A Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-Signal Image

Khalid Saeed, *Member, IEEE*, and Mohammad Kheir Nammous

***Abstract*—This paper discusses a speech-and-speaker (SAS) identification system based on spoken Arabic digit recognition. The speech signals of the Arabic digits from zero to ten are processed graphically (the signal is treated as an object image for further processing). The identifying and classifying methods are performed with Burg's estimation model and the algorithm of Töeplitz matrix minimal eigenvalues as the main tools for signal-image description and feature extraction. At the stage of classification, both conventional and neural-network-based methods are used. The success rate of the speaker-identifying system obtained in the presented experiments for individually uttered words is excellent and has reached about 98.8% in some cases. The miss rate of about 1.2% was almost only because of false acceptance (13 miss cases in 1100 tested voices). These results have promisingly led to the design of a security system for SAS identification. The average overall success rate was then 97.45% in recognizing one uttered word and identifying its speaker, and 92.5% in recognizing a three-digit password (three individual words), which is really a high success rate because, for compound cases, we should successfully test all the three uttered words consecutively in addition to and after identifying their speaker; hence, the probability of making an error is basically higher. The authors' major contribution to this task involves building a system to recognize both the uttered words and their speaker through an innovative graphical algorithm for feature extraction from the voice signal. This Töeplitz-based algorithm reduces the amount of computations from operations on an $n \times n$ matrix that contains n^2 different elements to a matrix (of Töeplitz form) that contains only n elements that are different from each other.**

***Index Terms*—Communication, humatronics, linear predictive coding, processing and recognition, speaker recognition, speech analysis, Töeplitz matrix (TM) eigenvalues, understanding speech.**

I. Introduction

Voice recognition systems are, in general, very useful in many tasks. Among those very important applications in our everyday life are secure telephony, voice-based login, and voice locks. They are also used as a security key—we can use the voiceprint of every human being [1]. That is why voice recognition (both speech and speaker) plays its significant role in the field of human electronics (humatronics) and its wide applications.

We can classify speaker recognition into text-dependent and text-independent methods. The former requires that the speaker should provide utterances of the same text for both training and recognition, while the latter one does not depend on the specific text being spoken.

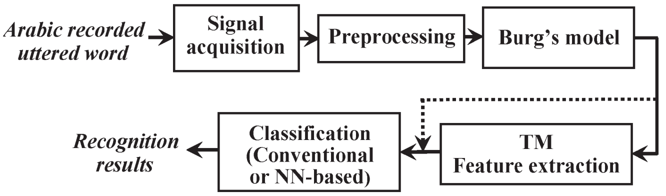


Fig. 1. Hybrid system suggested in this paper for speaker and speech recognition. The uttered word is recorded to have its signal for preprocessing. Then, MATLAB [6] is used to perform the needed computations and to evaluate Burg's model and its necessary graph. This, in turn, forms the input to Töeplitz matrix (TM) minimal eigenvalues algorithm whose output is fed to the neural networks (NNs) for classification. Some experiments, as explained in the paper and for the purpose of comparison, were performed without TM (follow the dashed line); Burg's model results were simply fed directly to the classification stage with either conventional methods or NN approaches.

Speaker recognition can also be classified into either speaker identification or speaker verification. *Speaker identification* is the process of determining one of the registered speakers from whom the given utterance comes, while *speaker verification* is the process of accepting or rejecting the identity claim of a speaker.

Speaker identification can then be categorized into open set or closed set. Open-set identification means that the system has to identify data from classes that are not a part of the training set data (the closed set). The problem of the open-set speaker identification is therefore similar to that of speaker verification [2].

This paper is categorized into a closed-set-based text-dependent speaker identification applied to spoken Arabic digits from zero to ten. Therefore, we will focus the research on the development of a text-and-its-speaker identification system, which means that we are recognizing the right words and identifying their speaker. In the suggested system, we treat the speech signal graphically [3]. Then, we apply TMs [4] to describe the speech signal characteristics by evaluating a sequence of minimal eigenvalues of these matrices [5] to form a feature vector for each signal image. Before entering this stage, the signal image needs some steps in speech preprocessing where Burg's model [6], [7] (the frequency spectral estimation method, based on the *linear predictive coding* principle [8], [9]) is applied. Burg's model is built on the idea of prediction error minimalization [9], [10] and is explained in detail together with its software and computer implementation in [6]. The obtained signal spectrum forms the basis to the analysis by the Töeplitz

Manuscript received January 6, 2006; revised August 21, 2006. Abstract published on the Internet January 14, 2007. This work was supported by the Rector of Bialystok Technical University under Grant W/WI/3/04.

The authors are with Bialystok Technical University, 15-351 Bialystok, Poland (e-mail: aida@ii.pb.bialystok.pl; nammous@mail.sy).

Digital Object Identifier 10.1109/TIE.2007.891647

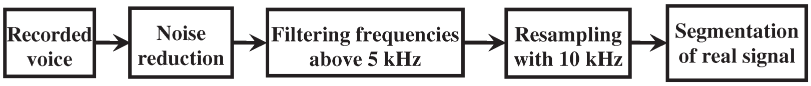

Recorded voice
Noise reduction
Filtering frequencies above 5 kHz
Resampling with 10 kHz
Segmentation of real signal

Fig. 2. Preprocessing procedure diagram used in the authors' work for speech signal preparation for the purpose of feature extraction and classification. All needed computations were made in MATLAB.

approach. Afterward, the Töeplitz-based feature vector enters the classification stage. For the purpose of classification, the authors mainly use probabilistic and radial basis function (RBF) NNs. Fig. 1 shows the whole procedure and the stages of speech signal recognition according to that presented in this paper system.

The succeeding sections will introduce the details of each stage together with some theoretical explanation wherever applicable.

II. Signal Preprocessing

The standard format used by the authors is pulse-code modulation, with a frequency of 22 050 Hz, 16-bit mono. Each file contains only one voice with a silence region before and after the right signal. The details of the continuous-speech segmentation techniques are beyond the scope of this paper. After resampling, the real signal is segmented and forwarded to the stage of Burg's model preparation. The process of the speech preparation for feature extraction is given in Fig. 2.

This and more information about how to prepare the signals for processing can be found in [8]. Now, the new signal is ready for further processing. The next step is feature extraction.

III. Feature Extraction

Among many methods of speech signal processing, the authors have chosen the method based on spectrum analysis [9], [10]. This method contributes to the speech-image feature extraction accomplished by spectral analysis. The authors' experiments showed that the power spectrum estimation of Burg's model (Fig. 3) is one of the best methods for smoothing irregular spectral shape resulting from applying the fast Fourier transform (FFT) and the linear predictive coding approach [8]–[10].

