EURASIP Journal on Applied Signal Processing 2004:8, 1078–1087

On the Determination of Optimal Model Order for GMM-Based Text-Independent Speaker Identification

M. F. Abu El-Yazeed
Department of Electronics & Communications Engineering, Faculty of Engineering, Cairo University, 12211 Giza, Egypt
Email: elbarawy@aucegypt.edu

M. A. El Gamal
Department of Engineering Physics & Mathematics, Faculty of Engineering, Cairo University, 12211 Giza, Egypt
Email: mhgamal@aucegypt.edu

M. M. H. El Ayadi
Department of Engineering Physics & Mathematics, Faculty of Engineering, Cairo University, 12211 Giza, Egypt
Email: mz_el_ayadi@masrawy.com

Received 1 October 2003; Revised 9 December 2003; Recommended for Publication by Chin-Hui Lee

Gaussian mixture models (GMMs) are recently employed to provide a robust technique for speaker identification. The determination of the appropriate number of Gaussian components in a model for adequate speaker representation is a crucial but difficult problem. This number is in fact speaker dependent. Therefore, assuming a fixed number of Gaussian components for all speakers is not justified. In this paper, we develop a procedure for roughly estimating the maximum possible model order above which the estimation of model parameters becomes unreliable. In addition, a theoretical measure, namely, a goodness of fit (GOF) measure is derived and utilized in estimating the number of Gaussian components needed to characterize different speakers. The estimation is carried out by exploiting the distribution of the training data for each speaker. Experimental results indicate that the proposed technique provides results comparable to other well-known model selection criteria like the minimum description length (MDL) and the Akaike information criterion (AIC).

Keywords and phrases: Gaussian mixture model, goodness of fit, minimum description length, Akaike information criterion, speaker identification, text-independent speaker identification.

1. INTRODUCTION

Speech signal is believed to be among the fast methods to transmit information between human and machine. In speech recognition, the emphasis is on recognizing words and phrases in a spoken utterance, while speaker recognition is concerned with extracting the identity of the person speaking the utterance. The latter has recently found many applications such as telephone financial transactions, machine voice commands, and voice stamp security applications.

Speaker recognition is divided into two main categories, verification and identification. Speaker verification concerns with deciding whether a certain voice sample belongs to a certain speaker or not. Speaker identification systems may be open set or closed set. Closed-set speaker identification addresses the following problem: given an unknown test utterance whose speaker is known a priori to be among a certain group of speakers, to whom does this utterance belongs? Open-set speaker identification includes the additional possibility where a speaker may be outside the given set of speakers [1].

Another distinguishing feature of speaker recognition is whether it is text-dependent or text-independent. In text-dependent systems, the underlying texts of training and testing are the same. On the other hand, the task is more difficult in the text-independent systems, where the utterances used in training phase differ from that used in testing phase [2]. This paper focuses on the closed-set text-independent speaker identification task.

Over the past several years, Gaussian mixture models (GMMs) have become the dominant approach for modeling in text-independent speaker recognition applications. This is evidenced by the numerous research works on the use of GMMs for speaker identification and verification tasks [3, 4, 5, 6]. GMMs are shown to efficiently represent speaker-dependent acoustic features. In this method, each speaker is represented by a single model. Learning is performed by adjusting the model parameters so that the likelihood function of the training pattern is maximized. Testing an unknown utterance is done by calculating its likelihood value with respect to each model. A decision is made to the speaker whose model gives the largest likelihood value.

The motivation behind this work is to enhance the identification performance of systems that use GMMs. In particular, the number of Gaussian components for each model is not specified a priori. Instead, it is determined according to the goodness of fit (GOF) measure of the training data to the model.

The paper is organized as follows. In Section 2, a brief review of the GMM is given. Section 3 describes a procedure for roughly estimating the maximum model order based on the theory of conventional sampling. In addition, a model-order selection technique, based on the GOF measure, is deduced. In Section 4, computer simulation results are presented and justified. Finally, conclusions are drawn in Section 5.

2. GAUSSIAN MIXTURE MODEL

This section is divided into three parts. In the first part, the mathematical representation of a GMM is given. The training procedure using GMM is explained in the second part, followed by a brief description of the conventional GMM-based speaker identification technique in the third part.

2.1. Model description

A GMM is a convex linear combination of multivariate Gaussian probability distributions with different mean vectors and covariance matrices. It can be represented mathematically as

$$\mathrm{p}(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i \mathrm{p}(\mathbf{x}|i,\lambda), \tag{1}$$

where M is the number of Gaussian components, $\mathbf{x} \in \mathbb{R}^D$, w_i is the weight of the ith Gaussian component, and $\mathrm{p}(\mathbf{x}|i;\lambda)$ is the multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Gamma}_i$ and is given by

$$\mathrm{p}(\mathbf{x}|i,\lambda) = \frac{1}{(2\pi)^{D/2}\left|\boldsymbol{\Gamma}_i\right|^{1/2}} \exp\left(- \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Gamma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right). \tag{2}$$

Thus, a GMM with M Gaussian components is parameterized by a set of M positive weights, M mean vectors, and M covariance matrices. These parameters are collectively mathematically represented by the notation

$$\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Gamma}_i\}, \quad i = 1,2,\ldots,M. \tag{3}$$

2.2. Parameter estimation and model training

For a sequence of vectors $\{\mathbf{x}_t\}$, $t = 1,2,\ldots,T$, the likelihood value for that sequence is given by

$$\mathrm{p}(\{\mathbf{x}_t\}|\lambda) = \prod_{t=1}^{T} \mathrm{p}(\mathbf{x}_t|\lambda). \tag{4}$$

