Text-Independent Speaker Recognition Based on the Hurst Parameter and the Multidimensional Fractional Brownian Motion Model

Ricardo Sant'Ana, Rosângela Coelho, *Member, IEEE*, and Abraham Alcaim

***Abstract*—In this paper, a text-independent automatic speaker recognition (ASkR) system is proposed—the SR_{Hurst}—which employs a new speech feature and a new classifier. The statistical feature pH is a vector of *Hurst* (H) parameters obtained by applying a *wavelet*-based multidimensional estimator (*M_dim_wavelets*) to the windowed short-time segments of speech. The proposed classifier for the speaker identification and verification tasks is based on the multidimensional fBm (*fractional Brownian motion*) model, denoted by *M_dim_fBm*. For a given sequence of input speech features, the speaker model is obtained from the sequence of vectors of H parameters, means, and variances of these features. The performance of the SR_{Hurst} was compared to those achieved with the Gaussian mixture models (GMMs), autoregressive vector (AR), and Bhattacharyya distance (dB) classifiers. The speech database—recorded from fixed and cellular phone channels—was uttered by 75 different speakers. The results have shown the superior performance of the *M_dim_fBm* classifier and that the pH feature aggregates new information on the speaker identity. In addition, the proposed classifier employs a much simpler modeling structure as compared to the GMM.**

***Index Terms*—Automatic speaker recognition, Hurst parameter, multidimensional fractional Brownian motion, wavelet-based estimation.**

I. Introduction

AUTOMATIC speaker recognition systems can be classified into speaker identification and speaker verification [1] processes. In the closed-set speaker identification process, a speech utterance has to be identified as to which of the registered speakers it belongs. On the other hand, speaker verification systems attempt to accepting or rejecting the identity of a claimed speaker. A generic automatic speaker recognition (ASkR) system is based on a speaker modeling and involves three basic steps: speech acquisition/preprocessing, feature extraction, and classification. This paper proposes a text-independent ASkR system—referred to as SR_{Hurst} (*Speaker Recognition with Hurst*)—that employs a new feature and a new classifier.

The most commonly used features employed in speaker recognition are the linear prediction coefficient (LPC) -derived cepstral parameters and the mel-cepstral coefficients. Generally, these physiological features are not robust to the channels acoustic distortion and their extraction from the speech signal requires a high computational load. This is due to the fact that these features model the spectral characteristics of the human vocal mechanism. A new statistical feature (pH) is proposed in this paper. It consists of a vector of *Hurst* (H) parameters for ASkR systems. Unlike the physiological features, the pH feature tends to be robust to channel distortions, since it models the stochastic behavior of the speech signal. The pH feature is not related to the transfer functions of the vocal tract and needs less complex extraction/estimation methods. Additionally, it can be obtained in real-time, i.e., during speakers' activity.

When comparing the pH feature with the standard features, the data mismatch between training and recognition data was not considered. Hence, there was no need of channel compensation procedures. Only in this situation, the use of time derivatives features, such as delta cepstrum, could be interesting. Besides, even in this case, Reynolds showed in [2] that compensation techniques (e.g., "mean normalization") are much more effective than the use of delta cepstrum. For all these reasons, delta parameters were not considered in this work.

The proposed classifier (*M_dim_fBm*) is based on the fractional Brownian motion (fBm) stochastic process. However, the speech signal was not considered as a fractal or self-similar process. The classifier can be applied to any feature matrix. The *M_dim_fBm* exploits the relationship and the evolution of the matrix elements to derive a speaker model. Only for fractal or self-similar processes, the H parameter can be related to a fractal dimension (D_h) [3] through the equation $D_h = 2 - H$, where D_h is the fractal Hausdorff dimension [4]. The fractal dimension was previously used in pattern recognition studies in [5] and [6]. In [7], the fractal dimension was applied for discriminating fricative sounds. A speaker identification system using cepstral coefficients is compared in [8] to a system based on the joint use of cepstral coefficients and the fractal dimension. In contrast to our assumption, these studies assume the hypothesis that the speech is a fractal signal.

Several classification procedures—such as the Gaussian mixture model (GMM) [9] , the autoregressive vector (AR) vector model [10] and the dB (Bhattacharyya distance) [11]—have been proposed for ASkR systems. These techniques are used for comparative purposes in the performance evaluation of the proposed classifier *M_dim_fBm*. The *M_dim_fBm* models the speech characteristics of a particular speaker using the H parameters along with the statistical means and variances of the

Manuscript received July 19, 2004; revised March 30, 2005. This work was supported in part by FAPERJ. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Timothy Hazen.

R. Sant'Ana and R. Coelho are with the Electrical Engineering Department, Instituto Militar de Engenharia, Rio de Janeiro, 2290-270 Brazil (e-mail: ricksant@ime.eb.br; coelho@ime.eb.br).

A. Alcaim is with the Center for Telecommunication Studies of the Pontífica Universidade Católica do Rio de Janeiro (CETUC-PUC/Rio), 22453-900 Brazil (e-mail: alcaim@cetuc.puc-rio.br).

Digital Object Identifier 10.1109/TSA.2005.858054

input speech matrix features. Moreover, in this paper, it is shown that the proposed classifier yields a better modeling accuracy with a lower computational load. The *M_dim_fBm* is characterized by only three scalar parameters (i.e., mean, variance, and H) while the Gone 1 mean vector, and one covariance matrix to achieve comparable performance results [2].

This paper is organized as follows. Section II describes the *Hurst* parameter and the methods to extract/estimate the pH feature. Section III provides a brief description of the GMM, AR, and dB classifiers which performance is compared with the proposed *M_dim_fBm* classifier. Section IV defines the fractional Brownian motion process and presents the proposed classifier *M_dim_fBm*. Performance results obtained for the speaker identification and speaker verification tasks are reported and discussed in Section V. Finally, Section VII presents the main conclusions of this paper.

II. Hurst Parameter

The *Hurst* parameter[1] expresses the time-dependence or scaling degree of a stochastic process. It can also be defined by the decaying rate of the autocorrelation coefficient function $\rho(k)$ $(-1 < \rho(k) < 1)$ as $k \to \infty$. Let the speech signal be represented by a stochastic process $X(t)$, with finite variance and normalized autocorrelation function (ACF) or autocorrelation coefficient

$$\rho(k) = \frac{Cov\left[X(t), X(t+k)\right]}{Var\left[X(t)\right]}, \quad k = 0, 1, 2, \ldots \tag{1}$$

where the $\rho(k)$ belongs to [-1,1] and $\lim_{k\to\infty} \rho(k) = 0$. The asymptotic behavior of $\rho(k)$ is given by

$$\rho(k) \sim H(2H-1)k^{2(H-2)}. \tag{2}$$

This means that $\rho(k)$ is a slowly decaying function and that when $k \to \infty$, $\rho(k) \sim H(2H-1)k^{2(H-2)}$ and, hence, $\rho(k)/H(2H-1)k^{2(H-2)} \sim 1$. The H parameter is then the exponent of the ACF [12] of a stochastic process. According to the value of H $(0 < H < 1)$, stochastic processes can be classified as the following.

