization. Normalization is the process by which the signal is brought into a range consistent with expected values. Normalization technique has been considered for audio signals to prevent clipping. It is therefore good practice to scale audio being processed to a specific range.

2.1.2. Zero Padding. Zero padding is a useful process that can be used in many applications for various reasons. In this

paper, zero padding was used because the correlation process used later cannot be done unless both the database signal and the test signal are equal in length.

2.2. Cepstrum Correlation Algorithm (CCA). Classically, spectral techniques making use of a short-time Fourier transform have been the dominant solution for classification and recognition problems.

The term "cepstrum" was coined by Bogert et al. [8] by swapping the order of the letters in the word "spectrum."

The Cepstrum is a common transform used to gain information from a person's speech signal. It can be used to separate the excitation signal (which contains the words and the pitch) and the transfer function (which contains the voice quality).

The cepstrum is defined as the inverse DFT of the log magnitude of the DFT of a signal. Consider

$$c[n] = \mathcal{F}^{-1}\left\{\log\left|\mathcal{F}\left\{x[n]\right\}\right|\right\}, \quad (1)$$

where $\mathcal{F}$ is the DFT and $\mathcal{F}^{-1}$ is the IDFT. For a windowed frame of speech $x[n]$, the cepstrum is

$$c[n] = \sum_{k=0}^{N-1}\left(\log\left|\sum_{m=0}^{N-1} x[m]\, e^{-j(2\pi/N)km}\right|\right) e^{-j(2\pi/N)kn}. \quad (2)$$

Figure 1 shows how a signal would be converted to the Cepstral domain. Consider the magnitude spectrum $|\mathcal{F}\{x[n]\}|$ of the periodic signal in Figure 1; this spectrum contains harmonics at evenly spaced intervals, whose magnitude decreases quite quickly as frequency increases. By calculating the log spectrum, however, we can compress the dynamic range and reduce amplitude differences in the harmonics.

Now if we were told that the log spectrum was a waveform, in this case, we would describe it as quasiperiodic with some form of amplitude modulation, to separate both components; we could then employ the DFT, and we would expect the DFT to show a large spike around the "period" of the signal and some "low-frequency" components due to the amplitude modulation. Figure 2 shows the flow chart of the Cepstrum Algorithm.

2.3. Pitch Detection Algorithm (PDA). PDA is an algorithm designed to estimate the pitch or fundamental frequency of a quasiperiodic or virtually periodic signal, usually a digital recording of speech.

When a segment of a signal is correlated with itself, the distance between the positions of the maximum and the second maximum correlation is defined as the fundamental period (pitch) of the signal.

The modified autocorrelation pitch detector based on the center clipping method and infinite clipping is used in our implementation.

2.3.1. Clipping. Two techniques are considered.

Center Clipping. Center clipping works by clipping a certain percentage of the waveform. Let Amax be the maximum amplitude of the signal and let CL be the clipping level. CL