Now, we can use the obtained power spectrum acoustic images directly, or we can apply the algorithm based on minimal eigenvalues of TMs [8], [11] to analyze these acoustic images. When using Burg's method of estimation, however, we need to specify the prediction order P and the FFT size, which is called the length of FFT $(NFFT)$. The FFT length must give the smoothest shape of the spectrum (the more samples we have, the smoother shape we get), and it cannot be a case where too many samples are considered. This, as very well known, would definitely lower the efficiency of the algorithm. Prediction order is also an important parameter. When it is too low, the envelope does not match with the FFT shape, and when it is too high, it causes a decrease in the speed of the algorithm. Thus, it is very important, although very difficult, to choose the best prediction order. This had already been proven and shown in some previous work [8], where more explanations and details about Burg's method are explained. The TM minimal eigenvalues algorithm in its model for image description and feature extraction, however, is given in [11]. For convenience, a brief description of the Töeplitz approach for obtaining the signal-image feature vector is given here.

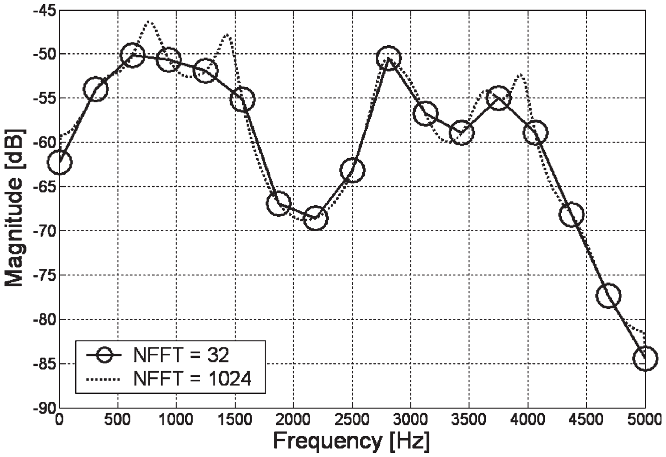


Fig. 3. Spectral analysis for Burg's method with $NFFT = 32, 1024$ and $P = 20$. The circles show the places of the characteristic points accomplished by Burg's model. Notice that both values of $NFFT = 32$ and $NFFT = 1024$ give almost the same features. The circles represent the start and end points for each of the straight lines forming the solid estimating curve. All computations and graphs were performed in MATLAB.

A. TM Minimal Eigenvalues Algorithm

This algorithm has shown its successful performance in object-image recognition [11]. The success rate of machine-typed script recognition by this algorithm reached 99%, while for the case of more complicated handwritten and cursive scripts, the algorithm is under development, although a success rate of about 97% had already been achieved [11]. In signal recognition, however, the work on TM minimal eigenvalues theory has quickly been developed, although the success rate is still not as high as with that of object-image applications. This comes from two facts: The first lies in the complicated nature of voice and speech signals, and the second is the short age of the application of the theory of minimal eigenvalues to speech recognition [8]. Nevertheless, applying Töeplitz and radial NNs gave a 95.82% successful recognition for single Arabic words spoken by people selected from different Arabic countries (of different accents) at different ages and gender [3]. The main advantage of the Töeplitz approach lies in its elasticity of fusing with other tools in a hybrid system [12]. In this section, a brief discussion of the theory is given.

The main idea is to obtain the rational function in

$$f(s) = \frac{x_0 + x_1 s + x_2 s^2 + \cdots + x_n s^n + \cdots}{y_0 + y_1 s + y_2 s^2 + \cdots + y_n s^n + \cdots} \qquad (1)$$

with its numerator and denominator coefficients being the coordinates (or other alternative parameters derived from the coordinates—remarks in B of this section) of the feature points extracted from the Burg's spectral analysis graph of the voice image in Fig. 3. The number of considered feature points is n, and they are the points marked with circles in Fig. 3.

Dividing the numerator by the denominator of $f(s)$ in (1), a power series, e.g., Taylor series, is obtained. The coefficients of the resulting series in the following expression can very easily be expressed [13] by the coefficients of the polynomials in the numerator and denominator of $f(s)$:

$$T(s) = \alpha_0 + \alpha_1 s + \alpha_2 s^2 + \cdots + \alpha_n s^n + \cdots \tag{2}$$

where [17]

$$\alpha_0 = \frac{x_0}{y_0}$$

$$\alpha_i = (y_0)^{-i-1} \cdot \begin{vmatrix} x_i & y_1 & y_2 & y_3 & \cdots & y_i \\ x_{i-1} & y_0 & y_1 & y_2 & \cdots & y_{i-1} \\ x_{i-2} & 0 & y_0 & y_1 & \cdots & y_{i-2} \\ x_{i-3} & 0 & 0 & y_0 & \cdots & y_{i-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0 & 0 & 0 & 0 & \cdots & y_0 \end{vmatrix}, \qquad \text{for } i = 1, \ldots, n.$$

Then, TMs are formed from these coefficients, i.e.,

$$A_0 = \alpha_0 = \frac{x_0}{y_0}, \qquad A_1 = \begin{bmatrix} \alpha_0 & \alpha_1 \\ \alpha_1 & \alpha_0 \end{bmatrix}, \ldots$$

where the general form of the TM for real numbers is given by

$$[A_i] = \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_i \\ \alpha_1 & \alpha_0 & \cdots & \alpha_{i-1} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_i & \alpha_{i-1} & \cdots & \alpha_0 \end{bmatrix} \tag{3}$$

for $i = 1, \ldots, n$ and assuming that $\alpha_{-i} = \alpha_i$.

Now, the minimal eigenvalues $\lambda_{\min}^{(i)}$ of these TMs are evaluated for $i = 1, \ldots, n$, in such a way that, for each submatrix, the ith minimal eigenvalue is computed as follows.

Therefore,

$$A_0 = \alpha_0 = \frac{x_0}{y_0} \rightarrow \lambda_0,$$

$$A_1 = \begin{bmatrix} \alpha_0 & \alpha_1 \\ \alpha_1 & \alpha_0 \end{bmatrix} \rightarrow \lambda_1,$$

$$A_2 = \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 \\ \alpha_1 & \alpha_0 & \alpha_1 \\ \alpha_2 & \alpha_1 & \alpha_0 \end{bmatrix} \rightarrow \lambda_2$$

and so on.

For the $k \times k$ submatrix,

$$A_k = \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \cdots & \alpha_k \\ \alpha_1 & \alpha_0 & \alpha_1 & \cdots & \vdots \\ c_2 & \alpha_1 & \alpha_0 & \vdots & \alpha_2 \\ \vdots & \vdots & \vdots & \ddots & \alpha_1 \\ \alpha_k & \cdots & \alpha_2 & \alpha_1 & \alpha_0 \end{bmatrix}_k \rightarrow \lambda_k.$$

The feature vector is formed from these eigenvalues. It is given by

$$\Phi_i = (\lambda_0, \lambda_1, \lambda_2, \ldots, \lambda_n). \tag{4}$$

The elements of the sequence in (4) have been proven [3], [11], [14]–[16] to form a monotonically nonincreasing series, as shown in the following:

$$\lambda_0 \geq \cdots \geq \lambda_i \geq \cdots \geq \lambda_n, \qquad \text{for } i = 1, \ldots, n. \tag{5}$$

The theories in [5] and [13]–[17] as well as heuristic methods and experiments [3], [5], [8], [11], [16] have shown that in signal-image description and processing, (5) is as feasible and practical as in digital filter realization applications. In speech-signal image processing, the feature vector Φ_i in (4) with the characteristics in (5) introduces a distinguished plot for a given signal image within the same family class of signals. Its role in the system is to act as the input data to the classifying algorithms when applying the known methods of similarity and comparison for the sake of recognition. It has experimentally been shown [3], [8] that, in a class of voices, an individual voice feature vector is quite unique and has its independent series of minimal eigenvalues in (4) among other signals within the tested class.