The training vectors $\{\mathbf{x}_t\}$ are fed to the GMM in order to set up the model parameters λ such that the likelihood value is maximum. The likelihood value is, however, a highly nonlinear function in the model parameters and direct maximization is not possible. Instead, maximization can be done through iterative procedures. Of the many techniques developed to maximize the likelihood value, the most popular is the expectation maximization (EM) algorithm [7] which proceeds as follows. Starting with a model λ, we need to find another model λ' such that $\mathrm{p}(\{\mathbf{x}_t\}|\lambda') \geq \mathrm{p}(\{\mathbf{x}_t\}|\lambda)$. This is done by maximizing the following auxiliary function,

$$Q(\lambda';\{\mathbf{x}_t\},\lambda) = \sum_{I} \mathrm{p}(\{\mathbf{x}_t\},I|\lambda)\log\left(\mathrm{p}(\{\mathbf{x}_t\},I|\lambda')\right), \tag{5}$$

with respect to λ', where I is a particular sequence of Gaussian component densities which produces $\{\mathbf{x}_t\}$, and Σ_I denotes the summation over all possible sequences of Gaussian component densities. The reestimation formulae for the jth Gaussian component parameters takes the following form [7], $j = 1,2,\ldots,M$,

$$\begin{aligned} w'_j &= \frac{1}{T}\sum_{t=1}^{T} \mathrm{p}(i_t = j|\mathbf{x}_t,\lambda), \\ \boldsymbol{\mu}'_j &= \frac{\sum_{t=1}^{T} \mathrm{p}(i_t = j|\mathbf{x}_t,\lambda)\mathbf{x}_t}{\sum_{t=1}^{T} \mathrm{p}(i_t = j|\mathbf{x}_t,\lambda)}, \\ \boldsymbol{\Gamma}'_j &= \frac{\sum_{t=1}^{T} \mathrm{p}(i_t = j|\mathbf{x}_t,\lambda)\mathbf{x}_t\mathbf{x}_t^T}{\sum_{t=1}^{T} \mathrm{p}(i_t = j|\mathbf{x}_t,\lambda)} - \boldsymbol{\mu}'_j\boldsymbol{\mu}'^T_j. \end{aligned} \tag{6}$$

If diagonal covariance matrices are used, then only the diagonal elements in the covariance matrices need to be updated. For the dth diagonal element $\sigma_j^2(d)$ of the covariance matrix of the jth Gaussian component, the variance update becomes

$$\sigma'^2_j(d) = \frac{\sum_{t=1}^{T}\mathrm{p}(i_t = j|\mathbf{x}_t,\lambda)x_{td}^2}{\sum_{t=1}^{T}\mathrm{p}(i_t = j|\mathbf{x}_t,\lambda)} - \mu'^2_{jd}, \tag{7}$$

where the a posteriori probability for the jth Gaussian component is given by

$$\mathrm{p}(i_t = j|\mathbf{x}_t,\lambda) = \frac{w_j \mathrm{p}(\mathbf{x}_t|j,\lambda)}{\sum_{k=1}^{M} w_k \mathrm{p}(\mathbf{x}_t|k,\lambda)}, \tag{8}$$

in which $x_t(d)$ and $\mu_j(d)$ refer to the dth element of $\mathbf{x}_t$, and μ_j, respectively.

Training a model to a certain pattern $\{\mathbf{x}_t\}$ can be done in the following way. First, appropriate initial values are assigned to the model parameters. The model parameters are updated using equations (6), (7), and (8). The new model parameters become the initial parameters for the next iteration. Updating is continued until no significant increase occurs to the likelihood value or a maximum allowable number of iterations has been exceeded.

2.3. Speaker identification (conventional technique)

Given a group of S speakers represented by GMMs $\lambda_1, \lambda_2, \dots, \lambda_S$ and an unknown test pattern $\{\mathbf{x}_t\}$, $t = 1, 2, \dots, T$, it is required to find the model that best matches this pattern, that is, the model that gives the largest a posteriori probability. Formally, the index of the selected speaker is

$$\hat{s} = \arg\max_{1\le k\le S} \Pr\left(\lambda_k|\{\mathbf{x}_t\}\right) = \arg\max_{1\le k\le S} \frac{\mathrm{p}\left(\{\mathbf{x}_t\}|\lambda_k\right)\Pr\left(\lambda_k\right)}{\mathrm{p}\left(\{\mathbf{x}_t\}\right)}. \tag{9}$$

Assuming equi-probable speakers (i.e., $\Pr(\lambda_k) = 1/S$) and noting that $\mathrm{p}(\{\mathbf{x}_t\})$ is the same for all models, (9) reduces to

$$\hat{s} = \arg\max_{1\le k\le S} \mathrm{p}\left(\{\mathbf{x}_t\}|\lambda_k\right), \tag{10}$$

where $\mathrm{p}(\{\mathbf{x}_t\}|\lambda_k)$ is given by (4). Since $\mathrm{p}(\{\mathbf{x}_t\}|\lambda_k)$ is the product of a large number of small values, direct implementation of (4) on a digital computer will result in an underflow. Instead, maximization is done over $\log(\mathrm{p}(\{\mathbf{x}_t\}|\lambda_k))$, yielding

$$\hat{s} = \arg\max_{1\le k\le S} \log\left(\mathrm{p}\left(\{\mathbf{x}_t\}|\lambda_k\right)\right) = \arg\max_{1\le k\le S} \sum_{t=1}^{T} \log\left(\mathrm{p}\left(\mathbf{x}_t|\lambda_k\right)\right), \tag{11}$$

where $\mathrm{p}(\mathbf{x}_t|\lambda_k)$ is given by (1).

3. THE PROPOSED TECHNIQUE

The GMM-based speaker identification technique proposed by Reynolds and Rose [3] assumes a fixed order for each speaker model. This assumption ignores the fact that the actual distribution of the training data is speaker dependent. In other words, some speaker patterns need to be fitted with a large number of Gaussian components, while others require only a small number of Gaussian components. In [8], two different methods are proposed in order to determine the relationship between the amount of training data and the model order. In the first one, a nonlinear transformation with different parameters was proposed. In the other method, exhaustive experiments (train and test) with different lengths of the training utterance were performed so that a linear relation between speech signal duration and model order could be established. However, the training time of the two methods is very large.