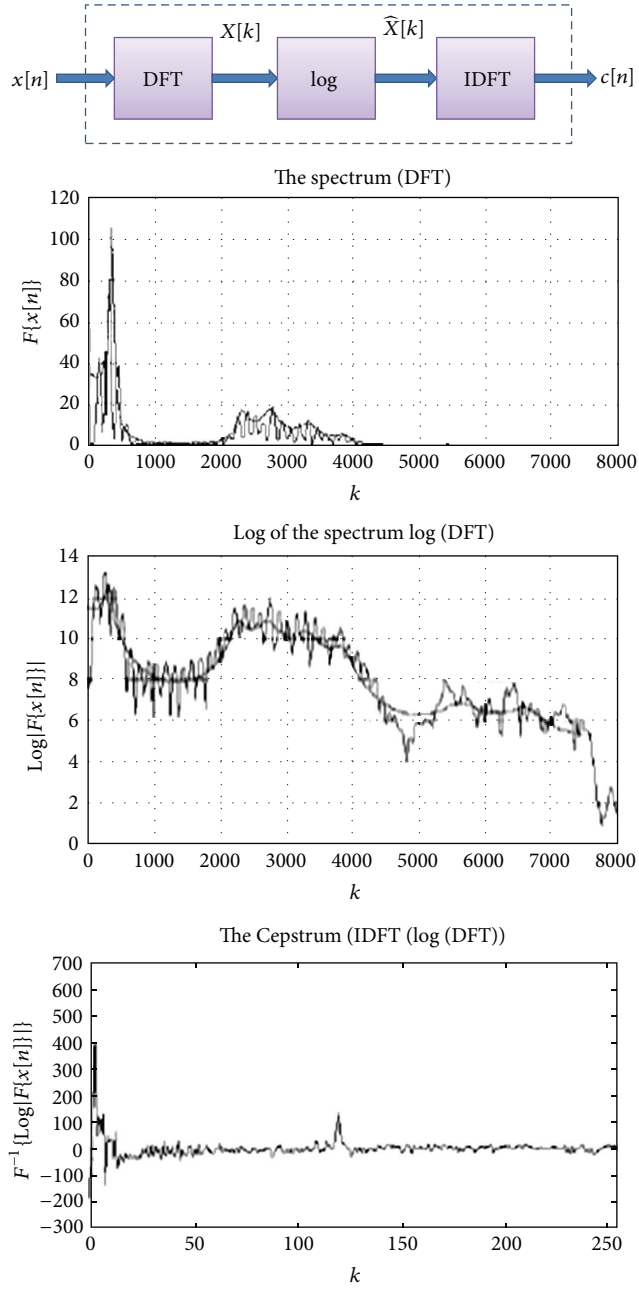

Figure 1: Computation of the Cepstrum of a signal $x(n)$.

is a fixed percentage of $A_{\max}$. Therefore, the output of this approach is as follows:

$$f(x) = \begin{cases} (x(n) - \mathrm{CL}), & x(n) \geq \mathrm{CL}, \\ 0, & x(n) < \mathrm{CL}, \\ (x(n) + \mathrm{CL}), & x(n) \leq -\mathrm{CL}, \end{cases} \quad (3)$$

where CL is the clipping threshold.

Infinite Peak Clipping. Infinite peak clipping works as follows:

$$f(x) = \begin{cases} 1, & x(n) \geq \mathrm{CL}, \\ 0, & x(n) < \mathrm{CL}, \\ -1, & x(n) \leq -\mathrm{CL}. \end{cases} \quad (4)$$

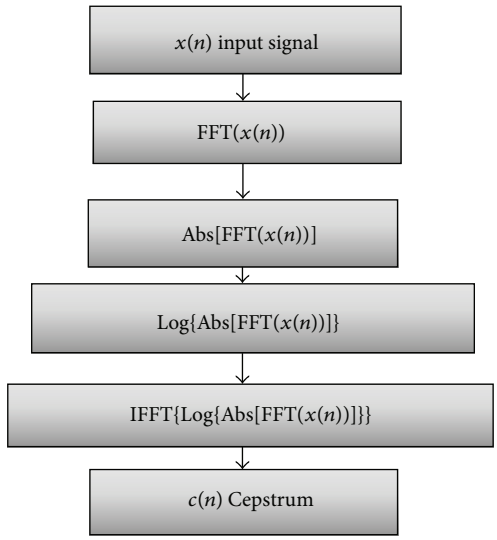

Figure 2: Flow chart of the Cepstrum Algorithm.

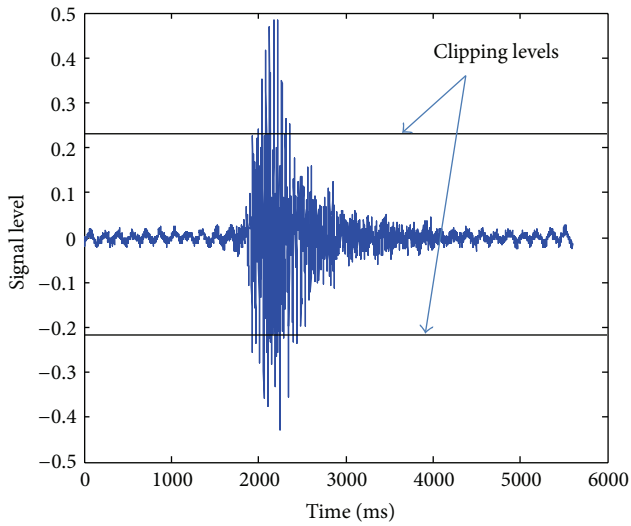

Figure 3: The result of the clipping value is determined by 50% clipping of the signal.

