Arabic word dependent speaker identification system using artificial neural network

Aws Al-Qaisi
Communication Engineering Department, Faculty of Engineering Technology,
Al-Balqa Applied University
Jordan

Received: May 1, 2020. Revised: June 5, 2020. Accepted: July 7, 2020. Published: July 9, 2020.

Abstract — The security of systems is a vital issue for any society. Hence, the need for authentication mechanisms that protect the confidentiality of users is important. This paper proposes a speech based security system that is able to identify Arabic speakers by using an Arabic word (شكرا) which means "Thank you". The pre-processing steps are performed on the speech signals to enhance the signal to noise ratio. Features of speakers are obtained as Mel-Frequency Cepstral Coefficients (MFCC). Moreover, feature selection (FS) and radial basis function neural network (RBFNN) are implemented to classify and identify speakers. The proposed security system gives a 97.5% accuracy rate in its user identification process.

Keywords— NN, MFCC, Speaker Identification, RBFNN.

I. Introduction

THE security of systems is an essential issue for all communities; as a result, this raises the need for a mechanism which protects the confidentiality of users. Authentication techniques are used in diverse systems and are recommended for many commercial applications such as credit cards, passwords, signatures, and voice features. Biometric authentication is one of the aurthentication techniques that is reasonably accepted to verify personal identities, Hence, biometric techniques become a hot topic of research. Human speech signals contain important information that can be applied in various practical purposes. Speech signals can be used to identify speakers and their related emotions and gender.

A speaker identification (SI) system employs features that distinctively stand for the features of individuals. SI system has several purposes; such as identifying individuals. In addition, by embedding the model of speakers into a chip, the speaker voice could be considered as an access tool for the uniqueness of users. Furthermore, speakers can be identified through any communication system since speech signals are transferred via communication channels. Over the past few decades, a significant study has been carried out on the identification of speakers under different conditions. Hence, many techniques have been proposed to solve SI problem such as factor analysis [1], Gaussian Mixture Model (GMM) [2], vector quantization (VQ) [3], linear prediction cepstral coefficient (LPCC) [4]. Nonetheless, the efficiency of these approaches significantly degrades in the case of channel variations and background noise. In addition, an important crisis facing the SI system is the data size. Let say that the actual time required to evaluate the voice of an unidentified speaker is 2 seconds, consequently, more than 48 hours is needed to recognize this speaker from 10,000 other speakers. In SI process, MFCC and LPCC are comonly used to extract features [5, 6]. The mentioned MFCC and LPCC are to enhance the performance of SI process [7] [8]. For MFCC method, further researches are concentrating on (i) the type of window used, (ii) sufficient numbers of coefficients, (iii) the number of adaptive filter banks and finally (iv) the outcome of implementing additional coefficients [9]. Many classifiers that can be used for dependent system identification such as back propagation, Multilayer perception and RBFNN [10], [11], [12]. A few number of investigations have been carried out to identify and verify Arabic speakers [17]. The main reason behind this lack of research lied on the bounded Arabic databases. The speaker identification scheme for Arabic words is implemented by utilizing GMMs [18], [19]. SI system for Arabic speakers is designed by combining the Linear Predictive Coding (LPC) and discrete Wavelet Transform techniques for extraction of speech features [20], [21].

This paper is to propose an efficient, accurate and less complex SI approach for the Arabic language speakers which tackles the weakness of the above mentioned conventional approaches. The efforts of this work can be classified into three main parts:

- Firstly; a single Arabic word data-base is built, this data base captures all potential voice phonemes.
- Secondly; MFCC is to extract features.

- Thirdly; RBFNNs are to identify the speaker.

In the proposed system, the MFCC is used to extract the speech feature and the RBFNN is used to classify the obtained feature extraction. The performance of SI systems is improved by enhancing the signal to noise ratio and reducing the environmental distortion. This can be done by using different pre-processing techniques such as removing the DC-offset and Pre-emphasize filters. The results will prove the effectiveness and precision of the proposed system.