- Antipersistent processes—$0 < H < (1/2)$. The ACF rapidly tends to zero and $\sum_{k=-\infty}^{\infty} \rho(k) = 0$.
- Processes with short-range dependence (SRD)[2]—$H = 1/2$. The ACF $\rho(k)$ exhibits an exponential decay to zero, such that $\sum_{k=-\infty}^{\infty} \rho(k) = c$, where $c > 0$ is a finite constant.
- Processes with long-range dependence (LRD)—$(1/2) < H < 1$. The ACF $\rho(k)$ is a slowly-vanishing function, meaning a dependence degree even between samples that are far apart or $\sum_{k=-\infty}^{\infty} \rho(k) = \infty$.

Some authors [12] classify stochastic processes with long-range dependence $(H > (1/2))$ presence as self-similar or fractal processes. However, an LRD process can only be considered as self-similar if it also shows distribution-invariance for any process time-increment. Mandelbrot [3] proposed the most important self-similar process known as *fractional Brownian motion* (fBm). This is derived from a pure Brownian motion where $H = 1/2$. Only for fractal or self-similar processes, one can relate the H parameter to a fractal dimension (D_h) through the equation $D_h = 2 - H$. Examples where the fractal dimension are used in pattern recognition studies can be found in [5] and [6].

As previously mentioned, the fractal dimension has already been used for discriminating fricative sounds and for speaker identification. The studies presented in [7] and [8] assumed the hypothesis that speech is a fractal signal. In the present work, however, although a vector of H parameters is adopted as a speech feature, it is not assumed that the speech signal is a fractal or self-similar signal.

In order to analyze the impact of the speech time-dependence on the performance of a speaker recognizer, accurate methods are needed to estimate the H parameter. Several schemes are presented in the literature ([13]–[15]) and have been widely used in the area of signal traffic modeling[3] [16], [17]. The appropriate method to be used in the speaker recognition task has to take into account the need of automatically estimating pH and the overall computational complexity of the estimator. The estimators[4] considered in this work are the ReScaled adjusted range (R/S) statistic [13], [14], the Higuchi [14], [18], and the wavelet-based Abry–Veitch (AV) [15] estimator. The R/S estimator can be used for any type of speech signal distribution. However, the R/S estimation of the H parameter is a time-consuming procedure since it depends on the user visual intervention to define the linear regression region. The Higuchi estimator is only appropriate for fractal stochastic processes, and it cannot be proven that speech signals are fractals. For these reasons, both the R/S and the Higuchi methods were used only in a preliminary study. The AV estimator was the choice in the present work because it allows an automatic estimation of the pH feature. Moreover, it enables the pH feature extraction in real-time and presents a low computational cost when compared to the standard physiological features extraction.

A. Abry–Veitch Estimator (AV)

The AV estimator[5] [19] uses the discrete wavelet transform (DWT) to successively decompose a sequence of samples into approximation $(a(j,k))$ and detail $(d(j,k))$ coefficients, where j is the decomposition scale and k is the coefficient index of each scale. Fig. 1 illustrates an example of the AV estimator using three decomposition scales to obtain a single H value.

The multiresolution analysis [20] adopted in the DWT of the AV estimator is a powerful theory that enables the detail and approximation coefficients to be easily computed by a simple discrete time convolution.

For the simulations experiments, *Daubechies* [15], [20] filters with four, six, and 12 coefficients were employed to obtain the detail and approximation sequences. The linear computational

[1] The H notation is used for a single *Hurst* parameter. The proposed feature is a vector of H parameters and is denote by pH.

[2] Markov processes are well known examples of SRD processes.

[3] The presence of the traffic time-dependence or scaling degree may cause severe impact on the communications performance. See *IEEE Signal Processing Magazine*, Vol. 19, no. 3, May 2002.

[4] These estimators are included in a *Hurst Estimator Package* (HEP) developed in [17]. The HEP is available and interested readers shall send a request to coelho@ime.eb.br.

[5] The AV estimator is available at http://www.emulab.ee.mu.oz.au/darryl.

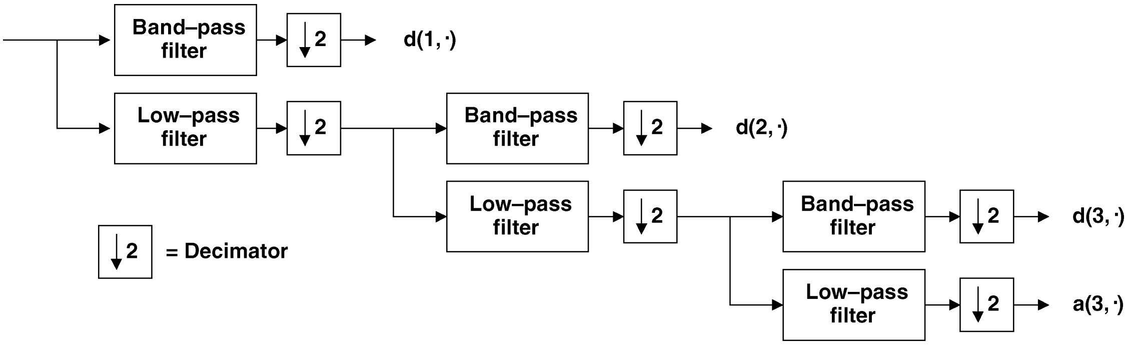

Fig. 1. Example of the AV estimator using three decomposition scales.

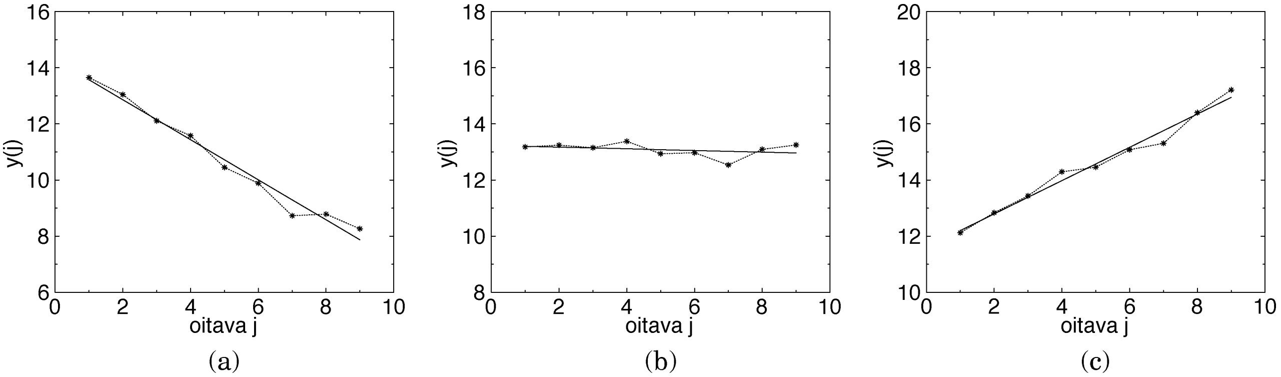

Fig. 2. Example of the AV estimation of the *Hurst parameter*. (a) $H = 0.2$. (b) $H = 0.5$. (c) $H = 0.8$.

complexity of the pyramidal algorithm to obtain the DWT is $O(n)$, where n is the signal samples length [19]. It is important to note that the computational complexity of the fast Fourier Transform (FFT), used to obtain the mel-cepstral coefficients, is $O(nlog(n))$. The AV estimation can be described in three main phases

1) Wavelet decomposition: the DWT is applied to the sample data generating the detail coefficients $d(j,k)$.
2) Variance estimation of the detail coefficients: for each scale j, the variance $\sigma_j^2 = (1/n_j)\sum_k d(j,k)^2$ is evaluated, where n_j is the number of available coefficients for each scale j. It can be shown [19] that

$$E\left[\sigma_j^2\right] = c_\gamma j^{2H-1}$$

where c_γ is a constant.
3) *Hurst* parameter estimation: plot $y_j = \log_2(\sigma_j^2)$ versus j. Using a weighted linear regression, one get the slope α of the plot, and the H parameter is estimated as $H = (1+\alpha)/2$. An example of the final stage of the estimation process for a speech signal with $H = 0.2$, $H = 0.5$, and $H = 0.8$ is depicted in Fig. 2.