B. Remarks on Taylor Series Coefficients—Feature Extraction

Experiments have shown that the transformation from (1) to (2) requires long and costly calculations. The mathematical evaluation of Taylor-series coefficients simply proves this conclusion [13], [17]. Therefore, unless (1) fulfills the geometric parameters of the image [5] (the signal image here), there exist other alternatives to define the coefficients of the Taylor series. This is followed particularly when certain classes of images need a higher number of these coefficients. Here are some of these alternatives. Nevertheless, it is worthwhile to notice that none of them is better or preferable. All of them have been proven experimentally, and therefore, it is a matter of experience and time calculation for each case.

1) Alternative 1—Modulus (Absolute Value): After evaluating the coordinates of the feature points, (2) is achieved by considering the coefficients $\alpha_i^{'s}$ as the modulus $\sqrt{x_i^2 + y_i^2}$, where $x_i^{'s}$ and $y_i^{'s}$ are the coefficients of the numerator and denominator of function $f(s)$ in (1), respectively. Hence,

$$\begin{aligned}\alpha_0 &= |r_0| = \sqrt{x_0^2 + y_0^2}\\ \alpha_1 &= |r_1| = \sqrt{x_1^2 + y_1^2}\\ &\vdots\\ \alpha_n &= |r_n| = \sqrt{x_n^2 + y_n^2}.\end{aligned}$$

2) Alternative 2—Successive Modulus Differences: This method considers the Taylor coefficients in (2) as the differences between the successive vector lengths r_i in the following manner: $\alpha_0 = |r_0| - |r_1|, \alpha_1 = |r_1| - |r_2|, \ldots, \alpha_n = |r_n|$.

In some cases, it is needed to interchange α_0 and α_n : $\alpha_0 = |r_0|, \alpha_1 = |r_0| - |r_1|, \ldots, \alpha_n = |r_{n-1}| - |r_n|$.

3) Alternative 3—Polar Representation: Another interesting alternative is by considering the *polar form* of the xy-coordinates representation. Therefore, for $i = 1, \ldots, n$, we have $r_i = |r_i|e^{j\varphi}$, where $|r_i| = \sqrt{x_i^2 + y_i^2}$ and $\varphi_i = \tan^{-1}(y_i/x_i)$.

Having the quantities $|r_i|$ and φ_i, two possibilities for Taylor coefficients then exist. The first possibility is to have $\alpha_i = |r_i|$, while the second one is to consider $\alpha_i = \varphi_i$, i.e., $\tan \varphi_i$ can also replace $\alpha_i^{'s}$ for $i = 1, \ldots, n$.

The current experiments are being conducted to maximally make use of the alternative that presents the least sensitiveness to changes. Although each of them has its applications where it serves better than others, it seems that the third one is a better alternative than the original approach. This comes from the fact that not only does the polar form give the system strong invariance to transformations such as rotation, scaling, or shear (shear is usually the basic factor responsible for changing normal writing to a cursive one) but it also assures insensitiveness to local changes such as the absence of a characteristic point. All these factors and other additional specific ones are demonstrated experimentally with the use of a large number of examples given in [5], [11] and many other published researches of the authors and their research team. The results have shown proper performance.

C. From Voice Recognizer to Speaker Identifier

We have directed our previous researches on voice recognition [3] to the situation of recognizing the speaker himself. Before presenting the results we have achieved and the high efficiency of the implemented security system, we will first prove experimentally that the algorithm is valid for speaker identification as well as for his speech.

First, we present Burg's spectral graph of the word *Ashr* (ten) uttered by three speakers (Fig. 4).

We can see that each curve in Fig. 4 has a specific shape that differs from the other two, which makes the identification process theoretically possible [8], [11]. One would recall that all of the three graphs, from the other side, represent the same word but are spoken by three different people. One may ask then how to distinguish between them; the answer is simply in that although the shapes of these curves are *almost* similar and, therefore, furnish a similar feature vector for the same word (*Ashr*), which in turn shows the characteristic behavior *differing much* from all the other digits' feature vectors, they (the curve shapes) belong to different speakers. This is because they differ from each other in only some features (coming from the voice origin) that are never absolutely similar to each other or to other digits' characteristics. These *almost*-similar features are the key for speaker identification. Therefore, the small differences between the curves in Fig. 4, such as the positions of the maxima, minima, and points of inflection, form the essence of the algorithm for distinguishing between the speakers' curves. The techniques followed are similar to those used in the graph presentation for speech (uttered words) recognition. The details are given in [8].

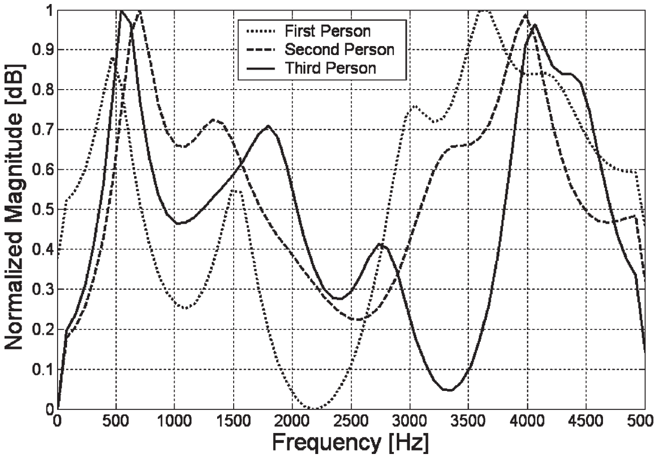


Fig. 4. Burg's curves for the Arabic word *Ashr* (one of the grammatically possible pronunciations of the digit *ten*) spoken by three different people. The signals are performed for the purpose of speaker identification. The small differences between the curves (see the positions of the maxima, minima, and points of inflection) form the basis for distinguishing between speakers. This is done in a similar way to the graph presentation for speech (uttered words) recognition given in [8]. All computations are performed in MATLAB.

The acquired data from the curves of Fig. 4, after computing the feature vector by the TM minimal eigenvalues algorithm [11], will enter the stage of classification (conventional or NN-based ones). The succeeding two sections will describe all the preceding issues demonstrating exactly how the system works.