This paper is arranged as follows: Section 2 explains the structure of the SI system, Section 3 presenst an detailed analysis of the proposed SI system, Section 4 analyses results in details, Finally, Section 5 concludes the paper.

II. The Proposed Speaker Identification System

The role of SI is to find the identity of an unknown speaker, this process links the speach utterance of the unidentified speaker with the utterances of registered speakers in the database.

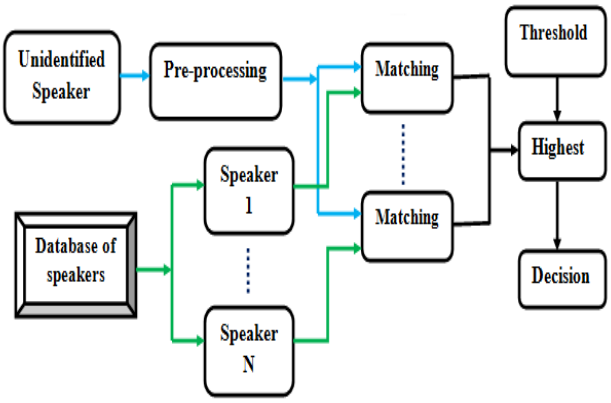

Fig. 1 The SI system

Figure 1 demonstrates the fundamental construction of spokesperson identification system. From Figure 1, it can be seen that a single Arabic word from an unidentified speaker is preprocessed and matched to speaker signal in the database of speakers. The unidentified speaker is identified as the one whose voice has the highest matches with the speaker voice. SI system includes multiple matching between the type of unidentified speaker and the type of speakers which is saved in the speaker database.

In SI technology, the system needs to learn the speaker voice patterns. Hence, the users are asked to enroll in the system to provide samples of their voices. SI system could be classified into two types: Text-Independent Speaker identification (TI-SI) and Text-Dependent Speaker identification (TD-SI). TD-SI needs the speaker to say precisely the word that has been enrolled. TI-SI is the method that identifies a speaker without any constraint on the content of the speech. Furthermore, SI can be categorized into open set and closed set according to the bunch of spokespersons who utilize the system. In the closed set scenario, the spokesperson is incorporated in the training database; in this scenario the highest matching outcome supposed to be sufficient to formulate the right decision without using the threshold in the process. In open set scenario, the speaker may arrive from outer training sets. Hence, the optimal value of threshold should be chosen to keep security and effectiveness of the system. In this paper, The TD-SI system is implemented and examined to assess its efficiency through a closed sets scenario.

III. Proposed SI System

The suggested system is made up of four major phases (see Figure 2):

- Phase I: the speech data are collected by recording the voice of different people. The chosen word in Arabic is (شكرا) which means “Thank you”.

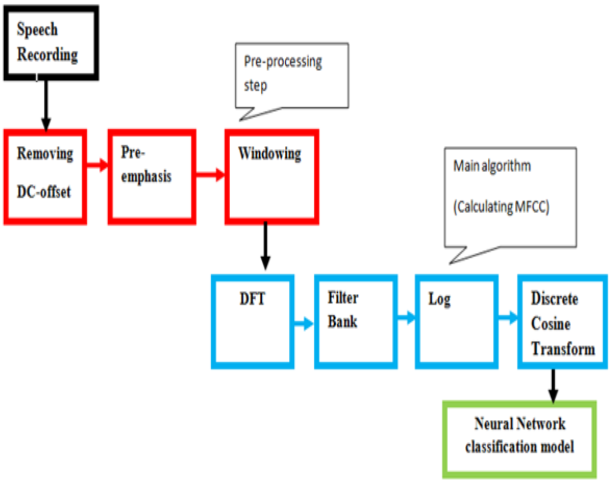

Fig. 2 Block diagram of proposed identification algorithm

It can be shown from Table 1 that the word (شكرا) was recorded twenty-eight times by both male and female with diffrent ages. Hence the speech data set for the system is built.