In this work, a new approach for determining the optimum order for each speaker model is presented. The main idea is to employ a well-known statistical measure called GOF measure to decide whether the training data fits well into the GMM distribution or not. This section is divided into three subsections. In Section 3.1, a GOF measure for the GMM distribution is introduced. In Section 3.2, a simple way for estimating the maximum allowable model order, is presented. The method is based on the theory of Monte Carlo simulation (conventional sampling). The final algorithm for determining the optimum order for a speaker model is given in Section 3.3.

3.1. GOF measure

In many statistical applications, it is important to establish a measure of the closeness between the frequency distribution of observations in a sampled space and a hypothesized distribution. Such measures are called GOF measures. Some GOF measures are applicable for any hypothesized distribution like the chi-squared test [9]. However, the focus here will be on a test devoted to the Gaussian distribution since the GMM is a convex combination of Gaussian densities. A popular test for examining Gaussianity is the kurtosis test [10] which measures the ratio between the fourth-order central moment and the squared variance. This ratio should be exactly three for Gaussian random variables. Assuming that random samples $\{x_t\}$, $t = 1, 2, \dots, T$ are taken from a Gaussian distribution $N(\overline{\mu}, \overline{\sigma}^2)$, a modified version of the kurtosis test is established as

$$\mathrm{GOF} = \frac{\sum_{t=1}^{T}\left(x_t-\mu\right)^4/T}{E\left\{\left(x-\mu\right)^4\right\}} = \frac{\sum_{t=1}^{T}\left(x_t-\mu\right)^4/T}{3\sigma^4}, \tag{12}$$

where μ, and σ^2 are the population mean and variance, respectively. The numerator is a good estimator of $3\overline{\sigma}^4$ if the distribution is Gaussian, but may overestimate or underestimate $3\overline{\sigma}^4$ when there is departure from Gaussianity. Thus, values of GOF differing considerably from one indicate that the hypothesis of Gaussianity should be rejected. The above test can be generalized to the case of the multivariate Gaussian distribution in the following way. In this case, we test the hypothesis that the random samples $\{\mathbf{x}_t\}$, $t = 1, 2, \dots, T$ are drawn from the multivariate Gaussian distribution $N(\overline{\boldsymbol{\mu}}, \overline{\boldsymbol{\Gamma}})$. The GOF will be the ratio between a sample estimate and the model estimate of a fourth-order statistic centered at $\overline{\boldsymbol{\mu}}$. Therefore, it can be expressed by the following formula:

$$\mathrm{GOF} = \frac{(1/T)\sum_{t=1}^{T}\left(\left(\mathbf{x}_t-\boldsymbol{\mu}\right)^T\boldsymbol{\Gamma}^{-1}\left(\mathbf{x}_t-\boldsymbol{\mu}\right)\right)^2}{E\left\{\left((\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^2\right\}}, \tag{13}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$ are the population mean vector and covariance matrix, respectively. In the appendix, we show that

$$E\left\{\left((\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^2\right\} = D^2 + 2D. \tag{14}$$

Substituting (14) in (13), we get

$$\mathrm{GOF} = \frac{(1/T)\sum_{t=1}^{T}\left(\left(\mathbf{x}_t-\boldsymbol{\mu}\right)^T\boldsymbol{\Gamma}^{-1}\left(\mathbf{x}_t-\boldsymbol{\mu}\right)\right)^2}{D^2+2D}. \tag{15}$$

As mentioned before, the GMM is a convex combination of multivariate Gaussian distributions with different mean vectors and covariance matrices. Therefore, if we partition the given training data set into M clusters using the k-means algorithm, it will be reasonable to assume that the data vectors of each cluster follow a unimodal multivariate Gaussian distribution. Therefore, we may test the hypothesis that the given data vectors follow the GMM distribution by testing the Gaussianity of each cluster. Formally, the GMM parameters are estimated using the EM algorithm. Each data vector is assigned to the cluster with the nearest mean vector (in a log-likelihood sense) and thereby M clusters are formed. Denoting the ith cluster by C_i, its GOF value is given by

$$\mathrm{GOF}_i = \frac{(1/T_i)\sum_{\mathbf{x}_t \in C_i}\left(\left(\mathbf{x}_t - \boldsymbol{\mu}_i\right)^T \boldsymbol{\Gamma}_i^{-1}\left(\mathbf{x}_t - \boldsymbol{\mu}_i\right)\right)^2}{D^2 + 2D}, \quad (16)$$

where T_i is the number of data points in the ith cluster. Clearly, we can construct an $M \times 1$ column vector; $\mathbf{g} = [\mathrm{GOF}_i]$, $i = 1, 2, \ldots, M$. Ideally, all elements of $\mathbf{g}$ should be equal to unity. Denoting this ideal vector by $\mathbf{g}_{\text{ideal}}$, we suggest defining a global measure of GOF, GGOF, as

$$\mathrm{GGOF} = 1 - \frac{\left\|\mathbf{g}_{\text{ideal}} - \mathbf{g}\right\|_1}{\left\|\mathbf{g}_{\text{ideal}}\right\|_1} = 1 - \frac{1}{M}\sum_{i=1}^{M}\left|1 - \mathrm{GOF}_i\right|. \quad (17)$$

The last term in the above equation represents an average value of the errors caused by individual components.

3.2. Finding an upper bound for the model order

The relatively limited number of training vectors imposes a constraint on the maximum allowable number of Gaussian components, above which the estimation of model parameters will not be reliable. We suggest the following simple method to obtain a rough estimate of the maximum possible model order.