The AV estimator is a good choice for the ASkR application due to its simplicity, low computational cost and the possibility of real-time implementation. However, the AV estimator is appropriate to stochastic processes that have only a single value of the time-dependence or scaling degree. They are known as monodependent processes. The proposed method—*M_dim_wavelets*—is based on the AV estimator but provides more than one H parameter for each segment of speech. The *M_dim_wavelets* is a cascade of AV stages applied to each decomposition scale or sequence. This multidimensional estimation leads to a vector of H parameters that composes the pH feature. The *M_dim_wavelets* method is described in the following subsection. Fig. 3 shows an example of the *M_dim_wavelets* estimation also considering three decomposition stages.

B. Wavelet-Based Multidimensional Estimator

Like the AV estimator, the *wavelet-based multidimensional* proposed estimator—*M_dim_wavelets*—uses the DWT to successively decompose a sequence of samples into the detail and approximation coefficients. From each detail sequence $d(j,k)$ generated by the filter bank in a given scale j, an H parameter is estimated, H_j. The set of H_j values and the H value obtained for the entire speech signal (H_0) compose the pH feature. The *M_dim_wavelets* estimator can be described in two main steps.

1) Wavelet decomposition: the DWT is applied to the speech samples generating the detail sequences $d(j,k)$.
2) pH estimation: application of the AV estimator (see Fig. 1) to the entire speech signal (H_0) and then to each of the J detail sequences obtained in the previous step (see Fig. 3). The resulting $(J+1)$ H values will compose the pH feature.

C. pH Extraction

Two procedures were examined for the pH extraction: *frame-by-frame* and *cumulative*.

1) Frame-by-Frame: The speech signal is split into N frames (with overlapping) and the proposed estimator *M_dim_wavelets* is applied to each speech frame. This means

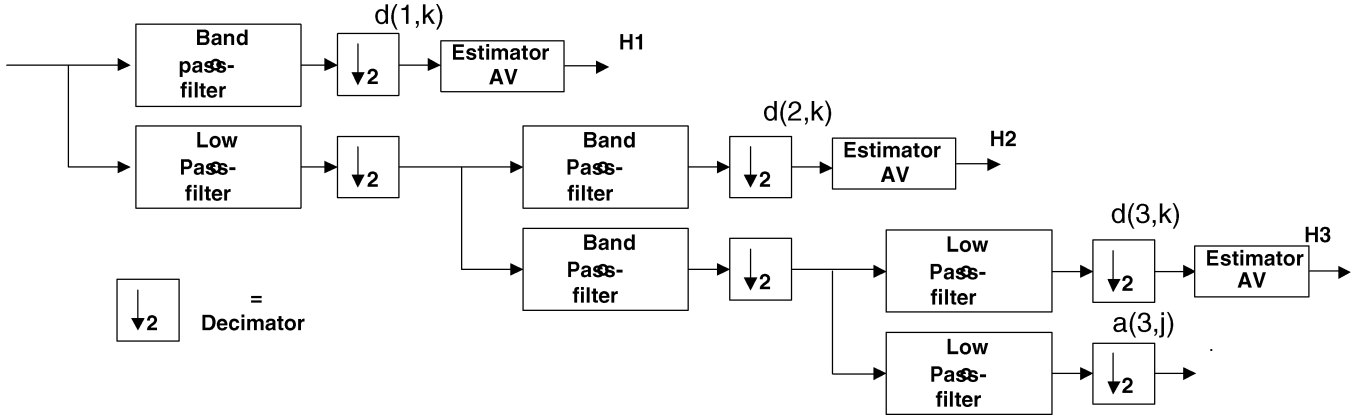

Fig. 3. *M_dim_wavelets* estimation example.

that at each frame n, several H parameter values are estimated. In this study, the pH matrix—containing the pH vectors along the frames—was obtained from 80 ms frames with 50% overlapping.

2) Cumulative: In this procedure, the proposed *M_dim_wavelets* estimator is applied at different time instants taking into account all the past history of the signal. This means that if the time instants $t_1, t_2, \ldots, t_n$ are chosen, the estimator will be applied to the 0 to t_1, 0 to t_2,…, 0 to t_n intervals.

In the cumulative procedure, the evolution of the H parameter values converges to one H value. This single H result is equal to the H_0 value obtained from the frame-by-frame extraction, i.e., an estimation of H obtained directly from the entire speech signal samples without a previous windowing procedure. Preliminary tests revealed that due to H values variability, the pH estimation employed on a frame-by-frame basis provides better speaker recognition results. Hence, this approach was adopted in the remaining experiments of the present work.

From several experiments, it was found that a good configuration for extraction of the pH feature matrix is given by the following specifications:

1) frame duration: 80 ms;
2) *Daubechies* wavelets [21] with 12 coefficients;
3) number of decomposition scales for the H_0: 6;
4) coefficient range from 3 to 5.

III. GMM, AR, AND dB CLASSIFIERS

Before introducing the *M_dim_fBm*, a brief description of the GMM, AR, and dB classifiers is presented in this section. The performance of these classifiers will be latter compared to the proposed scheme.

A. GMM

A mixture of Gaussian probability densities is a weighted sum of M densities, and is given by

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{3}$$

where $\vec{x}$ is a random vector of dimension D, $b_i(\vec{x})$, $i = 1, \ldots, M$ are the density components, and $p_i, i = 1, \ldots, M$ are the mixture weights. Each component density is a D variate Gaussian function of the form

$$b_i(\vec{x}) = \frac{e^{\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T K_i^{-1}(\vec{x}-\vec{\mu})\right)}}{(2\pi)^{\frac{D}{2}}\sqrt{|K_i|}} \tag{4}$$

with mean vector $\vec{\mu}_i$ and covariance matrix K_i, where T denotes the transpose operation and $|.|$ is the determinant.