IV. Classification

The authors have chosen two simple methods for classification: classical and neural-based ones. Both methods have their input data from the *TM minimal eigenvalues algorithm* (or from Burg's curve directly)—the data being extracted from the signal images shown in Fig. 3. As stated before, conventional (classical) classifying methods were used. They imply matching the given feature vector with the mean vector of each class; having 20 speakers with 11 digits means that we have 220 different classes, and the given vector is classified to the most similar one from the training set. Radial and probabilistic NNs (PNNs) were used as the second method of classification, showing better performance in less time-consuming work, as will be shown in the succeeding sections of this paper. During the training stage, we use all the samples from the training set. At the stage of classification, we have 11 classes for voice recognition (digits from zero to ten) and 20 for speaker identification (the number of speakers in the authors' base). In the experiments, we use different NNs for the two purposes (voice recognition and speaker identification).

A. NNs Used in the Authors' Approaches

The power and utility of the artificial NNs have been demonstrated in several applications including speech synthesis, robotic control, signal processing, computer vision, and many other problems related to the category of pattern recognition. Generally, different kinds of NNs have been tested by the authors, showing promising results in achieving good performance over techniques of more traditional artificial intelligence character, especially when it comes to their use for the purpose of Arabic speech recognition [3], [18].

In this section, we will skip the details and present only a brief theoretical consideration of the used NNs [19], [20].

1) PNNs: In PNNs [19], there are at least three layers: input, radial, and output ones. The radial units are copied directly from the training data, one per case. Each of them models a Gaussian function centered at the training case. There is one output unit per class; each is connected to all the radial units belonging to its class, with zero connections from all other radial units. Hence, the output units simply add up the responses of the units belonging to their own class. The outputs are proportional to the estimates of the probability density functions of the various classes. The only control factor that needs to be selected for PNN training is the smoothing factor. This factor needs to be selected in such a way that it would only cause a reasonable portion of overlapping. An appropriate figure is easily chosen by experiment, by selecting a number, which produces a low selection error, and, fortunately, PNNs are not too sensitive to the precise choice of smoothing factor.

The greatest advantage of PNNs is the fact that the output is probabilistic, which makes the interpretation of the output easier and the training speed higher [19]. Training a PNN actually consists mostly of copying training cases into the network, and so, it is as close to the instantaneous value as can be expected. The greatest disadvantage is the network size: a PNN network actually contains the entire set of training cases and is therefore space consuming and slow to execute.

2) RBF NNs: An RBF network [20] has three layers: input, radial, and output layers. The hidden (radial) layer consists of radial units; each actually is modeling a Gaussian response surface. The units will always be sufficient to model any function. The RBF in each radial unit has a maximum of 1 when its input is 0. As the distance between the weight vector and the input decreases, the output increases. Thus, a radial basis neuron acts as a detector that produces 1 whenever the input is identical to its weight vector; additionally, there is a bias neuron, which allows the sensitivity of the radial basis transfer function of the hidden neurons to be adjusted. The standard RBF NN has an output layer containing dot product units with identity activation [20].

Radial basis networks may require more neurons than standard feedforward backpropagation networks, but often they can be designed in a fraction of the time it takes to train standard feedforward networks. They work best when many training vectors are available.

RBF networks have a number of advantages [20]. First, they can model any nonlinear function using a single hidden layer, which eliminates some design decisions about the number of layers. Second, the simple linear transformation in the output layer can be optimized fully using traditional linear modeling techniques, which are fast and do not pose problems. RBF networks can therefore be trained extremely quickly; training of RBFs takes place in distinct stages. The centers and deviations of the radial units must be set up before the linear output layer is optimized.

B. Base of Voices

For the sake of comparison of the results of voice recognition and speaker identification, in this paper, the authors use the same base used in their last works and experiments [3]. The base has recorded voices for 20 people. The total number of samples (5472) is divided into two groups; for each person and voice, we choose five samples to be the test set (1100 samples), while the remaining samples (4372 samples) are taken for the teaching set.

C. Suggested Speech-and-Speaker (SAS) Identifying System

The performance of the recognition system is given here. It is shown how the right voice and its correct speaker are identified from a spoken three-digit password. The developed model introduces a simple-to-use security system—it has two kinds of protection: the spoken digits and their speaker. The system identifies first the speaker and then the spoken password using only three spoken digits.

Here, we will study and evaluate the effectiveness of our algorithm by demonstrating only one spoken digit; hence, we have four possibilities for each spoken word.

1) Success rate—correct recognition of the spoken word and/or the speaker.
2) Miss word rate or miss speaker rate—wrong recognition of the word or speaker, respectively.
3) False speech rejection—correct speaker-voice identification but wrong spoken-word recognition.
4) Miss speaker identification—false rejection or false acceptance of the speaker but recognition of the right uttered word.

These possibilities will be dealt with through Experiments 3–6 showing both the success rate and the miss cases, explaining their practical meaning (false rejection, false acceptance, etc.).

Having these possibilities, we can now proceed to perform two kinds of human–computer communication:

Part a: Speaker identification. Here, we use the spoken words only to identify the speaker, without necessarily trying to recognize the spoken words. The efficiency of the algorithm will be higher in spite of

TABLE I
SPEAKER IDENTIFICATION BY CONVENTIONAL AND NN CLASSIFICATION METHODS. BURG'S GRAPH CHARACTERISTIC POINTS ARE TAKEN DIRECTLY FROM FIG. 3 TO THE CLASSIFICATION STAGE WITHOUT FURTHER FEATURE EXTRACTION BY TMs. ALL COMPUTATIONS ARE PERFORMED IN MATLAB

Parameters		Recognition Results, Classification by:					
		Conventional Method		Radial NN		Probabilistic NN	
The length of FFT [*NFFT*]	Prediction order [*P*]	Recognized samples in 1100	Success Rate %	Recognized samples in 1100	Success Rate %	Recognized samples in 1100	Success Rate %
32	12	985	89.54	689	62.64	1054	95.82
	20	1027	93.36	708	64.36	1069	97.18
	28	1021	92.82	674	61.27	1071	97.36
64	12	1006	91.45	678	61.64	1063	96.64
	20	1039	94.45	995	90.45	1077	97.91
	28	1041	94.64	1069	97.18	1080	98.18
128	12	1002	91.09	891	81.00	1063	96.64
	20	1038	94.36	1012	92.00	1077	97.91
	28	1034	94.00	1052	95.64	1084	98.54
256	12	1000	90.91	1036	94.18	1062	96.54
	20	1043	94.82	1039	94.45	1082	98.36
	28	1042	94.73	1026	93.27	1086	98.73
512	12	1001	91.00	815	74.09	1061	96.45
	20	1041	94.64	776	70.54	1083	98.45
	28	1036	94.18	979	89.00	1087	98.82
1024	12	1001	91.00	671	61.00	1061	96.45
	20	1038	94.36	552	50.18	1083	98.45
	28	1041	94.64	580	52.73	1087	98.82

the decrease in the security level. This is very important especially when we know that the algorithm is classifying the given voice from a concrete uttered word but, at the same time, without trying to verify if the word is the right one or not.