The data is grouped into M clusters using the k-means algorithm [11]. Assume that the prior probability of the ith cluster is w_i, that is, each training vector is classified to the ith cluster with probability w_i. Evidently, the number of data points in the ith cluster, T_i, follows the binomial distribution $b(T, w_i)$. Thus, the mean and the variance of T_i are

$$\begin{aligned} E\{T_i\} &= Tw_i, \\ \operatorname{var}\{T_i\} &= Tw_i(1 - w_i). \end{aligned} \quad (18)$$

According to Monte Carlo simulation (conventional sampling) theory [12], the number of iterations (data points) is sufficient if the ratio between the standard deviation of T_i to its mean value does not exceed a specified threshold, usually taken as 0.1, that is,

$$\frac{\sqrt{\operatorname{var}\{T_i\}}}{E\{T_i\}} = \frac{\sqrt{Tw_i(1 - w_i)}}{Tw_i} \leq 0.1 \quad (19)$$

or

$$w_i \geq \frac{1}{0.01T + 1}, \quad i = 1, 2, \ldots, M. \quad (20)$$

Thus, the model parameters are reliably estimated if all prior probabilities are greater than $1/(0.01T + 1)$. The prior probability of the ith cluster is estimated as $\hat{w}_i = \hat{T}_i/T$, where $\hat{T}_i$ is the actual number of data points in the ith cluster after performing the k-means algorithm.

3.3. GOF-based training algorithm

Having established the GOF measure for the GMM distribution, the number of Gaussian components can be determined while training via the following algorithm.

(1) Start with model order $M = 1$.

(2) Group the training data into M clusters. Estimate the prior probability of the ith cluster as $\hat{w}_i = \hat{T}_i/T$, $i = 1, 2, \ldots, M$.

(3) If $\min_{i=1,2,\ldots,M} w_i < 1/(0.01T + 1)$, go to step (8).

(4) Apply the EM algorithm to find the model $\lambda_M \in \Lambda_M$ that maximizes the likelihood function of the training data, where Λ_M is the set of all GMMs with M Gaussian components.

(5) Partition the data set into M clusters. Calculate the GOF value of each cluster using (16).

(6) Calculate the global GOF value for model order M, GGOF_M, using (17).

(7) Increment M by one and return to step (2).

(8) The optimum model order M_{opt} is determined by

$$M_{\text{opt}} = \arg\max_{k=1,2,\ldots,M-1} \mathrm{GGOF}_k. \quad (21)$$

A couple of points should be addressed in this context. First, the GOF can play the role of an "educated guess" of the appropriate number of Gaussian components. The main aspect of the GMM technique is that both the training and test patterns, though different, share similar underlying distributions. As a result, using the GOF measure to find this optimum distribution from the training data is appropriate for the test utterance. Moreover, increasing the number of Gaussian components does not necessarily lead to an increase in the GOF. If this number is too large, then the number of data points close to each center will decrease, resulting in a bad fit for each component and a small value for the GOF.

4. EXPERIMENTAL EVALUATION

In this section, the GOF measure is applied to a particular example in order to determine the optimal number of Gaussian components required for characterization of different speakers.

4.1. Database development

The speech database contains speech time samples of 95 speakers, 50 males, 45 females. Each speaker has recorded five phrases, one for training and the other four for testing.

In order to ensure that the speaker identification algorithm is text-independent, the five recorded utterances are completely different. All phrases are recorded using a high quality microphone in almost noise-free conditions. The sampling rate of the digital recorder is 11025 Hz and each sample is represented by 16 bits. The speech samples of each speaker are grouped into frames. Each frame contains 256 successive samples from which short-time energy is computed. A frame is discarded if its energy is less than some specified threshold. In our database, this threshold is taken as 0.01 of the maximum frame energy. Low energy frames represent from 30 to 40% of the total frames of each utterance. This is the simplest form of speech silence discrimination [13]. The number of frames in each training utterance is kept fixed at 1721 frames, extracted from 20 seconds of pure speech. Each testing utterance contains 429 frames extracted from 5 seconds of pure speech.

In order to make experiments more realistic, we created another version of the testing phrases in which the effects of the telephone channel environment, for example, noise and band-limitation, are simulated. In our simulation, the signal-to-noise ratio (SNR) is taken as 20 dB. The passband of the telephone channel is from 300 to 3300 Hz.

4.2. Feature extraction

Although there are no speech features that completely identify the speaker, the speech spectrum has shown to be very effective, since it reflects a person's vocal tract structure which distinguishes one's utterance from another [3]. For this reason, Mel frequency cpestrum coefficients (MFCC) has been used extensively and shown to be very efficient in speaker identification tasks. In our database, 25 MFCC's, derived from the linear prediction (LP) polynomial, are calculated for each frame. The overlapping between frames is 50%. Hamming window is used in time domain, triangular-shaped filters are used in Mel domain, and filters act in the absolute magnitude domain. Cepstral analysis is performed only over the telephone passband (300–3300 Hz).

4.3. Performance evaluation

This section compares the performance of the proposed algorithm to two well-known model order selection criteria, minimum description length (MDL) [14] and Akaike information criterion (AIC) [15]. The MDL for a model λ is given by

$$\text{MDL} = -\log\left(\text{p}\left(\{\mathbf{x}_t\}|\lambda\right)\right) + \frac{1}{2}N(\lambda)\log T, \tag{22}$$

where $N(\lambda)$ is the number of parameters of the model λ. The AIC objective function for a model λ is given by

$$\text{AIC} = -2\log\left(\text{p}\left(\{\mathbf{x}_t\}|\lambda\right)\right) + 2N(\lambda). \tag{23}$$

Each of the above three techniques was applied on two GMM-based speaker identification systems, employing the database described in Section 4.1. One of the two systems utilizes full covariance matrices in speaker models, while the other utilizes diagonal covariance matrices. The EM algorithm, used for model training, stops if the difference between two successive log-likelihood values is less than 5×10^{-7} or the number of iterations exceeds 200, whichever condition comes first. Because of the limited size of training data, the parameter estimates, obtained by the EM algorithm, were very sensitive to the model initialization. In order to overcome the above deficiency, the training data of each speaker is clustered using the k-means algorithm twenty times, each with a different random initialization. For each trial, the model parameter values are stored and the average quantization error is computed. The model that attains the smallest value for the quantization error is selected as an initial model to the EM algorithm. So as to demonstrate the effects of telephone channel on the recognition accuracy, both the clean and noisy versions of the utterances, assigned for testing, were used in the identification phase. In each case, the average CPU testing time for each technique was also measured. All simulations were carried out on a Pentium III PC with a processor clock speed of 1 GHz.