Figure 3 shows an example of clipping using 50% clipping of the signal.

The autocorrelation pitch detector is one of the most robust and reliable pitch detectors based on detecting the highest value of the autocorrelation function. The autocorrelation is calculated as follows:

$$Rx(m) = \frac{1}{N}\sum_{n=0}^{N'-1} x(n+1)\, w(n)\, [x(n+I+m)\, w(n+m)], \quad 0 \le m \le M_0, \tag{5}$$

where: $w(n)$ = approximate window for analysis, N = section length being analyzed, M_0 = number of autocorrelation points to be computed, I = index of the starting sample of the frame, and N' = number of signal samples in computation of $Rx(m)$, for pitch detection applications.

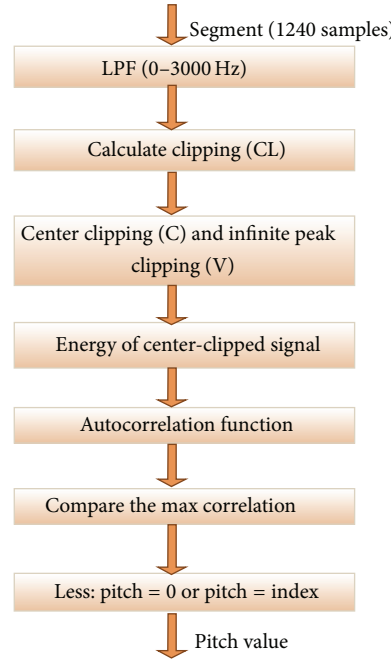

Figure 4: Flow chart of Pitch Detection Algorithm.

N' is generally as follows:

$$N' = N - m. \tag{6}$$

2.3.2. Voiced/Unvoiced Detection. The autocorrelation function is searched for its maximum value. If the maximum exceeds 0.61 (energy as the threshold) of the autocorrelation value at 0 delay, the section is classified as voiced and the location of the maximum is the pitch period. Otherwise, the section is classified as unvoiced.

2.3.3. Formant. A formant is a concentration of acoustic energy around a particular frequency in the speech wave. There are several formants, each at a different frequency, roughly one in each 1000 Hz band. Each formant corresponds to a resonance in the vocal tract.

Figure 4 shows the flow chart of Pitch Detection Algorithm.

3. Results

3.1. Database Preparation. In order to evaluate the performance of PDA and CCA algorithms, two databases of the spoken Arabic digits (0 to 9) were created: three males (User 1, User 2, and User 3) and three females (User 4, User 5, and User 6).

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 for each user and each file contained all the Arabic digits.

Every speech file contained both speech signals and nonspeech signals. 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

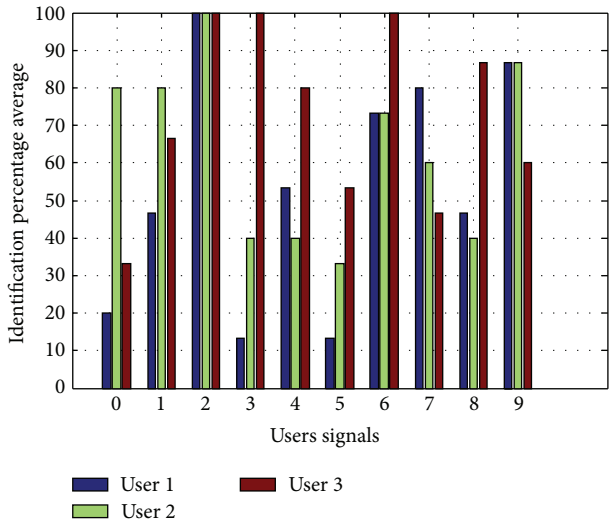

Figure 5: The percentage of all males' voices without normalization.