- Phase II is called preprocessing step; this phase is an essential part for any speaker identification system. Hence, this phase consists of three main parts:

(i) The first part is used to eliminate the DC offset from the recorded data as shown in Figure 3. Usually the recorded signal can be DC-offset from its zero point; this can be caused difficulties in time domain processing. To solve this problem, the dc-offset is eliminated by calculating the mean of the signal and take off the mean successfully from the signal as shown in equation (1).

$$s = \frac{s-\bar{s}}{\max[|s-\bar{s}|]} \qquad (1)$$

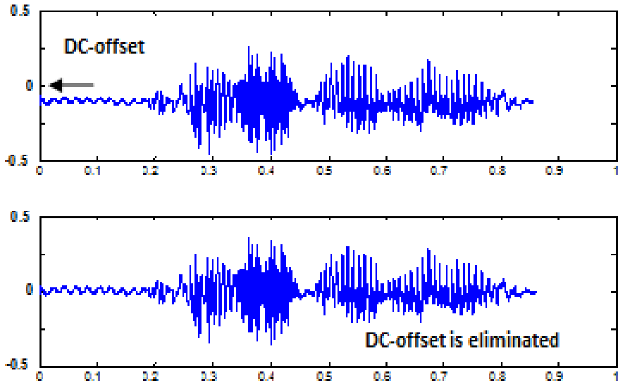

Fig. 3 DC Offset elimination

Table 1. Speaker's data

	Gender	Age			Gender	Age
1	F	30		15	F	18
2	F	25		16	F	21
3	M	22		17	M	34
4	M	30		18	M	36
5	M	50		19	M	30
6	F	18		20	F	23
7	F	36		21	F	34
8	M	17		22	F	16
9	M	22		23	M	38
10	F	26		24	M	44
11	F	20		25	M	31
12	F	38		26	M	45
13	M	42		27	F	36
14	M	40		28	F	27

(ii) The second step in preprocessing phase is called the Pre-emphasis step. In this step the first order FIR filter is used to put emphasis on the higher frequencies as the voice spectrum is reduced in high frequency area as shown in Figure 4.

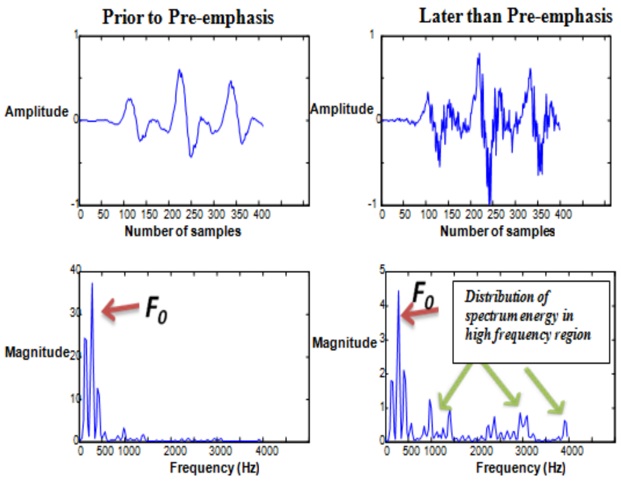

Fig. 4 Pre-emphasis Process

The FIR filter is expressed as follows:

$$H(z) = 1 - bz^{-1} \qquad 0 < b < 1 \qquad (2)$$

Where, parameter 'b' represents the pre-emphasis parameter that manages gradient of filter.

(iii) The third part of preprocessing phase is called windowing. The main idea behind widowing is to reduce the spectral distortion by utilizing different type of window like rectangular, Hanning, Hamming and Blackman window. The Hamming and Hanning window are mainly used in SI system (see Figure 5), these windows are expressed by u (n).

$$u(n) = 0.53836 - 0.46164\cos\left(\frac{2\pi n}{N-1}\right) \qquad (3)$$

$$u(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right) \qquad (4)$$

Where equation 3 and 4 represents Hamming and Hanning Windows respectively, where N represents the number of samples and u(n) has an interval between 1 and N (1<n<N).