The global GOF, MDL, and AIC values of the patterns of three typical speakers are plotted versus the model order in Figures 1, 2, and 3, respectively. In the upper subplots of each figure, the speaker GMMs have diagonal covariance matrices while full covariance matrices are used in the lower subplots. In all figures, the value at which the model order is considered optimum is marked by an asterisk. The vertical dotted line refers to the maximum model order, afforded by the training data. It can be noticed from all figures that the optimum model orders for GMMs with full covariance matrices are somewhat smaller than those corresponding to diagonal covariance matrices. In Figures 2a, 2b, 2c, 3a, 3b, and 3c, we see that optimum GMM orders obtained by the MDL and AIC criteria with diagonal covariance matrices coincide with the maximum allowable model order. This indicates the fact that the limited amount of training data does not support the use of diagonal covariance matrices. In other words, the optimum model order for the diagonal covariance case may be so large that it requires amount of training data too large to be available in many situations.

Table 1 compares the identification performances of the six available systems in terms of the identification accuracy (success ratio), the CPU testing time, and the mean and the standard deviation of the number of Gaussian components. As shown from the table, the GOF provides the greatest identification accuracy in general. When using full covariance matrices, the GOF requires slightly more time to identify the speaker than the AIC technique. However, the GOF is superior to the other two techniques when diagonal covariance matrices are used. Although the GOF gives the largest identification accuracy, it requires the least identification time. From the table, one may establish that the identification accuracy is somewhat proportional to the average number of Gaussian components. The standard deviation of model orders is relatively small (about 1–3 Gaussian components), indicating a small fluctuation in the number of Gaussian components over all speaker models.

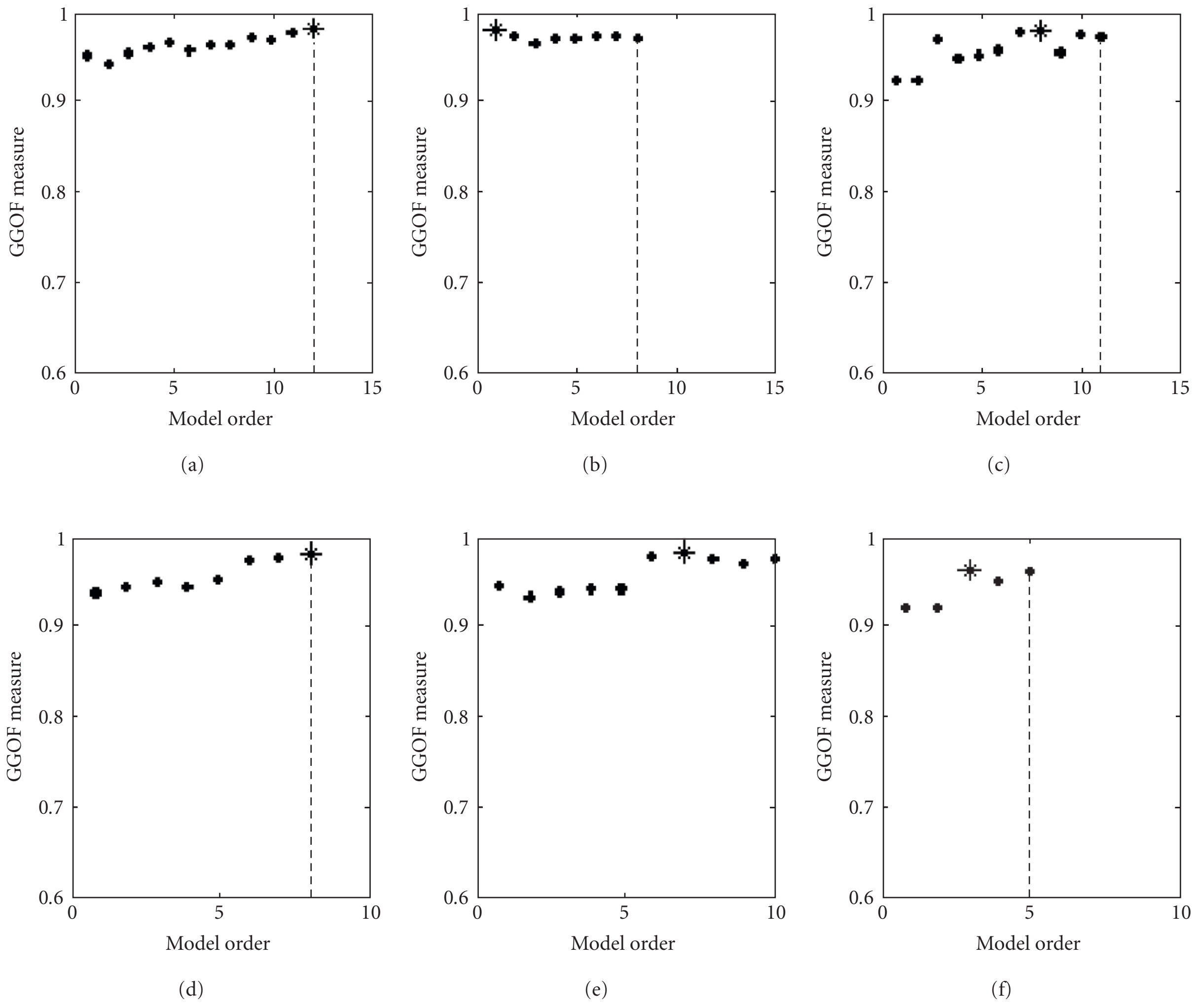

Figure 1: Global GOF versus model order for three typical speakers. (a), (b), and (c) The speaker GMMs have diagonal covariance matrices. (d), (e), and (f) The speaker GMMs have full covariance matrices.

It is also evident that the use of full covariance matrices provides an increase of about 3–7% in the identification accuracy, with a slight increase in the testing time in the case of employing either the proposed GOF technique or the AIC technique. However, the MDL techniques achieve a higher identification accuracy with a less identification time when using full covariance matrices. One can also deduce that GMMs with full covariance matrices achieve a better fit to the distribution of training data than do GMMs with diagonal covariance matrices, especially when the size of the training data is limited.