Part b: Multilevel security for both the spoken words and their speaker. Here, we use the spoken word as a password after identifying the speaker. This means identifying the speaker first and then using his spoken words as a password. The identification rate will be lower, while the level of security will be higher.

V. SYSTEM DEMONSTRATION—EXPERIMENTS AND RESULTS

In this section, we introduce the results of different experiments. For each case, we give the success rate as the percentage of efficiency of the suggested security system. This will be demonstrated through the verification of the two system parts (*Part a* and *Part b*) given in Section IV: the identification of only the right speaker (*Experiments* 1–2) and then the recognition of the right words after identifying their right speaker (*Experiments* 3–6). The experiments will show the performance of the suggested system for human–computer communication using different methods of signal-image description and feature extraction (Burg's and Töeplitz approaches) with both conventional and NN-based classifying tools.

TABLE II
SPEAKER IDENTIFICATION BY CONVENTIONAL AND NN METHODS APPLYING TM: THE BURG'S GRAPH CHARACTERISTIC POINTS ARE FIRST APPLIED TO THE TM ALGORITHM TO COMPUTE THE FEATURE VECTOR (MINIMAL EIGENVALUES) AND FEED THEM INTO THE CLASSIFYING SYSTEM. ALL COMPUTATIONS ARE PERFORMED IN MATLAB

Parameters		Recognition Results, Classification by:			
		Conventional Method		Radial NN	
The length of FFT [*NFFT*]	Prediction order [*P*]	Recognized samples in 1100	Success Rate %	Recognized samples in 1100	Success Rate %
32	12	663	60.27	896	81.45
	20	651	59.18	901	81.91
	28	586	53.27	852	77.45
64	12	829	75.36	1044	94.91
	20	849	74.27	1075	97.73
	28	795	72.27	1070	97.27
128	12	821	74.64	1061	96.45
	20	932	84.73	1064	96.73
	28	956	86.91	1066	96.91
256	12	728	66.18	1054	95.82
	20	877	79.73	1046	95.09
	28	933	84.82	1039	94.45
512	12	586	53.27	1050	95.45
	20	785	71.36	1052	95.64
	28	838	76.18	1029	93.54
1024	12	467	42.45	1054	95.82
	20	619	56.27	1050	95.45
	28	683	62.09	1033	93.91

We will first present two examples in order to show how the system identifies the right speaker and to demonstrate the effect of applying the TM minimal eigenvalues and NNs to Burg's graphs in speaker identification, i.e., identifying the speaker irrespective of what he is saying.

1) Experiment 1—Classification Without the Application of the TM Minimal Eigenvalues Algorithm: Table I shows the results of this experiment. Again, it may seem convenient to emphasize that in this and the second experiments, we are experiencing *right speaker identification* regardless of his speech. Through different techniques of classification with Burg's method in speaker identification, the number of wrongly recognized samples was reduced to 13 in 1100 samples to have a miss rate of only 1.18%. The achieved successful recognition rate of 98.82% by this method was with the PNN classifiers, as can be observed in Table I, which shows all different conditions. Notice that the input data to the NN are applied directly from Burg's graphs.

The obtained results showed a success rate of 94.82% by using classical classification methods, which are based on point-to-point matching to distinguish between the examined signal features and the features of the signal images taken from the teaching set in the database, and then classifying the signal to the most similar class. The use of the NN has increased the success rate to 97.18% in the case of radial NNs and to 98.82% by the probabilistic neural ones.

TABLE III

SAMPLE OF CALCULATION FOR EXPERIMENT 3 RESULTS: IT SHOWS THE AVERAGE SUCCESS RATE FOR SPEECH (93.64%) AND SPEAKER (94.82%) OVERALL IDENTIFICATION TOGETHER WITH THE PERCENTAGE OF THE AVERAGE FALSE REJECTION RATE IN THE CONSIDERED CASES OF 1100 SAMPLES OF THE TEST SET IN A DATA SET OF 5472 SAMPLES. ALL COMPUTATIONS ARE PERFORMED IN MATLAB

Identification/Misclassification	Recognizing the right word %	False Word Rate Rejection % (Misclassifying the right word)	*Overall Speaker Identification* % (*Speaker Success Rate*)
Identifying the right speaker	92	2.82	*94.82*
False Rate for Speaker Rejection (Misclassifying the right speaker)	1.64	3.55	
Overall Speech recognition (*Speech Success Rate*)	*93.64*		

2) Experiment 2—Classification With the Application of TM Minimal Eigenvalues: Now, we will introduce the results after applying the minimal eigenvalues as input to the classifying system to show how to identify the right speaker. Table II shows the results of this experiment.

Following the classical method of classification, Table II shows that the number of wrongly recognized samples is 144 in 1100 samples (a miss rate of about 13.1%). The recognition rate achieved by this method is only 86.91%. The radial NN, however, has increased this rate to 97.73%, with the minimal eigenvalues as their input. RBF NNs are very popular in SAS recognition [20]. Moreover, they work very well with the algorithm of minimal eigenvalues.

Now, Experiments 3–6 will deal with the second kind of human–computer communication in our suggested system (*Part b* in Section IV—multilevel security for both the spoken words and their speaker). Here, we are using the spoken word as a password after having identified the speaker.

3) Experiment 3—Burg's Model and Conventional Speech Classification Methods: Table III shows the results of this experiment. For $NFFT = 256$ and $P = 20$ for Burg's method and through the classical method of classification, the number of correctly recognized samples for both spoken words and their speaker was 1012 in 1100 samples (92%). The SAS recognition average rate achieved when uttering a word results in recognizing both the correct word and its correct speaker successfully, as can be seen in Table III.

However, the rate of identifying the right speaker with a speech misclassifying possibility has reached 94.82% in average. This comes from the fact that in an average of 2.82% word misclassification coming from false word rejection, the right speaker is still successfully recognized; hence, this percentage should be added to the speaker recognition computation result to have an overall speaker identification of 94.82%. The details of this part of the experiment are given in Table IV. All examined cases are shown with their success and miss cases. It is also given which speaker was misclassified with whom. There still exists another important measuring factor, namely, the false acceptance rate. This will be presented in Table V when explaining the consideration of uttered word success rate regardless of who the words were spoken by.

From the other side, Table III has shown that identifying the right uttered word (speech recognition) results in an average of 93.64%. Again, this is because there is an average of 1.64% coming from the *false acceptance rate*, where the right speaker is misclassified and another speaker is accepted instead, while the right spoken word is still recognized successfully. Table V shows the details of speech recognition together with the misclassification cases, resulting in unsuccessfully (wrongly) recognized words. In all the cases of Table V, the concentration is on identifying the right spoken word whoever spoke it.