The Gaussian mixture model λ is parameterized by mean vectors, covariance matrices, and mixture weights. These parameters are jointly represented by the following notation: $\lambda = \{p_i, \vec{\mu}_i, K_i\}\ i = 1, \ldots, M$. The model parameters are estimated for a set of training data as the ones that maximize the likelihood of the GMM. In this paper, we obtain the parameter estimates using a special case of the expectation-maximization (EM) algorithm [2]. For a sequence of T independent training vectors $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$, the normalized log-likelihood of the GMM is given by

$$\log p(X|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log p(\vec{x}_t|\lambda). \tag{5}$$

The decision rule for the speaker identification system chooses the speaker model for which this value is maximum. The speaker verification system requires a binary decision, accepting or rejecting a pretense speaker. It uses two models, which provide the normalized logarithmic likelihood with input vectors $\vec{x}_1, \ldots, \vec{x}_T$: one from the pretense speaker and another one trying to minimize the variations not related to the speaker (*background* model), providing a more stable decision threshold [2]. If the system output value (difference between the two likelihoods) is higher than a given threshold θ, the speaker is accepted; otherwise, it is rejected. The *background* is built with a hypothetical set of false speakers and modeled via GMM. The threshold is calculated on the basis of experimental results.

B. AR-Vector

The AR-vector is actually an extension of the LPC in the sense that it carries out a prediction among vectors (not samples), modeling the time evolving of the vectors. The order p

AR-vector model for a sequence of N vectors of dimension D, in time domain, is given by

$$X_n = \sum_{k=1}^{p} A_k X_{n-k} + E_n \tag{6}$$

where X_n and E_n are D dimensional vectors, with E representing the linear prediction error and A_k being the $D \times D$ prediction matrix. The set of prediction matrices can be represented by a $D \times (p+1)$ matrix $\mathbf{A} = [\mathbf{A}_0\ \mathbf{A}_1\ \mathbf{A}_2 \ldots \mathbf{A_p}]$, with $A_0 = I$ (identity matrix).

From the vectors X_n, it can be defined an estimate of the autocorrelation matrix

$$R_k = \sum_{n=1}^{N-k} X_n X_{n+k}^T \tag{7}$$

where N is the number of vectors X. Note that R_k results in a $D \times D$ matrix. The matrices A_k are obtained by solving a set of equations that depends on R_k.

The use of the AR in speaker recognition requires a measure to evaluate the similarity between two autoregressive models. A widely used measure is the Itakura distance [22], which provides the distance between two all-pole LPC models, based on the linear prediction coefficients and on the autocorrelation matrix [23]

$$\mathbf{R} = \begin{pmatrix} R_0 & R_1^T & \ldots & R_p^T \\ R_1 & R_0 & \ldots & R_{p-1}^T \\ \vdots & \vdots & \ddots & \vdots \\ R_p & R_{p-1} & \ldots & R_0 \end{pmatrix}. \tag{8}$$

Assuming a stored model $\mathbf{A}$ previously estimated from a given speaker with autocorrelation matrix $\mathbf{R_A}$ and a model $\mathbf{B}$ from a pretense speaker with autocorrelation matrix $\mathbf{R_B}$, different distance measures between these two models can be defined. In this paper, the symmetric distance (d_{sim}) was used and it is given by

$$d_{\mathrm{sim}} = \frac{1}{2}\left[\log\left(\left[\frac{tr(\mathbf{AR_B A^T})}{tr(\mathbf{BR_B B^T})}\right]\right) + \log\left(\left[\frac{tr(\mathbf{BR_A B^T})}{tr(\mathbf{AR_A A^T})}\right]\right)\right] \tag{9}$$

where tr is the trace matrix function. In the speaker identification system, the identified speaker is the one for which this distance is minimum. The speaker verification system provides a binary output, acceptance or rejection of a pretense speaker. Hence, an estimation of a threshold θ, based on true and false utterances, is required. The threshold is estimated taking into account false acceptance and false rejection errors. When a speaker is to be analyzed, he/she will be accepted if the resulting distance is lower than the threshold, and rejected otherwise.

C. Bhattacharyya Distance

The *Bhattacharyya distance*, or simply dB, is described in several pattern recognition texts, such as [24]. Consider the following notation:

- ω_i: class i, i $= 1, 2, \ldots$;
- μ_i: mean vector of ω_i;
- C_i: covariance matrix of ω_i.

The dB measures the separability between two Gaussian distributions and is defined by

$$dB = \frac{1}{2}\ln\frac{\frac{|C_i+C_j|}{2}}{|C_i|^{\frac{1}{2}}|C_j|^{\frac{1}{2}}} + \frac{1}{8}(\mu_i-\mu_j)^T\left(\frac{C_i+C_j}{2}\right)^{-1}(\mu_i-\mu_j). \tag{10}$$

This equation can be rewritten as

$$dB = d_C + d_M \tag{11}$$

where d_C represents the distance due to the difference between the covariance matrices of the two classes, and d_M characterizes the difference between the mean vectors of these classes.

For applications in speaker recognition systems, dB directly compares the characteristics of the reference (test) pattern with the target speaker. Therefore, the lower is the dB value, the higher is the probability that the test pattern belongs to the target speaker.

IV. M_dim_fBm Classifier

The *M_dim_fBm* classifier proposed in this paper models each speaker on the basis of the speech features time-dependence or scaling characteristics. The speech signals are not assumed to be fractals. The most important stochastic processes that can represent time-dependence or scaling characteristics are the*fractional Brownian motion* (fBm), the *fractional Gaussian noise* (fGn), and the *fractional Autoregressive Moving Average* (f-ARIMA) models.[6] The *M_dim_fBm* is based on the fBm process.

A. Fractional Brownian Motion

The fBm [3] is a Gaussian stochastic process $(X_H(t))$ indexed in $\Re$ with zero mean and continuous sample path (null at origin). The fBm is known as the unique gaussian H-sssi, i.e., self-similar with self-similarity parameter and stationary increments. The variance of the independent increments is proportional to its time interval accordingly to the expression

$$Var\left[X_H(t_2) - X_H(t_1)\right] \propto |t_2 - t_1|^{2H} \tag{12}$$

for all instants t_1 and t_2 and

1) $X_H(t)$ has stationary increments;
2) $X_H(0) = 0$ and $E[X_H(t)] = 0$ for any instant t;
3) $X_H(t)$ presents continuous sample paths.

The fBm is considered a self-similar process since its statistical characteristics[7] hold for any time scale. In other words, for any τ and $r > 0$

$$\left[X_H(t+\tau) - X_H(t)\right]_{\tau \leq 0} \overset{d}{\approx} r^{-H}\left[X_H(t+r\tau) - X_H(t)\right]_{\tau \leq 0} \tag{13}$$

where $\overset{d}{\approx}$ means equal in distribution, and r is the process scaling factor ($r = \tau = |t_2 - t_1|$). Note that $X_H(t)$ is a Gaussian

[6] The fGn and f-ARIMA processes are appropriate only for stochastic processes with $H > (1/2)$ (long-range dependence), while the fBm models stochastic processes can represent any value of H $(0 < H < 1)$.

[7] *Statistical characteristics* means marginal distribution and time-dependence degree.

process completely specified by its mean, variance, H parameter, and ACF given by [12]

$$\rho(k) = \frac{1}{2}\left[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}\right] \qquad (14)$$

for $k \geq 0$ and $\rho(k) = \rho(-k)$ for $k < 0$.