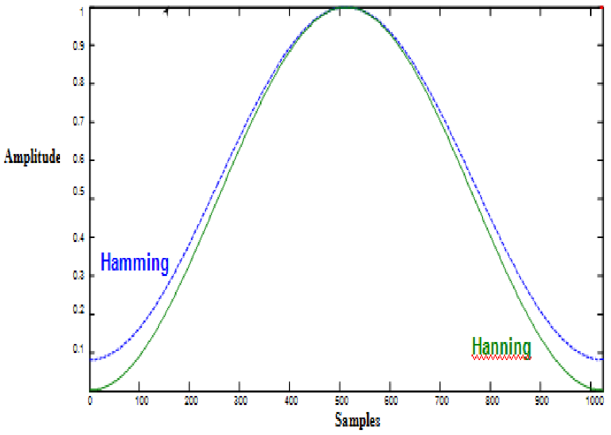

Fig. 5 Hanning and Hamming function in time domain

Then the windowed signal is stated as follow

$$y(n) = s(n)u(n) \qquad (5)$$

- Phase III in the block diagram is the most important part where the features of the sound are extracted by calculating their MFCC. MFCC is obtained as a result of calculating the logarithm of the discrete frequency transform of the windowed speech signal. The central dissimilarity between the MFCC and LPCC can be stated as follow: the MFCC has the ability to chart the logarithmic range to the Mel range by utilizing triangular windows. A filter bank is used to build up Mel-spectrum, with minimum one filter per each required Mel-frequency component. These filters have triangular shape and applied in frequency domain. Hence, it has band pass frequency response as shown in Figure 6.

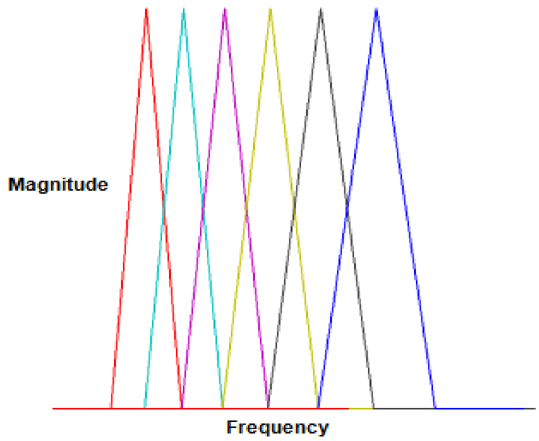

Fig. 6 Filters bank

In order to calculate MFCCs, the final step is to implement Discrete Cosine Transform (DCT).

- Phase IV applies RBFNN as a classification model. The RBFNN contain three layers as shown in Figure 7. The first layer corresponds to the input layer where the Frequency Cepstral Coefficients is inserted. The second layer represents the hidden layer which applies radial basis functions. The third layer corresponds to the output layer, it implements linear summation functions.

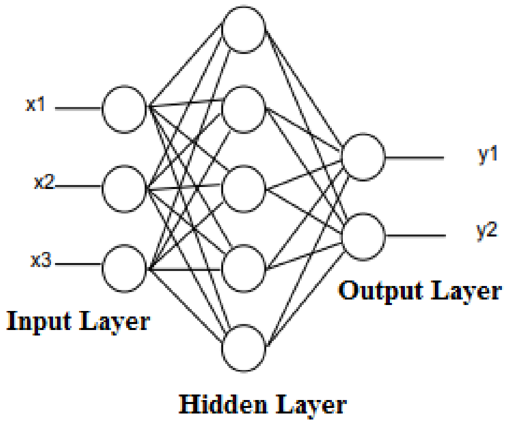

Fig. 7. RBFNN Structure

IV. Result and Discussion

A database for Arabic speakers is produced in order to assess the system efficiency and accuracy as shown in Table 1. An Arabic word (شكرا) is chosen to be recorded for each speaker within a two-second period. Figure 8 shows some samples of recorded speech signals, where the x-axis denotes time which is two seconds and y axis denotes amplitude.