4) Experiment 4—Burg's Model and NN Classification: The considered parameters in this experiment are $NFFT = 512$ and $P = 20$. The followed techniques are Burg's method for feature extraction and *probabilistic* NN for classification. After the spectral analysis for the Burg's model has been extracted, the graph data are fed directly to the NN without TM minimal eigenvalues. The number of correctly recognized samples for both spoken words and speaker was 1072 in 1100 samples. The recognition success rate has therefore achieved 97.45%. The false word rejection rate was only 1%, while the false speaker rejection rate was 0.82%.

The same experiment was repeated with *radial* NN as a classifier. The results showed a successful recognition rate of 90.36% (994/1100) for $NFFT = 128$ and $P = 28$. The false word rejection rate was higher than with PNN—It has reached 5.27%. However, the speaker rejection rate was only 1.36%.

5) Experiment 5—TM Minimal Eigenvalues and Conventional Classification: When using the conventional methods in classifying the data directly from Burg's model, the success rate was 92% (1012/1100). Applying the TM algorithm, the success rate was 81.64% (898/1100) using the conventional methods as classifiers. The results are shown in Table VI.

However, with the addition of NNs instead of the conventional ones as classifying tools, the resulting TM-NN system has increased the success rate to an absolutely higher value (speech recognition to 95.64% and speaker recognition to 97.73%). This is shown in the next experiment (Experiment 6), with the NN as classifiers.

6) Experiment 6—TM Minimal Eigenvalues and NN Classification: Not only have the TM minimal eigenvalues rapidly

TABLE IV
LIST OF THE 20 SPEAKERS CONSIDERED IN THE RESEARCH—MEN AND WOMEN OF DIFFERENT NATIONALITIES AGED BETWEEN 23 AND 45 YEARS OLD, WITH TWO MALE AND FEMALE TEENAGERS (SPEAKERS 10 AND 15, RESPECTIVELY). MOST OF THE SAMPLES ARE NATIVE ARABIC-SPEAKING PEOPLE, BUT SOME ARE ARABIC-SPEAKING EUROPEAN PEOPLE. THE RESULTS SHOW THE SUCCESS RATE IN RECOGNIZING THE RIGHT PERSON AND THE PERCENTAGE OF MISS CASES, REVEALING THAT THE FALSE REJECTION TOOK PLACE. ALL COMPUTATIONS ARE PERFORMED IN MATLAB

Uttering Speaker	**Identifying the right speaker (Success Rate %)**	**False Rejection Rate %**																			
		misclassifying the right speaker – they are recognized as:																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	96.36	-	0	0	1.8	0	0	0	0	0	0	0	0	0	1.8	0	0	0	0	0	0
2	98.18	0	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.8	0	0	0
3	96.36	0	0	-	0	0	0	0	0	0	0	0	0	0	1.8	0	1.8	0	0	0	0
4	92.73	1.8	0	0	-	1.8	0	0	0	0	0	0	1.8	0	1.8	0	0	0	0	0	0
5	94.54	0	0	1.8	1.8	-	0	0	0	0	0	0	0	0	0	0	0	1.8	0	0	0
6	96.36	0	0	0	0	0	-	0	0	0	1.8	1.8	0	0	0	0	0	0	0	0	0
7	100	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0	0	0
8	90.91	1.8	1.8	0	0	0	0	0	-	1.8	0	3.6	0	0	0	0	0	0	0	0	0
9	89.09	1.8	0	0	0	0	0	0	3.6	-	0	1.8	0	0	0	0	0	1.8	1.8	0	0
10	87.27	0	0	0	0	0	7.3	1.8	0	1.8	-	0	1.8	0	0	0	0	0	0	0	0
11	98.18	0	0	0	0	0	0	1.8	0	0	0	-	0	0	0	0	0	0	0	0	0
12	89.09	0	0	0	0	3.6	0	1.8	1.8	0	0	0	-	0	0	0	0	0	0	1.8	1.8
13	98.18	0	0	0	0	0	0	0	0	0	0	0	0	-	1.8	0	0	0	0	0	0
14	90.91	0	0	1.8	1.8	3.6	0	0	1.8	0	0	0	0	0	-	0	0	0	0	0	0
15	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0	0
16	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-	0	0	0	0
17	98.18	0	0	0	0	0	0	0	0	0	0	1.8	0	0	0	0	0	-	0	0	0
18	90.91	0	0	0	0	0	0	0	0	0	0	0	3.6	3.6	0	0	0	0	-	1.8	0
19	96.36	0	0	0	0	0	0	0	0	1.8	0	1.8	0	0	0	0	0	0	0	-	0
20	92.73	0	0	0	0	0	0	1.8	0	0	0	1.8	0	0	0	0	0	0	0	3.6	-
Average	94.82																				

improved the system performance but they have also led to a very high success rate (94.64%) when used with the *radial function NNs* in the hybrid manner discussed in Section III. Notice that the false rate of speaker rejection is only 1%. The false word rejection rate (still identifying the right speaker), however, is still high, showing an error of 3.09% in misclassifying the right word. The overall speaker recognition rate has then reached 97.73%. The results in Table VII show the computations for radial function NN classification and TM feature extraction to recognize the spoken Arabic digits and to identify their speakers for different combinations of $NFFT$ and prediction order P. The best results were achieved with $NFFT = 64$ and $P = 20$. The number of correctly recognized samples for both spoken words and speaker was 1041 in 1100 samples (94.64%).

VI. COMPARISON AND CONCLUSION

The SAS identifying system suggested in this paper is categorized into a text-dependent speaker identification based on a closed set of spoken Arabic digits from zero to ten. The work is based on developing a word-and-its-speaker identification to recognize the right words and identify their speaker. The achieved results have proven that the authors' Burg–Töeplitz–NN approach has introduced a high success rate in both speaker and speech identification, and recognition. The successful

TABLE V

IDENTIFICATION OF THE RIGHT UTTERED WORD (SPEECH RECOGNITION) RESULTS IN AN AVERAGE OF 93.64%. AGAIN, THIS IS BECAUSE THERE IS AN AVERAGE OF 1.64% COMING FROM THE FALSE ACCEPTANCE RATE, WHERE, ALTHOUGH THE RIGHT SPEAKER IS MISCLASSIFIED AND ANOTHER SPEAKER IS ACCEPTED INSTEAD, THE RIGHT SPOKEN WORD IS STILL RECOGNIZED SUCCESSFULLY. ALL COMPUTATIONS ARE IN MATLAB

Arabic Uttered Digit	Recognition % of the right word	**False Word Rejection %**										
		misclassifying the right word – they are recognized as:										
		0	1	2	3	4	5	6	7	8	9	10
Syfr – 0	93	-	0	0	0	1	0	0	5	0	1	0
Wahid – 1	94	0	-	0	1	3	1	0	0	1	0	0
Ethnan – 2	98	0	0	-	0	0	0	0	0	2	0	0
Thalath – 3	96	1	1	0	-	1	0	0	0	1	0	0
Arb'a – 4	96	0	3	0	0	-	0	0	0	1	0	0
Khams – 5	93	0	1	0	3	0	-	0	0	2	1	0
Syt – 6	93	0	0	0	1	0	1	-	0	0	5	0
Sab'a – 7	87	0	2	0	4	0	0	0	-	1	2	4
Thamaan – 8	95	0	0	4	1	0	0	0	0	-	0	0
Tes'a – 9	89	0	1	0	1	0	2	3	3	0	-	1
Ashr – 10	96	0	0	1	2	0	0	0	0	0	1	-
Average	93.64 %											