B. Description of the *M_dim_fBm* Classifier

As previously mentioned, the fBm is a monofractal stochastic process, i.e., it uses a single H parameter value. In order to be suitable for applications in ASkR systems, it was developed a novel classification scheme called *multidimensional fractional Brownian motion* (*M_dim_fBm*). Similarly to the GMM classification procedure, our statistical classification scheme is based on the input features models. The *M_dim_fBm* model of a given speaker is generated according to the following steps.

1) *Preprocessing*: the feature matrix formed from the input speech features[8] is split into r regions. This matrix contains c rows, where c is the number of feature coefficients per frame, and N columns, where N is the number of frames.
2) *Decomposition*: for each row of the feature matrix in a certain region, the *wavelet* decomposition is applied in order to obtain the *detail* sequences.
3) *Parameters Extraction/Estimation*: from each set of *detail* sequences obtained from each row of step 2), estimate the mean, the variance and the H parameters of the features being used by the ASkR system. For the H parameter estimation, use the AV *wavelet-based* estimator proposed in [15].
4) *Generation of fBm Processes*: using the *Random Midpoint Displacement* (RMD) algorithm [3] and the three parameters computed in step 3), generate the fBm processes. Therefore, c fBm processes are obtained for a given region.
5) *Determining the Histogram and Generating the Models*: compute the histogram of each fBm process of the given region. The set of all histograms defines the speaker c-dimensional model for that region.
6) *Speaker Model*: the process is repeated for all of the r regions. This means that a $r.c$-dimensional fBm process is obtained, which defines the speaker *M_dim_fBm* model.

In the phase of tests, the histograms of the speaker, obtained from the *M_dim_fBm* model, are used to compute the probability that a certain c-dimensional feature vector $\vec{x}$ belongs to that speaker. This is performed to the N feature vectors, resulting in N probability values: $p_1, p_2, \ldots, p_N$. Adding these values, the measure of the likelihood that the set of feature vectors under analysis belongs to that particular speaker is obtained.

V. Experiments Results and Discussions

In this section, the results of identification (Part A) and verification (Part B) performances of the proposed SR_{Hurst} system are compared to those of other ASkR schemes reported in the literature.

TABLE I
Average Duration of the Training and Tests Speech

Text type	Time Duration (in sec)
Training/Fixed phone	142
Test/Fixed phone	197
Training/Cellular phone	140
Test/Fixed phone	196

TABLE II
Results Precison (δ) for Different Number of Test Experiments for a 95% Confidence Degree

Test Duration	Number of tests	δ
20 s, Fixed phone	753	0.081
10 s, Fixed phone	1541	0.057
5 s, Fixed phone	3114	0.040
20 s, Cellular phone	730	0.083
10 s, Cellular phone	1493	0.058
5 s, Cellular phone	3028	0.041

It is important to remark that in all experiments with the SR_{Hurst}, the *M_dim_fBm* classifier was used with $r = 1$ region only. This means that the speaker model is defined by a c-dimensional fBm process, where c is the number of feature coefficients.

A. Database

The database[9] (BaseIME) used in the experiments is composed of 75 speakers (male and female, 2:1) from 27 Brazilian regions that read two different texts. To record the speech signal, the speakers called a free automatic communication center using first a fixed phone and then a cellular phone. Hence, two databases are available in which each speaker recorded four different text files (i.e., two different training and tests speech texts recorded from fixed and cellular phones). The speech average duration of the training and tests is given in Table I. Tests experiments were applied to 20-, 10-, and 5-s speech segments. A separate speech segment of 1-min duration was used to train a speaker model. Table II shows the recognition results precision for the different test durations and the number of tests. The precision values (δ), considering a confidence degree of 95%, were obtained by using the Chebyshev inequality [25].

B. Identification

The performance results of the identification systems—*M_dim_fBm*, GMM, AR, and dB—are presented in terms of the recognition accuracy. The results using the pH on a frame-by-frame basis, to be presented in Section V-B1, satisfy the low computational cost requirement [26]. On the other hand, the use of the mel-cepstrum and the joint use of the mel-cepstrum and the pH, to be presented in Section V-B2, are useful in applications where the computational cost is not of major concern [27].

1) pH Feature: Table III shows the speaker recognition accuracy of the identification systems based on the pH, for speech signals recorded from a fixed telephony channel.

[8] Note that the *M_dim_fBm* classifier is not constrained to the proposed pH. It can be used with any selected set of speech features.

[9] This database was developed by the Electrical Engineering Department of the *Instituto Militar de Engenharia (IME)* under a project jointly sponsored by FAPERJ and the Security Department of Rio de Janeiro. The database is available and interested readers can send a request to coelho@ime.eb.br.

TABLE III
RECOGNITION ACCURACY (%) OF THE IDENTIFICATION SYSTEMS BASED ON THE pH, FOR SPEECH SIGNALS RECORDED FROM A FIXED TELEPHONY CHANNEL

Testing Interval	M_dim_fBm	GMM	AR	dB
20 s	95.48	95.48	91.24	82.20
10 s	94.22	94.09	83.39	72.62
5 s	89.98	89.69	64.45	56.13

TABLE IV
RECOGNITION ACCURACY (%) OF THE IDENTIFICATION SYSTEMS BASED ON THE pH, FOR SPEECH SIGNALS RECORDED FROM A CELLULAR TELEPHONY CHANNEL

Testing Interval	M_dim_fBm	GMM	AR	dB
20 s	87.53	86.85	79.86	74.66
10 s	84.93	84.89	72.07	62.29
5 s	61.43	61.10	52.41	44.78

TABLE V
RECOGNITION ACCURACY (%) OF THE IDENTIFICATION SYSTEMS BASED ON THE 15 MEL-CEPSTRAL COEFFICIENTS, FOR SPEECH SIGNALS RECORDED FROM A FIXED TELEPHONY CHANNEL

Testing Interval	$MdimfBm$	GMM	AR	dB
20 s	98.54	97.95	96.81	95.48
10 s	97.99	97.99	95.13	93.38
5 s	97.59	97.46	59.47	84.17

TABLE VI
RECOGNITION ACCURACY (%) OF THE IDENTIFICATION SYSTEMS BASED ON THE FUSION USE OF THE pH AND THE MEL-CEPSTRAL COEFFICIENTS, FOR SPEECH SIGNALS RECORDED FROM A FIXED TELEPHONY CHANNEL

Testing Interval	$MdimfBm$	GMM	AR	dB
20 s	98.57	98.40	97.21	95.75
10 s	98.62	98.51	92.54	94.61
5 s	97.91	97.66	72.83	87.81

TABLE VII
RECOGNITION ACCURACY (%) OF THE IDENTIFICATION SYSTEMS BASED ON THE FUSION USE OF THE pH AND THE MEL-CEPSTRAL COEFFICIENTS, FOR SPEECH SIGNALS RECORDED FROM A CELLULAR TELEPHONY CHANNEL

Testing Interval	$MdimfBm$	GMM	AR	dB
20 s	98.19	98.14	88.22	85.75
10 s	92.56	92.03	84.80	74.68
5 s	89.96	89.96	76.09	52.28

Table IV presents the results for speech signals recorded from a cellular telephony channel.

As can be seen from these tables, the best results were obtained with the *M_dim_fBm* and the GMM classifiers for both cellular and fixed telephone recordings.