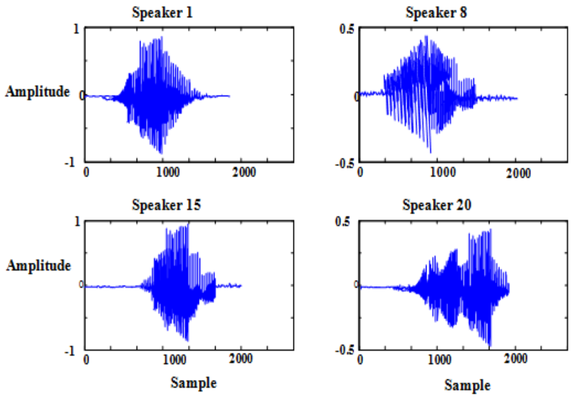

Fig. 8 Samples of recorded speech signal

The system parameters are initiated as shown in Table 2, where, the amount of enroll speakers is 14, a single Arabic term has been recorded for every speaker 25 times within a period of two seconds. The sampling rate is 32,000 samples/sec. The dc-offset is applied to all recorded speeches. In each 8 ms, a Hamming window with 64 ms length is performed on speaker signals. For every frame, 26 Mel filters are employed to take out 28 MFCC.

Table 2 System Parameters

Sampling rate	32000 sample/sec
The Pre-emphasis filter	$H(z) = 1\text{-}0.95Z^{-1}$
The length of Hanning window	64 ms, 1024 samples
Overlapping period of window	256 samples
Recording time	2 seconds
Number of enroll speaker	14 speaker
Number of records for word (شكرا) for each speaker	25 times
Number of records used for training for each speaker	15 records
Number of records used for testing for each speaker	10 records
Number of Mel-filters	26
Number of MFCC	28

In this experiment, fourteen speakers are enrolled to the system. The input matrix M to speaker identification system has P x N dimension. The row P represents the number of MFCC (P=28). The column N represents the speaker's numbers multiplied by the number of records for the word (شكرا) - which is used for training per speaker. Hence, N=14x15=210. The RBFNN receives the matrix (M) as an input matrix as shown in Figure 9.

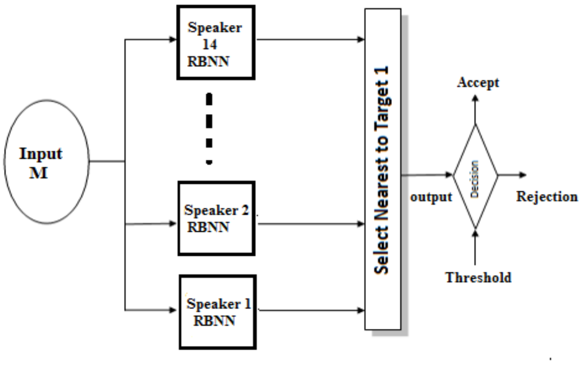

Fig. 9 Speaker identification system

The target (Gi) is represented as a matrix with dimension [1 x N], where parameter (i) represents the amount of speakers.

The target $G_{i,j}$ for the jth element is represented as:

$$G_{i,j} = \begin{cases} 1\,, & ((i-1)*15+1) \leq j \leq ((i*15)+1) \\ -1, & \text{otherwise} \end{cases} \qquad (6)$$

Where number fifteen related to the number of records used for training for each speaker. Hence, for the first speaker, the first fifteen elements have values of 1, and -1 for otherwise elements. In case of second speaker, the second fifteen elements (j =16 → 30) have values of one etc. For each speaker, the lasting ten records are utilized for testing. RBNN is employed for testing every record. It is worth to mention that in the case of trained networks, RBNN operates more robustly than conventional NNs especially when the speech input data set is corrupted with noise. Every RBNN gives an output, the close output to one is chosen by the system. If the space between the close output and one is fewer than one, the determination of SI system believes that the desired speaker is the holder of the NN. However, if the space is more than one, the desired speaker is rejected. (The threshold of acceptance = 1).