TABLE VI

RESULTS OF RECOGNITION AND IDENTIFICATION WITH TÖEPLITZ AND CLASSICAL CLASSIFICATION. BURG'S PARAMETERS: $NFFT = 128$, $P = 28$. THE NUMBER OF SUCCESSFULLY RECOGNIZED SAMPLES FOR BOTH SPOKEN WORDS AND SPEAKER WAS 898 IN 1100 USED SAMPLES, i.e., 81.64%. ALL COMPUTATIONS ARE PERFORMED IN MATLAB

Identification/Misclassification	Recognizing the right word %	False Word Rate Rejection % (Misclassifying the right word)	*Overall Speaker Identification* % (*Speaker Success Rate*)
Identifying the right speaker	81.64	5.27	*86.91*
False Rate for Speaker Rejection (Misclassifying the right Speaker)	4.82	8.27	
Overall Speech recognition (*Speech Success Rate*)	*86.46*		

TABLE VII

AVERAGE RESULTS FOR SAS CLASSIFICATION WITH THE RADIAL FUNCTION NNs AFTER APPLYING THE TM FEATURE EXTRACTION ALGORITHM. SPEAKER IDENTIFICATION IS 97.73%, WHILE SPEECH RECOGNITION IS 95.64% FOR $NFFT = 64$ AND $P = 20$. ALL COMPUTATIONS ARE PERFORMED IN MATLAB

Identification/Misclassification	Recognizing the right word	False Word Rate Rejection (Misclassifying the right word)	*Overall Speaker Identification* % (*Speaker Success Rate*)
Identifying the right speaker	94.64	3.09	*97.73*
False Rate for Speaker Rejection (Misclassifying the right Speaker)	1	1.27	
Overall Speech recognition (*Speech Success Rate*)	*95.64*		

recognition rate was very high for individually uttered words and for different methods of classification. The speaker identification was about 95% when classifying the speech signal obtained from Burg's model with classical classification methods without NNs or TM feature vector. It reached, however, about 97.73%, with the signal-image feature vector being extracted from the minimal eigenvalues of the TM and fed into a radial NN for classification (Table II), and 98.82% when feeding Burg's plot into a PNN for classification (Table I). This led to the fact that PNN worked very well with Burg's model only, while RNN gave better results when classifying the signal-image feature vector obtained from the Töeplitz model extracted from Burg's plot. Concerning the NN classification of the Burg's and Töeplitz models, this is not a final concluding result as we still are working on the whole model and its mathematical apparatus.

Comparing the authors' approach of SAS recognition system with the methods known to them [18], [20]–[29], we can claim promising heuristic results indeed. The performance of the newly designed and still-under-development system has shown comparable results with other voice recognition approaches, particularly, with the methods demonstrated in [18] and [21]–[24]. These methods of processing for both speech and speaker recognition have applied different approaches to achieve the best success rate. Alghamdi [22], for example, uses classical methods in studying the Arabic phonetics on the basis of the fact that each sound in Arabic including stop consonants has its specific place of articulation. He divides the Arabic sounds into three categories: tight/stop, loose/nonstop, and the sounds between them. The work he presents does not show numerical results, but he rather introduces his approaches in speech analysis and processing. Nofal *et al.* [23], however, have shown another approach in feature extraction and classification tools. They apply the Gaussian hidden Markov models for the purpose of building a system for speaker-independent continuous Arabic speech recognition. The authors develop a set of acoustic models that are based on nonsegmented speech data to segment the speech data used in generating new acoustics. Four language models are built and used in the tests. The authors claim results with 5.26% and 2.72% word error rates for 1340- and 306-word bigram-based language models, respectively, and 0.19% and 0.99% word error rates for 1340- and 306-word context-free grammar-based language models, respectively. In Alotaibi's work [24], a system based on recurrent artificial NN is designed and used in conjunction with spoken Arabic digits from a speech-recognition-field point of view. His study is concentrated on both multispeaker (the same set of speakers was used in both the training and testing phases) and speaker-independent modes using artificial NNs, namely, recurrent Elman network with the concentration of Arabic digits (from zero to nine) as isolated utterances. His recognition system has achieved 99.5% correct digit recognition in the case of the multispeaker mode and 94.5% in the case of the speaker-independent mode, as he claims.

Therefore, we can conclude that the success rate and the results that we obtained are really at a promising level.

Now, considering the system of SAS recognition with the security-system model as an example, the approach presented in this paper has achieved a success rate of about 92.5% in recognizing the right three-digit password together with their right speaker. Therefore, for the multilevel SAS system, after recognizing the speaker successfully, his voice is tested three more times for the password. More precisely, all of the three consecutively spoken digits of the password are recognized right in addition to the fourth digit used primarily to identify the right speaker before entering the system and uttering the password. This result is of high success rate indeed, as the identification of the speaker by uttering a three-element compound password is a procedure that requires probability theory to evaluate its performance. Since the password is composed of three words, then the n miss rate will increase to $n \times n \times n$, causing the success rate of the whole system to decrease in a noticeable manner.

All the experiments have proven that the image-based methods in speech recognition have given as good recognition results as other methods [18], [20]–[29]. It is worth noticing, and as was demonstrated in this paper, that the minimal eigenvalues algorithm, which has shown a very high success rate of recognition in varieties of pattern recognition, particularly, in the case of cursive scripts [11], was successfully applied to the introduced system, as observed in the presented examples and experiments. The authors and their team have been working on the whole system and other TM-based image description ways with the hope to develop the procedure of human–computer communication and reach a perfect security system. The current work, however, is concentrated on utilizing other alternative methods of data feeding into (1) or (2), as shown in the remarks of B in Section III, which are showing developing results. This is why this research can be treated as a primary step in building an open-set SAS identification approach, resulting in a practical security system or other important-in-our-life systems such as human-speech understanding by computer.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments and extremely helpful suggestions and the Editorial Staff of the IEEE Transactions on Industrial Electronics for their efforts in preparation of the proof for this paper.

References

[1] M. Alghamdi, "Speaker's identification—Voice print," *King Khalid Mil. Acad. Q.*, vol. 14, no. 54, pp. 24–28, 1997. (in Arabic).

[2] C.-N. Hsu, H.-C. Yu, and B.-H. Yang, "Speaker verification without background speaker models," in *Proc. IEEE-ICASSP*, Hong Kong, 2003, vol. 2, pp. 233–236.