It is important to remark that for the pH feature used, only $7H$ parameters per speech frame. This implies in a lower complexity of the classifiers as compared to the systems operating on 15 mel-cepstral coefficients per frame. Moreover, it should be reminded that the estimation of the pH feature demands less computational complexity $(O(n))$ than the extraction of the mel-cepstral coefficients (the FFT computational complexity is $O(nlog(n))$.

2) pH + Mel-Cepstral Coefficients: In this second set of experiments, the speaker recognition accuracy of the identification systems was examined for the mel-cepstral coefficients and for the fusion of the mel-cepstral coefficients and the pH. These results are presented in Tables V and VI, respectively, for speech signals recorded from a fixed telephony channel. From a comparative analysis of these tables it can be verified that, except for the 10-s tests with the AR classifier, the best results were achieved by the systems based on the joint use of the mel-cepstral coefficients and the pH. This means that the proposed pH feature aggregates new information on the speaker identity. Again, in all the cases, the GMM and the *M_dim_fBm* classifiers provided the best results with a slight superiority of the latter one.

Recognition accuracies of the identification systems based on the fusion use of the pH and the mel-cepstral coefficients, for speech signals recorded from a cellular telephony channel, are shown in Table VII. Comparing the results of Tables VI and VII, it can been seen that the *M_dim_fBm* and GMM classifiers are more robust to the effects of the cellular telephony channel than the AR and dB models. Note, for instance, the case of the 20-s testing interval. While the *M_dim_fBm* and GMM performances are degraded around 3%, the AR and dB classifiers performances decrease around 10%.

C. Verification

The performance results for the speaker verification systems were obtained by varying the threshold and computing the miss (false rejection) (f_r) and the false alarm (false acceptance) (f_a) probabilities. These error probabilities are plotted as *Detection Error Tradeoff* (DET) curves [28]. The *Universal Background Model* (UBM) was used as *background* model [29] constructed from speech material of 20 speakers that do not belong to the set of 75 speakers used for the testing experiments.

1) pH: In this section, the proposed SR_{Hurst} ASkR system is compared to a scheme that is also based on the pH but uses a GMM classifier. Fig. 4 shows the DET curves for the verification systems based on the pH. The *M_dim_fBm* (SR_{Hurst}) and GMM classifiers results are presented for 20-, 10- s, and 5-s tests.

Table VIII presents the equal recognition rates (ERR) for the operating point of the DET curve where $f_r = f_a$. This measure is given by $ERR = (1 - EER)100\%$, where EER is the equal error rate usually employed in the literature.

From Table VIII, it can be seen that the *M_dim_fBm* achieved better performance results for 20-s testing and fixed telephony channel. In this condition, the *M_dim_fBm* classifier is almost 2% superior to GMM in terms of ERR performance. For the other cases, the ERR is comparable for both systems. However, the DET curves show that for most of the operating points (miss probability x false alarm probability) the proposed classifier provides better results. Note that the performance gains are substantial for a wide range of medium to low false alarm probabilities. It is important to remark that in most ASkR applications, high false alarm probabilities must be avoided.

The results presented in this section show the superior modeling procedure of the *M_dim_fBm* strategy for the speaker verification task.

2) pH + Mel-Cepstral Coefficients: In this section, the fusion use of the pH and the mel-cepstral coefficients is examined for the best speaker verification systems: the *M_dim_fBm*

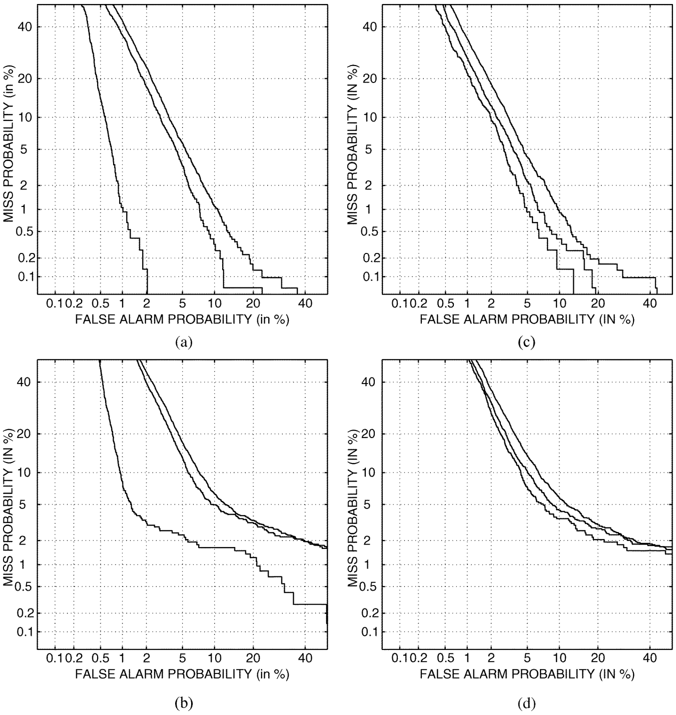

Fig. 4. DET curves for the system based on the pH using the *M_dim_fBm* and the GMM classifiers. From the lower to the upper curves are the results corresponding to 20-, 10-, and 5-s tests, respectively. (a) *M dim_fBm* fixed. (b) *M_dim_fBm* cellular. (c) GMM fixed. (d) GMM cellular.

TABLE VIII
ERR (%) OF THE VERIFICATION SYSTEMS BASED ON THE pH, FOR SPEECH SIGNALS RECORDED FROM FIXED AND CELLULAR TELEPHONY CHANNELS

	$MdimfBm$ fixed	GMM fixed	$MdimfBm$ cel	GMM cel
20 s	98.93	96.67	97.12	96.69
10 s	95.66	95.66	92.98	93.01
5 s	94.76	94.37	92.67	92.11

and the GMM. These systems were compared to the ones based only on the mel-cepstral coefficients.

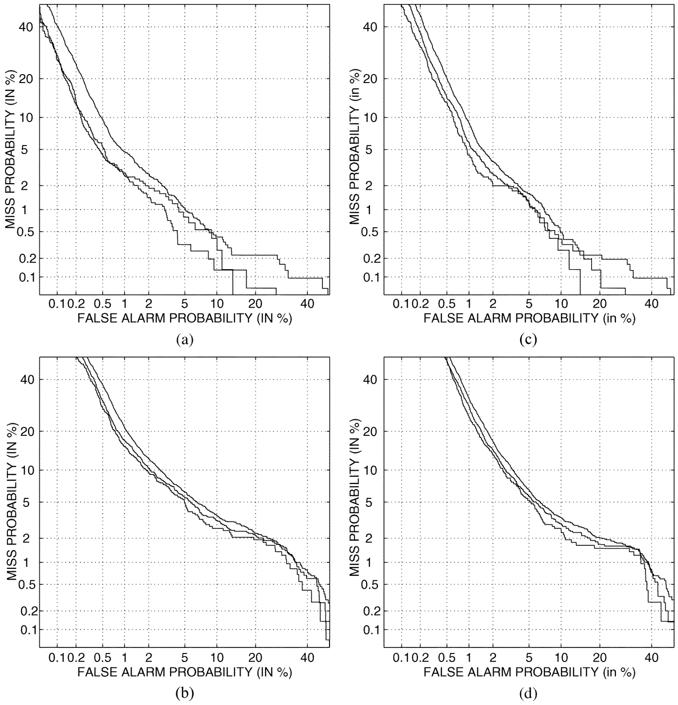

Fig. 5. DET curves for the systems based on the mel-cepstral coefficients using the *M_dim_fBm* and the GMM classifiers. From the lower to the upper curves are the results corresponding to 20-, 10-, and 5-s tests, respectively. (a) *M_dim_fBm* fixed. (b) *M_dim_fBm* cellular. (c) GMM fixed. (d) GMM cellular.