To determine the accuracy of the proposed identification system, a test is performed 20 times for specific spread values from 1 to 20. It is shown from Figure 10 that when the spreading values in the range from one to three, the accuracy rate increases from 20% to 92% respectively, then it becomes 97.5 % percent. This means that the accuracy rate of identification process is 97.5%. Hence, the proposed IS system is accurate and efficient comparing with conventional SI methods.

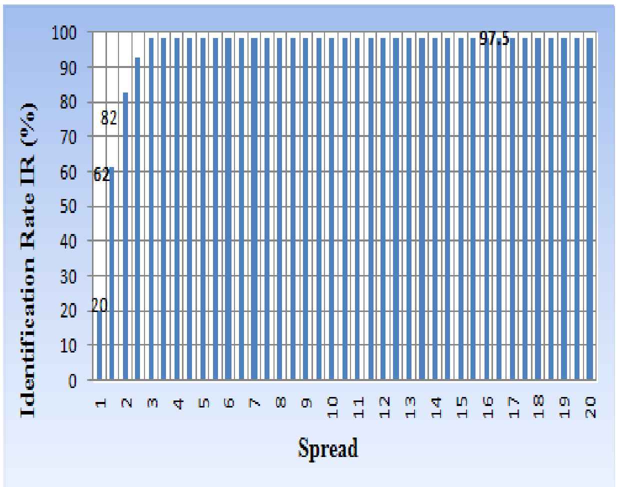

Fig. 10 Identification rate for Closed Set

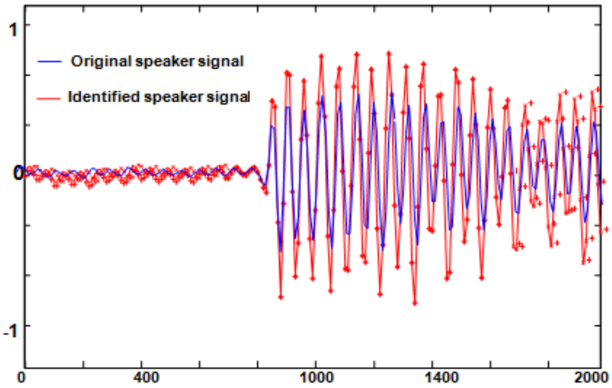

Fig. 11 identified speaker signal compared with original

Figure 11 shows the identified speaker signal compared with original signal. It can be shown from Figure 11 that the identification system has the ability to recover and identify the original speaker signal. The blue signal represents the original speaker signal which is saved in the speakers' database, where the red signal represents identified signal by the speaker identification system.

V. Conclusions

In this paper, speaker identification system for Arabic language using NNs is implemented. A database for Arabic speakers of a single word (شكرا) is used in IS. The pre-processing step was implemented to improve the system's accuracy. MFCC were extracted as feature parameters to characterize the voice of Arabic speakers. The Arabic desired speaker is identified using RBFNN as a classifier. Every speaker had a certain RBFNN to categorize his/her speaker signal. Hence, the assessment is made by choosing the adjacent output to certain targets. Comparing with conventional methods, the accuracy of the proposed method is 97.5%.

.

Acknowledgment

This work is supported by Al-Balqa Applied University during the period of sabbatical leave in the Private Applied Science University.

References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, " Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788 - 798, May. 2011.

[2] A. Reynolds, F. Quatieri, B. Dunn," Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 3, pp. 19–41, Jan 2000.

[3] K. Soong, E. Rosenberg, H. Juang, R. Rabiner, "Report: A vector quantization approach to speaker recognition," *AT&T technical journal*, vol. 66, no. 2, pp. 14–26, April 1987.

[4] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–80, April 1975.

[5] Z. Bhotto, R. Amin, "Bengali Text Dependent Speaker Identification using Mel-Frequency Cepstrum Coefficient and Vector Quantization," in 3rd International conference on Electrical and Computer Engineering, Dhaka, Bangladesh, 2004, pp.569-572.

[6] N. Balaska, Z. Ahmida, A. Goutas, "Speaker Recognition Using Artificial Neural Networks: RBFNNs vs. EBFNNs," in 4th International Conference on Computer Integrated Manufacturing, Algeria, 2007.