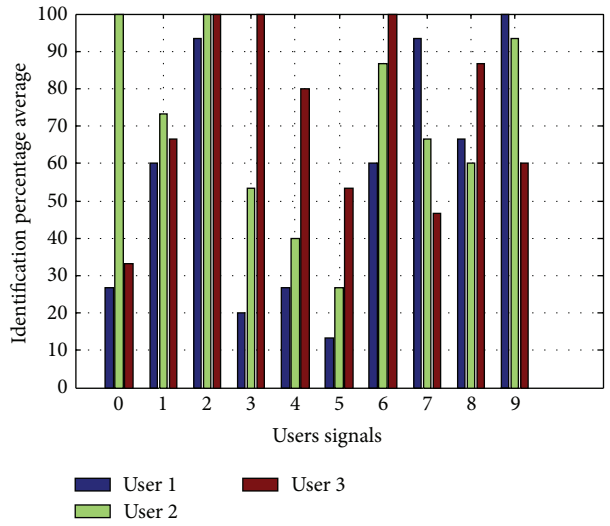

Figure 6: The percentage of all males' voices with normalization.

segmentation of the sound signals: the zero crossing and the signal energy.

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. The GT (Ground truth), which defines the original signal number for each spoken digit inside the database, has been used to compare between the correct spoken digit and the test signal. Based on the recognition result of each technique the percent of correct recognition can be calculated.

The recognition starts by taking two signals: one from the created database and the other from the test samples, and the objective is to recognize the spoken Arabic digit based on CCA and PDA techniques.

This procedure will be repeated among the whole 10 signals, taking the best five results.

3.2. Result of the CCA Technique. The results were taken in two groups, first one for males and the second for females. Each group has the recognition results with and without normalization.

The males' recognition results without normalization are shown in Figure 5. Figure 6 shows males' recognition results with normalization.

From Figures 5 and 6, we note the following.

(i) Recognition of User 2 and User 3 results was relatively better than other users because their data records were taken at a quite environment.

(ii) The results obtained with normalization were relatively better than the results without normalization.

(iii) Spoken Arabic digit 6 was the hardest number to be recognized because of the syllables nature in Arabic language.

(iv) Spoken Arabic digits 3, 7, and 10 had the best recognition results because they consist of strong Arabic syllables, which make them easy to be recognized.

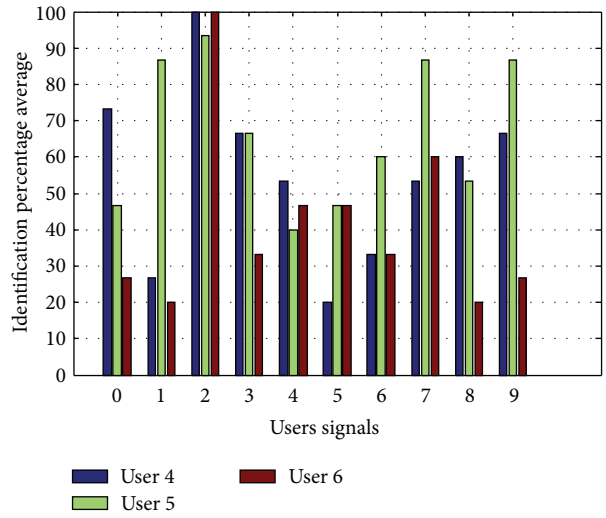

Figure 7: The percentage of all females' voices without normalization.

(v) The records of User 1 were taken in a different microphone than records of User 2 and User 3, and the microphone of the last 2 users had larger sensitivity to spoken words, which made them more recognizable by the recognition techniques.

The females' recognition results without normalization are shown in Figure 7. Figure 8 shows females' recognition results with normalization.

3.3. Result of the PDA Technique. The recognition results were considered using the center clipping and infinite peak clipping.

Results of Centre Clipping. The bar chart shown in Figure 9 represents the results for the three test signals of User 1 for each spoken digit.

Here we can see the "zero" with the lowest percentage whereas "eight" with the best result, it scored 100% twice. The

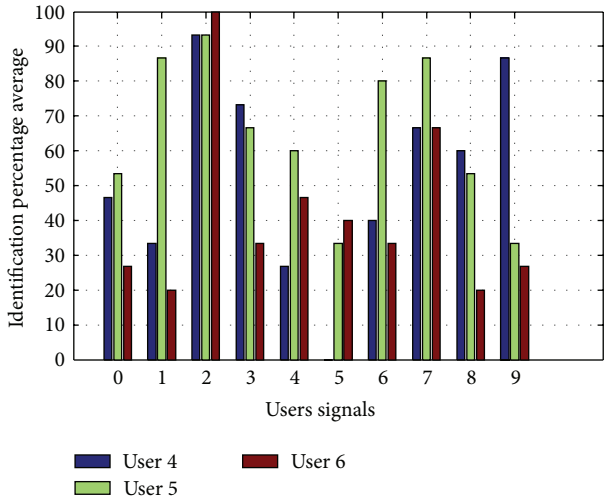

Figure 8: The percentage of all females' voices with normalization.