Fig. 5 shows the DET curves for the systems based only on the mel-cepstral coefficients, while Fig. 6 depicts the DET curves for the systems based on the fusion use of the pH and the mel-cepstral coefficients. Tables IX and X show the equal recognition rate performances, i.e., the ERR for $f_r = f_a$ operating points of the DET curves.

For the systems based on the mel-cepstral coefficients, once again, the DET curves show that for most of the operating points, the *M_dim_fBm* performance gains over GMM are also substantial for a wide range of medium to low false alarm probabilities. It should be reminded that this is a desirable range of operation for most speaker verification applications. On the other hand, the classifiers are comparable if they are based on the joint use of the pH and the mel-cepstral coefficients, when the performance are the highest ones.

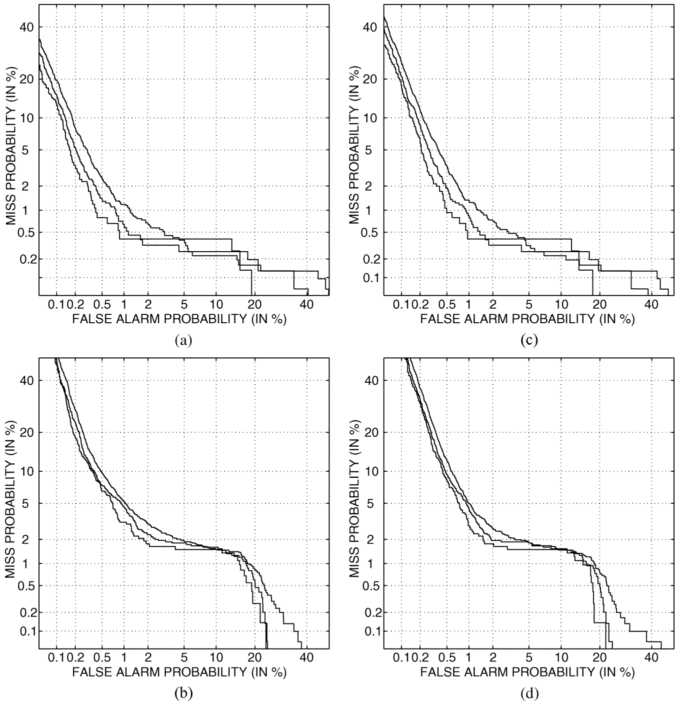

Fig. 6. DET curves for the system based on the fusion use of the pH and the mel-cepstral coefficients using the *M_dim_fBm* and the GMM classifiers. From the lower to the upper curves are the results corresponding to 20-, 10-, and 5-s tests, respectively. (a) *M_dim_fBm* fixed. (b) *M_dim_fBm* cellular. (c) GMM fixed. (d) GMM cellular.

These figures and tables show the improvement of the recognition accuracy when the systems are based on the fusion of the pH and the mel-cepstral coefficients. It is also important to note the 1% average recognition accuracy gain of the speaker verification systems based on the fusion of the pH and the mel-cepstral coefficients over the system based on the mel-cepstral coefficients for fixed telephone speech. This average gain increases to 3% for cellular telephone speech.

TABLE IX
ERR (%) OF THE VERIFICATION SYSTEMS BASED ON THE MEL-CEPSTRAL COEFFICIENTS, FOR SPEECH SIGNALS RECORDED FROM FIXED AND CELLULAR TELEPHONY CHANNELS

	M_dim_fBm	GMM	M_dim_fBm	GMM
	fixed	fixed	cel	cel
20 s	98.08	98.00	95.06	94.93
10 s	98.31	97.59	94.67	94.63
5 s	97.62	97.27	94.34	94.32

TABLE X
ERR (%) OF THE VERIFICATION SYSTEMS BASED ON THE FUSION USE OF THE pH AND THE MEL-CEPSTRAL COEFFICIENTS, FOR SPEECH SIGNALS RECORDED FROM FIXED AND CELLULAR TELEPHONY CHANNELS

	M_dim_fBm	GMM	M_dim_fBm	GMM
	fixed	fixed	cel	cel
20 s	99.33	99.20	98.21	98.09
10 s	99.15	99.09	97.96	98.01
5 s	98.87	98.81	97.49	97.42

The results presented in this section corroborate the superior modeling procedure of the *M_dim_fBm* strategy for the speaker verification task. Moreover, the *M_dim_fBm* results were achieved for a simpler model with dimension equal to 15. Each fBm is characterized by only three scalar parameters (i.e., mean, variance, and H). On the other hand, the GMM used 32 gaussians, each one characterized by one scalar parameter, one mean vector, and one covariance matrix to achieve the performance results. This means that the *M_dim_fBm* classifier yields a better modeling accuracy with a lower computational load.

VI. ISSUES ON MISMATCH CONDITIONS

The performance results presented in this paper, demonstrated the potential of the proposed new feature and the new classifier for speaker recognition. However, it is also very important that the performances of the speaker recognition systems be examined in conditions of training and testing mismatch.

Interesting results on mismatch conditions of speaker identification, based on cepstral features, have been reported in the literature [30], [31]. These works analyzed different mismatch situations concerning microphones, recordings sessions, languages [30], and channel mismatch [31]. The results have shown a severe degradation of the recognition rates under these mismatch situations. Furthermore, these works presented a detailed study on compensation techniques for mismatch conditions. Although, in several situations, a significant improvement of the recognition rate was achieved, in other situations, this was not possible. For instance, in [31], it was shown that for the N-TIMIT database, the recognition rate using the best compensation technique proposed in that paper, achieved an improvement of only 17.4% [i.e., 50.5% (without compensation) and 67.9% (with compensation)]. Note that 67.9% is not an acceptable performance for any application. The experiments carried out in [31] have also shown that the recognition rates using the classical cepstral mean substraction (CMS) technique were even worse than without any compensation technique [i.e., 50.5% (without compensation) and 43.2% (with CMS)].

Considering the above discussion, it is possible that the speaker recognition rates obtained in this paper may change in mismatch conditions. However, if this situation is confirmed, as in the cepstral features studies [30], [31], a detailed investigation on different compensation techniques, eventually including new proposals, will be required to improve the recognition rates of the speaker recognition systems. This is certainly a very interesting research topic for future works.

VII. CONCLUDING REMARKS

In this paper, a novel speaker recognition system—the SR_{Hurst}—is presented based on a new speech statistical feature, the pH, and a new classifier, the multidimensional fractional Brownian motion—the *M_dim_fBm*. The *M_dim_fBm* performance was compared to the GMM, AR, and dB classifiers.