[3] K. Saeed and M. K. Nammous, "Heuristic method of Arabic speech recognition," in *Proc. IEEE 7th Int. Conf. DSPA*, Moscow, Russia, 2005, pp. 528–530.

[4] R. M. Gray, "Töeplitz and circulant matrices: A review," Stanford Univ. Press, Tech. Rep., Stanford, CA, 2000.

[5] K. Saeed, "Computer graphics analysis: A criterion for image feature extraction and recognition," *MGV—Int. J. Mach. Graph. Vis.*, vol. 10, no. 2, pp. 185–194, 2001.

[6] V. K. Ingle and J. G. Proakis, *Digital Signal Processing Using MATLAB*. Pacific Grove, CA: Brooks/Cole, Jul. 1999.

[7] D. Rocchesso and F. Fontana, Eds., *The Sounding Object*. Florence, Italy: Mondo Estremo Publishing, 2003.

[8] K. Saeed and M. Kozłowski, "An image-based system for spoken-letter recognition," in *Proc. 10th Int. Conf. CAIP*, R. Petkov, R. Westenberg, Eds. Groningen, The Netherlands, Aug. 2003, vol. LNCS 2756, pp. 494–502.

[9] R. Tadeusiewicz, *Speech Signals*. Warsaw, Poland: WKiL, 1988. (in Polish).

[10] J. Durbin, "Efficient estimation of parameters in moving average models," *Biometrics,* pt. 1 and 2, vol. 46, no. 3/4, pp. 306–316, Dec. 1969.

[11] K. Saeed, *Image Analysis for Object Recognition*. Bialystok, Poland: Publications Bialystok Tech. Univ., 2004.

[12] K. Saeed and M. Tabędzki, "Intelligent feature extract system for cursive-script recognition," in *Proc. 4th IEEE Int. WSTST*, Muroran, Japan, 2005, pp. 192–201.

[13] E. A. Guillemin, *A Summary of Modern Methods of Network Synthesis—Advances in Electronics*, vol. III. New York: Academic, 1951, pp. 261–303.

[14] C. Carathéodory, "Über den Variabilitätsbereich der Koeffizienten von Potenzreihen die Gegebene Werte nicht Annehmen," *Math. Ann.*, vol. 64, no. 1, pp. 95–115, Mar. 1907. (in German).

[15] U. Grenander and G. Szegö, *Töeplitz Forms and Their Applications*. Berkeley, CA: Univ. California Press, 1959.

[16] K. Saeed, "On the realization of digital filters," in *Proc. 1st Int. Conf. IEEE/DSPA*, Moscow, Russia, 1998, vol. 1, pp. 141–143.

[17] F. H. Effertz, "On the synthesis of networks containing two kinds of elements," in *Proc. Symp. Modern Netw. Synthesis*, 1955, pp. 145–173.

[18] M. M. El Choubassi, H. E. El Khoury, C. E. J. Alagha, J. A. Skaf, and M. A. Al-Alaoui, "Arabic speech recognition using recurrent neural networks," in *Proc. IEEE/ISSPIT*, Darmstadt, Germany, 2003, pp. 543–547.

[19] T. Hill and P. Lewicki, *Statistics Methods and Applications*. Tulsa, OK: StatSoft Inc., 2006.

[20] M. W. Mak, W. G. Allen, and G. G. Sexton, "Speaker identification using radial basis functions," in *Proc. 3rd Int. Conf. Artif. Neural Netw.*, 1998, pp. 138–142.

[21] A. N. Fox and B. R. Reilly, "Audio-visual speaker identification based on the use of dynamic audio and visual features," in *Proc. 4th Int. Conf. AVBPA*, Guildford, U.K., 2003, pp. 743–751.

[22] M. Alghamdi, *Analysis, Synthesis and Perception of Voicing in Arabic*. Riyadh, Saudi Arabia: Al-Toubah, 2004.

[23] M. Nofal, E. Abdel-Raheem, H. El Henawy, and N. A. Kader, "Acoustic training system for speaker independent continuous Arabic speech recognition system," in *Proc. 4th IEEE ISSPIT*, Turin, Italy, 2004, pp. 200–203.

[24] Y. A. Alotaibi, "Spoken Arabic digits recognizer using recurrent neural networks," in *Proc. 4th IEEE ISSPIT*, Turin, Italy, 2004, pp. 195–199.

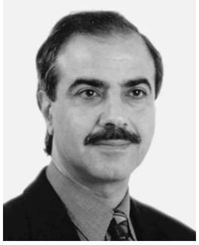
[25] R. Wouhaybi and M. A. Al-Alaoui, "Different implementations of neural networks for speaker recognition," in *Proc. ECCTD*, Stresa, Italy, Aug. 2–29, 1999, vol. II, pp. 1339–1342.

[26] L. Lisker, "'Voicing' in English: A catalogue of acoustic features signaling /b/ versus /p/ in Trochees," *Lang. Speech*, vol. 29, no. 1, pp. 3–11, Jan.–Mar. 1986.

[27] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*. New York: Marcel Dekker, 2001.

[28] Z. Wanfeng, Y. Yingchun, W. Zhaohui, and S. Lifeng, "Experimental evaluation of a new speaker identification framework using PCA," in *Proc. IEEE Int. Conf. Syst., Man and Cybern.*, 2003, vol. 5, pp. 4147–4152.

[29] S.-N. Tsai and L.-S. Lee, "Improved robust features for speech recognition by integrating time-frequency principal components (TFPC) and histogram equalization (HEQ)," in *Proc. IEEE Workshop Autom. Speech Recog. and Understanding*, 2003, pp. 297–302.



Khalid Saeed (M'93) received the B.Sc. degree in electrical engineering–electronics and communications from Baghdad University, Baghdad, Iraq, in 1976, and the M.Sc. and Ph.D. degrees from Wroclaw University of Technology, Wroclaw, Poland, in 1978 and 1981, respectively.

From 1981 to 1986, he was with Al-Fateh University, Tripoli, Libya. From 1987 to 1990, he was with the Higher Institute of Electrical Engineering, Zliten, Libya. From 1990 to 2002, he was the Director of the Center of General Education, Lapy, Poland. Since 1992, he has been with the Department of Real-Time Systems, Faculty of Computer Science, Bialystok Technical University, Bialystok, Poland, where he was the Head of the Department of Computer Engineering from 1994 to 1999 and the Faculty Vice-Dean from 1997 to 2000 and has been the Head of the Research Team of Information Processing Systems since 2000. He is the author of more than 70 publications, seven of which are text and reference books.

Mohammad Kheir Nammous was born in Damascus, Syria, in 1981. He received the M.Sc. degree in software engineering–computer science from Bialystok Technical University, Bialystok, Poland, in 2004. The topic of his thesis was Arabic voice recognition.

Since receiving the M.Sc. degree, he has been conducting experimental research in the field of voice recognition in Arabic language under the supervision of Dr. K. Saeed at Bialystok Technical University. He is the author of four papers on speech and speaker recognition.