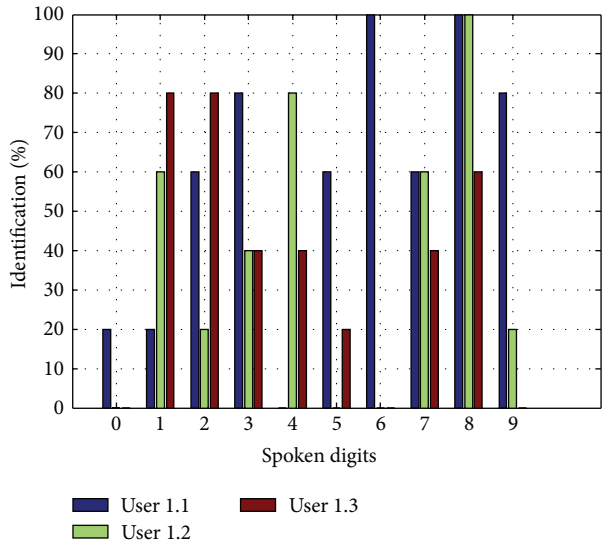

Figure 9: Recognition results of the three test signals of User 1.

digit "Six" is correct with a correct recognition percent of 100% with the first user, but it has 0% of correct recognition with users 2 and 3. Others are slightly the same.

Results of Infinite Peak Clipping. Figure 10 shows the results for all the text signals of males' voices.

As can be noted from the figure, the percentages are not the same but vary according to the users.

The recognition results for all male and female users can be summarized in Tables 1 and 2, respectively. Table 1 shows that the center clipping acquired the best result. Using 1/3 of the beginning and the end of the signal is much better than using the whole signal.

PDA recognition using center clipping was found better for males' (35.8% accuracy) compared to females' results (31.8% accuracy). On the other hand, infinite peak clipping gives about 25.57% accuracy. For females in both cases, the

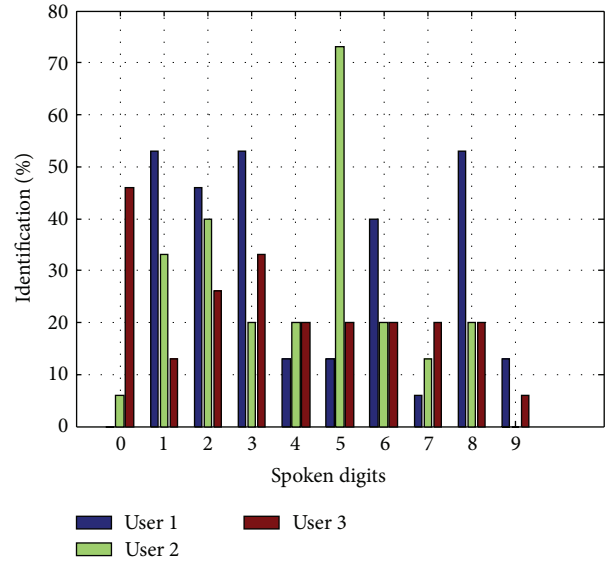

Figure 10: Percentage of all males' voices for all spoken digits.

Table 1: Recognition accuracy of the average male users.

	Whole signal	1/3 signal	Average
Center clipping	34.22%	37.33%	35.775%
Infinite peak clipping	25.554%	28.444%	25.497%

Table 2: Recognition accuracy of the average female users.

	Whole signal	1/3 signal	Average
Center clipping	31.996%	37.33%	31.828%
Infinite peak clipping	33.33%	28.444%	31.66%

recognition was slightly the same. For this reason center clipping is commonly used.

In general, the results of using CCA were relatively better than PDA approach.

3.4. Graphical User Interface Development. A Graphical User Interface (GUI) [9] has been designed using MATLAB software to analyze voice signals. Using the GUI, we can select the recognition technique and the parameters of the selected method.