Another important remark is about the great degradation in the identification accuracy caused by the telephone channel distortion effects. As shown in the table, the accuracy of identifying a speaker via a telephone channel is about 15–22% less than that via high fidelity channel providing clean undistorted utterances. The telephone channel causes mainly two undesirable effects, additive white Gaussian noise (AWGN) and band-limitation of the utterances to be tested. These two factors cause a mismatch between the distributions of the training and testing utterances. The effect of band-limitation is more severe, since it results in a distortion of the power spectral density of the testing utterances. From the table, it is evident that the GOF can be considered a technique against telephone channel effects. While both the GOF and AIC techniques achieve almost the same identification accuracy when using the high fidelity version of the testing utterances, the GOF attains about 6% increase in the identification accuracy for the case of the noisy version utterances.

Finally, it is worth comparing the performance of a speaker identification system employing a model order selection criterion to that of another with a fixed-order for all speaker GMMs. For this purpose, the following experiment was conducted. The training data of each speaker is used to train a six-component GMM with full covariance matrices.

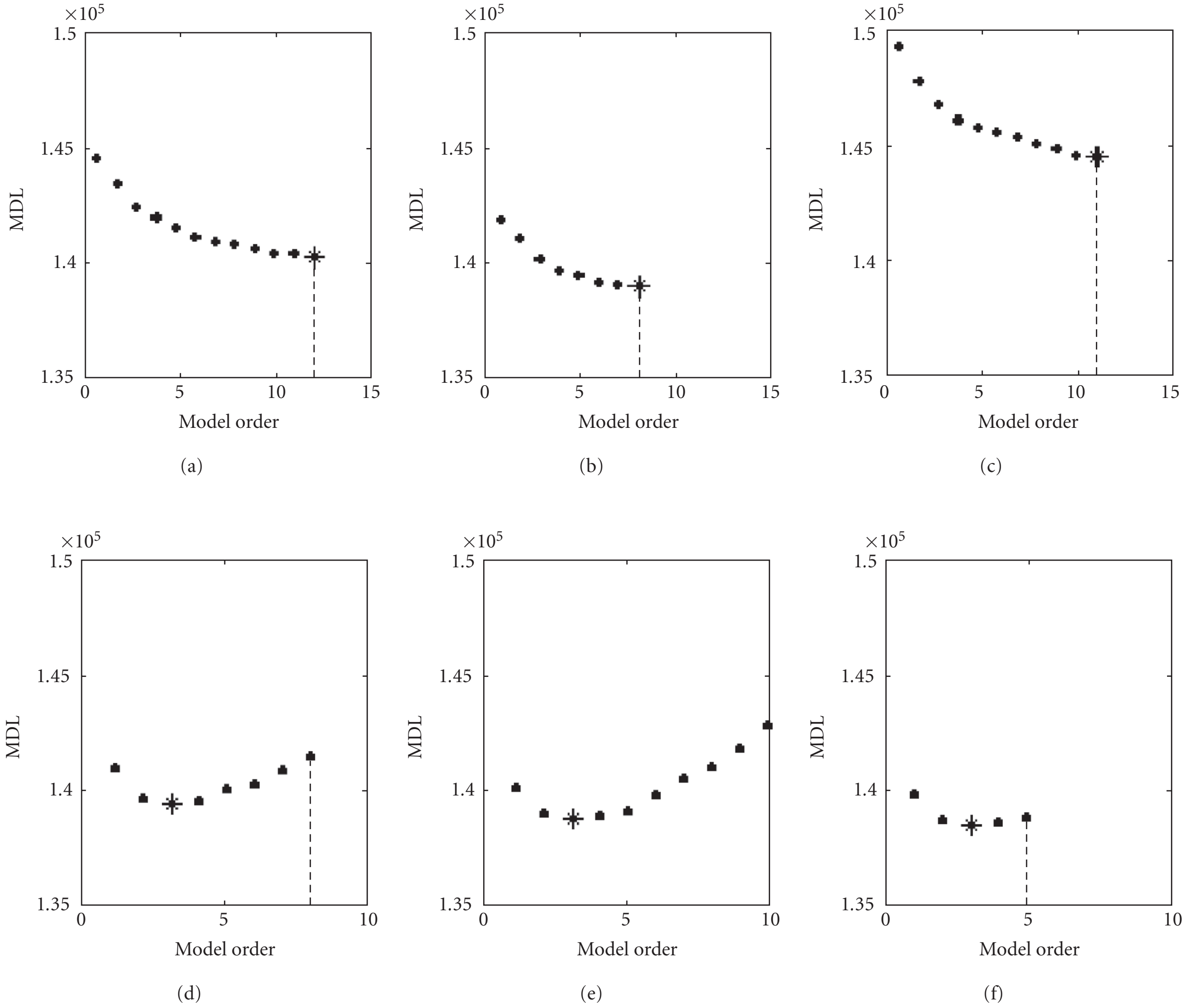

FIGURE 2: MDL versus model order for three typical speakers. (a), (b), and (c) The speaker GMMs have diagonal covariance matrices. (d), (e), and (f) The speaker GMMs have full covariance matrices.

Testing using five-second utterances with telephone channel quality, the identification accuracy was found to be comparable to the speaker identification system employing the GOF model order selection criterion. Decreasing the duration of the testing utterances to one second of pure speech, the identification accuracy was 59.58% for the fixed-order system and 60.16% for the GOF-based system. Thus, it can be concluded that the proposed GOF-based technique outperforms the conventional technique especially in the difficult task of short test utterances.

5. CONCLUSIONS

The determination of the appropriate number of Gaussian components per model is instrumental for the success of any GMM-based speaker identification technique. In this paper, a GOF measure for speaker identification is introduced, derived, and justified. The findings of this research are summarized in the following observations.

(i) The available amount of training data imposes a constraint on the range of possible values of model orders. For a limited size of the training data, increasing the model order (and the number of the unknown parameters of the classifier in consequence) decreases the reliability of the parameter estimates.