For applications requiring a low computational cost, the systems employing only the proposed pH feature have shown to be an attractive choice in terms of extraction and classification procedures. On the other hand, if computational complexity is not of major concern, the best strategy—due to its highest performance—is the one based on the fusion of the pH feature and the mel-cepstral coefficients.

The results presented in this paper show that the *M_dim_fBm* requires less computational load and provides a more accurate modeling strategy as compared to the GMM. It is also shown that the pH feature adds substantial information to the systems based on the mel-cepstral coefficients. It can be concluded, therefore, that the SR_{Hurst} is a very attractive tool to the area of automatic speaker recognition and represents an important contribution due to its performance and simplicity.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their important remarks and suggestions that helped to improve this work.

REFERENCES

[1] D. O'shaughnessy, *Speech Communication*, 2nd ed. Piscataway, NJ: IEEE Press, 2000.

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

[3] M. Barnsley *et al.*, *The Science of Fractal Images*. New York: Springer-Verlag, 1988.

[4] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, 2nd ed. New York: Wiley, 2003.

[5] R. Esteller, G. Vachtsevanos, and T. Henry, "Fractal dimensions characterizes seizure onset in epileptic patients," in *Proc. IEEE ICASSP*, vol. 4, 1999, pp. 2343–2346.

[6] T. Morimoto *et al.*, "Pattern recognition of fruit shape based on the concept of chaos and neural networks," *Comput. Electron. Agriculture*, vol. 26, pp. 171–186, 2000.

[7] S. Fernández, S. Feijóo, and R. Balsa, "Fractal characterization of spanish fricatives," in *Proc. ICPhS*, 1999, pp. 2145–2148.

[8] A. Petry and D. Barone, "Fractal dimension applied to speaker identification," in *Proc. IEEE ICASSP*, 2001, pp. 405–408.

[9] D. Reynolds, R. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, Jan. 2000.

[10] C. Montacié and J. L. LE Floch, "Ar-vector models for free-text speaker recognition," in *Proc. ICSLP*, 1992, pp. 611–614.

[11] A. Petry, A. Zanuz, and D. Barone, "Bhattacharyya distance applied to speaker identification," in *Proc. ICSPAT*, Dallas, TX, 2000.

[12] J. Beran, *Statistics for Long-Memory Processes*. London, U.K.: Chapman & Hall, 1994.

[13] E. Hurst, "Long-term storage capacity of reservoirs," *Trans. Amer. Soc. Civil Eng.*, pp. 770–799, Apr. 1951.

[14] M. Taqqu, V. Teverovsky, and W. Willinger, "Estimators for long-range dependence: An empirical study," *Fractals*, vol. 3, no. 4, pp. 786–789, Jul. 1995.

[15] D. Veith and P. Abry, "A wavelet-based joint estimator of the parameters of long-range dependence," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 878–897, Mar. 1998.

[16] W. Leland, W. Willinger, M. Taqqu, and D. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Trans. Netw.*, vol. 2, no. 1, pp. 1–15, Feb. 1994.

[17] R. Pontes and R. Coelho, "The scaling characteristics of the video traffic and its impact on acceptance regions," in *Proc. 17th Int. Teletraffic Congr.*, vol. 4, Dec. 2001, pp. 197–210.

[18] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physics D.*, vol. 31, pp. 277–283, 1988.

[19] M. Roughan, D. Veith, and P. Abry, "Real-time estimation of the parameters of long-range dependence," *IEEE/ACM Trans. Netw.*, vol. 8, no. 4, pp. 467–478, Aug. 2000.

[20] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Englewood Cliffs: Prentice-Hall, 1985.

[21] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.

[22] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoustics, Speech Signal Process.*, vol. 23, no. 1, pp. 42–47, Feb. 1975.

[23] F. Bimbot, L. Mathan, A. Lima, and G. Chollet, "Standard and target driven ar-vector models for speech analysis and speaker recognition," in *Proc. ICASSP*, Mar. 1992, pp. 5–8.

[24] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic Press, 1990.

[25] A. Allen, *Probability, Statistics and Queueing Theory*. New York: Academic Press, 1978.

[26] J. Kumagai, "Talk to the machine," *IEEE Spectrum*, vol. 39, no. 9, pp. 60–64, Sep. 2002.

[27] The NIST Year 2001 Speaker Recognition Evaluation Plan. NIST, Gaithersburg, MD. [Online] Available: http://www.nist.gov/speech/publications

[28] A. Martin *et al.*, "The det curve in assessment of detection task performance," in *Proc. EuroSpeech*, 1997, pp. 1895–1898.

[29] D. Reynolds, R. Rose, and E. Hosftetter, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 245–267, Apr. 1994.

[30] C. Alonso-Martinez and M. Faúndez-Zanuy, "Speaker identification in mismatch training and testing conditions," in *Proc. ICASSP*, 2000, pp. 1181–1184.

[31] A. Garcia and R. Mammone, "Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping," in *Proc. IEEE ICASSP*, May 1999, pp. 325–328.

Ricardo Sant'Ana received the B.S. and M.Sc. degrees in electrical engineering from the Instituto Militar de Engenharia (IME), Rio de Janeiro, Brazil, in 1997 and 2003, respectively.

In 2003, he joined the Systems Developing Center of the Brazilian Army. His main research interests include speech and speaker recognition.

Rosângela Coelho (M'97) received the Ph.D. degree from the Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, in 1995 and the M.Sc. degree from the Pontifícia Universidade Católica (PUC), Rio de Janeiro, Brazil, in 1991, both in electrical engineering.

She is an Associate Professor in the Electrical Engineering Department, Instituto Militar de Engenharia (IME), Rio de Janeiro. She heads the Optical Network and Systems Laboratory (LaRSO). In 2003, she received the University Research Program grant from CISCO. She also serves as Editorial Board Member of the IEEE Communications Surveys and Tutorials. Her main research interests include speaker and speech recognition, signal processing for communications, signals stochastic modeling, and optical communications.

Abraham Alcaim received the Dipl. and M.Sc. degrees in electrical engineering from the Pontifical Catholic University (PUC), Rio de Janeiro, Brazil, in 1975 and 1977, respectively, and the D.I.C. and Ph.D. degrees in electrical engineering from the Imperial College of Science and Technology, University of London, London, U.K., in 1981.

Since 1976, he has been with the Center for Telecommunication Studies of the Catholic University (CETUC-PUC/Rio), Rio de Janeiro, where he is currently Associate Professor. He has been working for more than 25 years in the area of digital processing of speech and image. He is the author of numerous papers in international conferences and journals. In 1984 he spent a short leave at the Centre National d'Etudes des Télécommunications (CNET), Lannion, France, where he worked on objective and subjective quality measures for speech coders. From December 1991 to September 1993, he was a Visiting Scientist at the IBM Brazil Scientific Center, where he worked on the design of image coders with special application to satellite images. His research interests are in speech, image and video compression, and speech and speaker recognition.

Dr. Alcaim was the Technical Program Chairman of the 1990 and the 1994 SBT/IEEE International Telecommunications Symposiums—ITS'90 and ITS'94—held in Rio de Janeiro in September 1990 and in August 1994, respectively. He was the Executive Chairman of the 1999 IEEE Global Telecommunications Conference—GLOBECOM'99—held in Rio de Janeiro in December 1999.