The operation "Performance evaluation" was used to evaluate the recognition result of the selected technique. This was done by loading a prerecorded signal and then performing the recognition based on the selected technique, as shown in Figure 11.

As can be seen in Figure 11, each recognized number has its own percentage of recognition. The overall percentage is also shown which represents the total evaluation of the recognition program.

Another example is shown in Figure 12 using PDA technique.

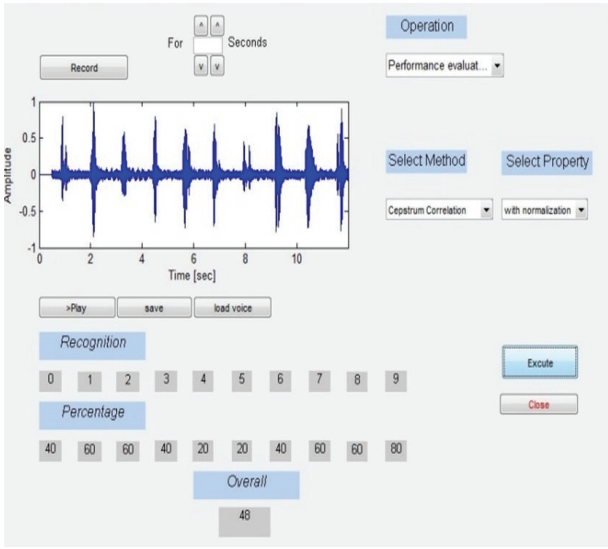

Figure 11: Performance evaluation using CCA.

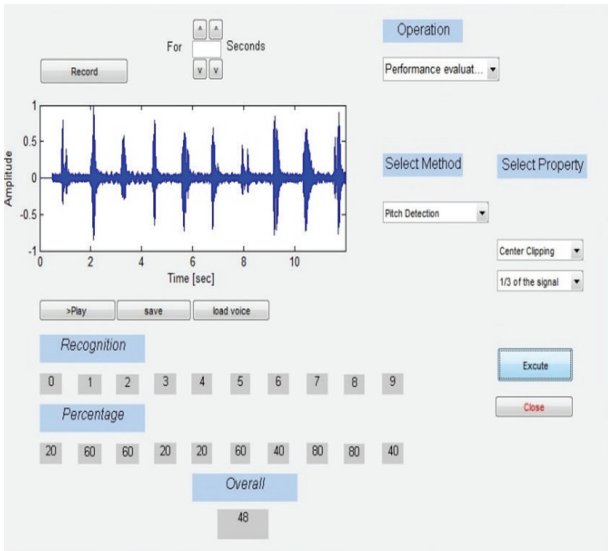

Figure 12: Performance evaluation using PDA.

4. Conclusion

The paper describes the recognition of spoken Arabic digits using two techniques: the PDA and CCA.

Spoken Arabic digits six and nine were especially difficult to be recognized. The reason for that is the complexity of the voiced signals of these two spoken digits.

As a future work on speech recognition, we will consider the following tasks:

(i) using DSP kits in recognition tasks;

(ii) using other recognition techniques such as hidden Markov model (HMM).

References

[1] I. Mcloughlin, *Applied Speech and Audio Processing with Matlab Examples*, 2009.

[2] J. P. Marques, *Pattern Recognition-Concepts Methods and Applications*, 2001.

[3] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Audio, Signal, and Speech Processing*, vol. 24, no. 5, pp. 399–418, 1976.

[4] S.-H. Chen and Y.-R. Wang, "Vector quantization of pitch information in Mandarin speech," *IEEE Transactions on Communications*, vol. 38, no. 9, pp. 1317–1320, 1990.

[5] M. M. Sondhi, "New methods of pitch extraction," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-16, pp. 262–266, 1968.

[6] M. J. Ross, H. L. Shaffer, and A. Cohen, "Average magnitude difference function pitch extractor," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 5, pp. 353–362, 1974.

[7] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 1, pp. 24–33, 1977.

[8] B. Bogert, M. Healy, and J. Tukey, "The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proceedings of the Symposium on Time Series Analysis*, pp. 209–243, 1963.

[9] V. K. Ingle and J. G. Proakis, *Digital Signal Processing Using MATLAB*, 1997.