(ii) The minimum number of Gaussian components required to adequately model the speaker data relies to a larger extent on the data distribution rather than its amount. Therefore, the GOF measure is a powerful tool that can be used in determining the appropriate number of Gaussian components.

(iii) In most cases, choosing too many Gaussian components has almost no significant effect on the final

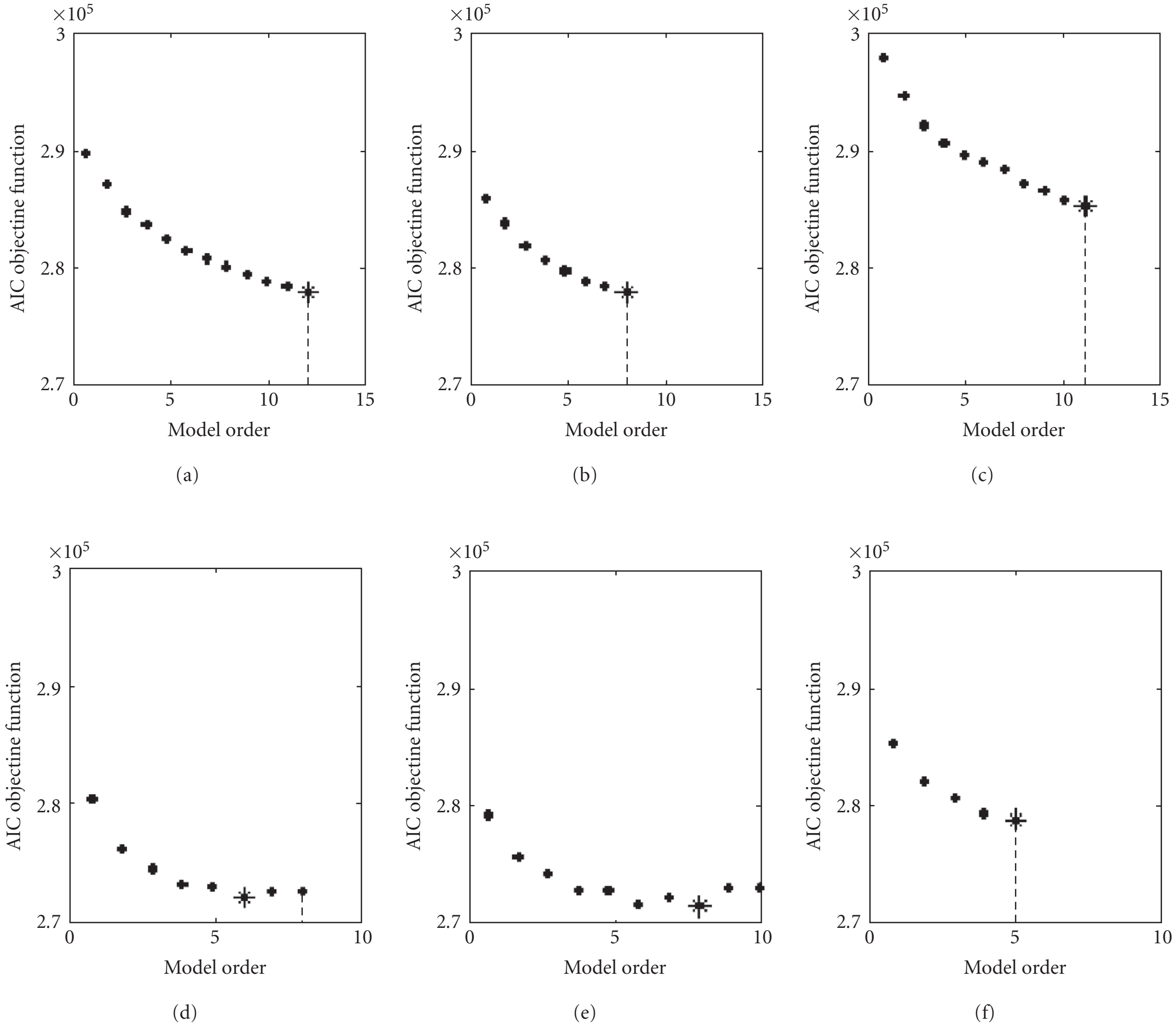

Figure 3: AIC objective function versus model order for three typical speakers. (a), (b), and (c) The speaker GMMs have diagonal covariance matrices. (d), (e), and (f) The speaker GMMs have full covariance matrices.

recognition performance. On the other hand, the testing time increases considerably.

(iv) In general, the GOF technique achieves a better identification performance than the MDL and AIC techniques. In some cases, the GOF provides a greater identification performance, with less time required to identify the speaker.

(v) Utilizing the GOF measure in determining the optimum model order increases the robustness of the speaker identification system against the telephone channel effects, like noise and band-limitation.

(vi) For the case of limited size of training data, the performance of GMM systems using full covariance matrices is superior to those using diagonal covariance matrices, in terms of the classification accuracy (and the identification time in the case of the MDL technique).

APPENDIX

First, it can be shown that if u, v, w, *and* z are four Gaussian random variables, each with a zero mean, then [16]

$$\begin{aligned} E\{uvwz\} &= E\{uv\}E\{wz\} + E\{uw\}E\{vz\} \\ &\quad + E\{uz\}E\{vz\}. \end{aligned} \tag{A.1}$$

The above relation can be extended to the vector case. In this context, if $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$, and $\mathbf{z}$ are four D-dimensional multivariate Gaussian random vectors, each with mean equal to the zero vector, an expression for $E\{\mathbf{u}^T\mathbf{v}\mathbf{w}^T\mathbf{z}\}$ is derived as follows,

$$\begin{aligned} E\{\mathbf{u}^T\mathbf{v}\mathbf{w}^T\mathbf{z}\} &= E\left\{\sum_{k=1}^{D}\sum_{l=1}^{D} u_k v_k w_l z_l\right\} \\ &= \sum_{k=1}^{D}\sum_{l=1}^{D} E\{u_k v_k w_l z_l\}. \end{aligned} \tag{A.2}$$

Table 1: Identification performances for several GMM-based speaker identification systems.