[7] B. Davis, P. Mermelstien, "Comparison of prarmetric representations for Monosyllabic word recognition in continuously spoken sentences," IEEE transactions on Acoustic, Speech & Signal Processing, vol. 28, no. 4, pp. 357 -366, Aug 1980.

[8] A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Transactions on Speech & Audio Processing*, vol. 2, no. 4, pp. 639 - 643, Oct 1994.

[9] M. Hassain, B. Ahmed, M. Asrafi, "A Real Time Speaker Identification using Artificial Neural Networks," in *10th international conference on computer and information technology*, Dhaka, Bangladesh, 2007.

[10] S. Al-Dahri, H. Al-Jassar, Y. Alotaibi, M. Alsulaiman, K. Al-Mamun, "A Word Dependent Arabic Speaker Identification System," in *IEEE International Symposium on Signal Processing and Information Technology*, Ajman, United Arab Emirates, 2008, pp.198-202.

[11] K. Yiu, M. Mak, S. Kung, "A Comparative Study on Kernel-Based Probabilistic Neural Networks for Speaker Verification," *International Journal of Neural Systems*, vol. 12, no. 5, pp. 381-391, 2002.

[12] S. Tirumalaa, S. Shahamiria, A. Garhwala, R. Wangb, "Speaker identification features extraction methods: A systematic review," *Expert Systems with Applications*, vol. 90, pp. 250-271. Dec 2017.

[13] J. Raitoharju, S. Kiranyaz, M. Gabbouj, "Training Radial Basis Function Neural Networks for Classification via Class-Specific Clustering," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 12, pp. 2458 -2471, 2015.

[14] N. Wang, L. Wang, "Robust speaker recognition based on multistream features," in *IEEE International Conference on Consumer Electronics*, China, 2016, pp.1-4.

[15] P. Sale, S. Jainar, B. Nagaraja, "A Comparison of Features for Multilingual Speaker Identification - A Review and Some Experimental Results," *International Journal of Recent Technology and Engineering*, vol. 7, no. 4, pp. 299-304, Dec 2018.

[16] Y. Wang, B. Lawlor, "Speaker recognition based on MFCC and BP neural networks," in *28th Irish Signals System Conference*, Ireland, 2017, pp.1-4.

[17] M. Alsulaimana, A. Mahmoodb, G. Muhammada, "Speaker recognition based on Arabic phonemes," *Speech Communication*, vol. 86, pp. 42–51, Feb 2017.

[18] M. Alkanhal, M. Alghamdi, Z. Muzaffar, "Speaker verification based on Saudi accented Arabic database," in *9th International Symposium on Signal Processing and Its Applications*, Sharjah - United Arab Emirates, 2007, pp.1-4.

[19] S. Nidhyananthan, R. Kumari, "Language and Text-Independent Speaker Identification System Using GMM," *WSEAS Transactions on Signal processing*, vol. 9, no. 4, pp. 185-194, Oct 2013.

[20] S. Shah, S. Ahsan, "Arabic speaker identification system using combination of DWT and LPC features," in *International Conference on Open Source Systems & Technologies*, Pakistan, 2014, pp. 176-181.

[21] K. Daqrouq, A. Morfeq, M. Ajour, A. Alkhateeb, "Wavelet LPC With Neural Network for Speaker Identification System," *WSEAS Transactions on Signal processing*, vol. 9, no. 4, pp. 216-226, Oct 2013

Aws Al-Qaisi is an associated professor in Communication Engineering Department, Al-Balqa Applied University, Jordan. Al-Qaisi was received his PhD and MSc in communication and signal processing from Newcastle university in 2006 and 2010 respectively. Dr. Aws research interest includes Digital signal processing, image processing, Feature Extraction, Wireless communication, Digital communication. Dr Al-Qaisi is a member of IEEE executive committee in Jordan section as the industrial relations coordinator from March/2020 till Dec/2021. He served as reviewer in many international journals.
Email: aws.alqaisi@bau.edu.jo