Model selection criterion	Type of covariance matrices	Identification accuracy (high fidelity version)	Identification accuracy (telephone channel version)	Average CPU testing time/speaker/utterance (seconds)	Average model order	Standard deviation of model orders
GOF	Full	95.00	77.89	3.9001	6.6632	1.5131
MDL	Full	95.79	74.21	2.4630	4.1474	1.2962
AIC	Full	94.21	71.58	3.5724	6.0632	1.6230
GOF	Diagonal	87.37	72.11	2.5973	9.1158	2.4705
MDL	Diagonal	87.37	70.53	3.3255	9.7789	1.3125
AIC	Diagonal	87.11	71.58	2.7550	9.9579	1.9181

Clearly, each of u_k, v_k, w_l, z_l is a Gaussian random variable with a zero mean. Using (A.1) and (A.2) takes the following form

$$\begin{aligned} E\{\mathbf{u}^T\mathbf{v}\mathbf{w}^T\mathbf{z}\} &= \sum_{k=1}^{D}\sum_{l=1}^{D}\big(E\{u_kv_k\}E\{w_lz_l\} + E\{u_kw_l\}E\{v_kz_l\} \\ &\qquad + E\{u_kz_l\}E\{v_kw_l\}\big) \\ &= E\{\mathbf{u}^T\mathbf{v}\}E\{\mathbf{w}^T\mathbf{z}\} + \operatorname{tr}\big(E\{\mathbf{u}\mathbf{w}^T\}E\{\mathbf{z}\mathbf{v}^T\}\big) \\ &\quad + \operatorname{tr}\big(E\{\mathbf{u}\mathbf{z}^T\}E\{\mathbf{w}\mathbf{v}^T\}\big). \end{aligned} \tag{A.3}$$

Substituting for $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$, and $\mathbf{z}$ in (A.3) by $\boldsymbol{\Gamma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$ and simplifying,

$$\begin{aligned} &E\Big\{\big((\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\big)^2\Big\} \\ &\quad = E^2\{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\} \\ &\qquad + 2\operatorname{tr}\big(E\{\boldsymbol{\Gamma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1/2}\}\big). \end{aligned} \tag{A.4}$$

By definition,

$$\boldsymbol{\Gamma} = E\Big\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\Big\}. \tag{A.5}$$

Hence,

$$\begin{aligned} E\Big\{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big\} &= E\Big\{\operatorname{tr}\big((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}\big)\Big\} \\ &= \operatorname{tr}\big(E\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}\}\big) \\ &= \operatorname{tr}\big(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{-1}\big) = D, \\ E\{\boldsymbol{\Gamma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1/2}\} &= \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{-1/2} = \mathbf{I}. \end{aligned} \tag{A.6}$$

Substituting (A.6) in (A.4) gives

$$E\Big\{\big((\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\big)^2\Big\} = D^2 + 2D. \tag{A.7}$$

REFERENCES

[1] K. R. Farrell, R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks and conventional classifiers," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, pp. 194–205, 1994.

[2] M. F. Abu El-Yazeed, A. H. Khalid, and M. A. El-Gamal, "A two stage classifier for speaker identification in multi-speaker data," in *Proc. International Conference on Industrial Electronics, Technology and Automation (IETA)*, vol. 1, pp. 164–169, Cairo, Egypt, December 2001.

[3] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech, and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[4] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[5] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[6] D. A. Reynolds, "Automatic speaker recognition using Gaussian mixture speaker models," *Lincoln Laboratory Journal*, vol. 8, no. 2, pp. 173–192, 1995.

[7] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.

[8] C. Tadj, P. Dumouchel, and P. Ouellet, "GMM based speaker identification using training-time-dependent number of mixtures," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, pp. 761–764, Seatle, Wash, USA, May 1998.

[9] R. E. Walpole and R. H. Myers, *Probability and Statistics for Engineers and Scientists*, Macmillan, New York, NY, USA, 1993.

[10] M. R. Spiegel and R. Meddis, *Probability and Statistics*, Mc-Graw Hill, New York, NY, USA, 1988.

[11] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[12] R. T. Mitchell, "Importance sampling applied to simulation of false alarm statistics," *IEEE Trans. on Aerospace and Electronics Systems*, vol. 17, no. 1, pp. 15–24, 1981.

[13] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1978.

[14] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[15] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[16] B. Picinbono, *Eléments de théorie du signal*, Dunod Université, Paris, France, 1977.

M. F. Abu El-Yazeed was born in Cairo, Egypt, in February 1959. He received his B.S. degree with honors in electronics and communication engineering from Cairo University, Egypt, in 1982. He also received his M.S. and Ph.D. degrees from Cairo University in 1986 and 1990, respectively. In 1982, he joined the Department of Electronics and Communication Engineering, Cairo University, where he is now an Associate Professor. From 1994 to 1999, he served as an Assistant Professor at the Department of Physics, Emirates University. His research interests are in the area of digital signal processing, fault diagnosis and testing of electronic circuits, and neural networks applications.

M. A. El Gamal received his B.S. degrees in electronics and communication engineering from Cairo University and in applied mathematics from Ain Shams University, in 1977 and 1980, respectively. He received his Ph.D. degree in Electrical and Computer Engineering from Ohio University, Athens, Ohio in 1990. In 1987, he worked at the National Institute of Standards and Technology, Gaithersburg, Maryland as a Guest Scientist. Since 1990, he has been with the Department of Engineering Physics and Mathematics, Cairo University, where he is now an Associate Professor. Dr. El Gamal's research interests include fault diagnosis of analog and mixed-signal circuits, computer-aided design of integrated circuits, applications of neural networks and fuzzy logic, optimization techniques, genetic algorithms, and statistical pattern recognition.

M. M. H. El Ayadi received his B.S. degree in electronics and communication engineering from Cairo University in 2000. He received his M.S. degree in engineering mathematics in 2003. He is currently a Teaching Assistant in the Engineering Physics and Mathematics Department at the same university. His research interests include statistical pattern recognition, speech processing, and statistical signal processing.