Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters

Juan Ignacio Godino-Llorente*, *Member, IEEE*, Pedro Gómez-Vilda, *Member, IEEE*, and Manuel Blanco-Velasco, *Member, IEEE*

***Abstract*—Voice diseases have been increasing dramatically in recent times due mainly to unhealthy social habits and voice abuse. These diseases must be diagnosed and treated at an early stage, especially in the case of larynx cancer. It is widely recognized that vocal and voice diseases do not necessarily cause changes in voice quality as perceived by a listener. Acoustic analysis could be a useful tool to diagnose this type of disease. Preliminary research has shown that the detection of voice alterations can be carried out by means of Gaussian mixture models and short-term mel cepstral parameters complemented by frame energy together with first and second derivatives. This paper, using the F-Ratio and Fisher's discriminant ratio, will demonstrate that the detection of voice impairments can be performed using both mel cesptral vectors and their first derivative, ignoring the second derivative.**

***Index Terms*—Cepstral parameters, F-Ratio, fisher's discriminant ratio, Gaussian mixture models, short-term analysis, voice disorders.**

I. Introduction

ACOUSTIC analysis is a noninvasive technique based on the digital processing of the speech signal, which is an efficient tool for the objective support of the diagnosis of voice disorders, the screening of vocal and voice diseases (and particularly their early detection), the objective determination of vocal function alterations and the evaluation of surgical as well as pharmacological treatments and rehabilitation. Its application is not restricted to the medical area alone, as it may also be of special interest in the control of voice quality for voice professionals such as singers, speakers, etc. It furthermore offers two main advantages: it is a noninvasive tool, and it provides an objective diagnosis. It can be a complementary tool for those methods based on the direct observation of the vocal folds using laryngoscopy, as this inspection technique is considered risky and it has to be carried out by a specialist under well controlled conditions. Moreover, it is expensive, time consuming and requires costly resources, such as special light sources, endoscopic instruments and specialized video-camera equipment.

This study is mainly focused on organic pathologies affecting the vocal folds, appearing as a modification of the morphology of the excitation (i.e., vocal folds, increasing the distribution of masses) and producing a more irregular vibration pattern. This group may include pathologies such as polyps, nodules, cysts, sulcus, edemas, carcinoma, etc.

Most of the approaches found in existing literature address the automatic detection of voice alterations by means of long-time signal analysis [1]–[11]. These long-time parameters are generally calculated by averaging local time perturbations measured from the speech, so providing estimations of the degree of normality. Amongst these parameters are the following: *pitch, jitter, shimmer, amplitude perturbation quotient (APQ), pitch perturbation quotient (PPQ), harmonics to noise ratio (HNR), normalized noise energy (NNE), voice turbulence index (VTI), soft phonation index (SPI), frequency amplitude tremor (FATR), glottal to noise excitation (GNE),* etc. Previous studies [12]–[14] indicate that the detection of voice alterations can be carried out by means of the above mentioned long-term acoustic parameters, enabling each individual voice utterance to be quantified by a single vector. However, some of these parameters are based on an accurate estimation of the fundamental frequency, a fairly complex task in the presence of certain pathologies [15], [16].

In recent years, more modern approaches have been devised which use short-time speech analysis [17]–[19] or electroglotographic (*EGG*) [20]–[22] signals. Some of these focus on the automatic detection of voice impairments from the excitation waveform collected with a laryngograph [5], [21] or extracted from the acoustic data by inverse filtering [18], [22]. However, due to the fact that *linear prediction coding (LPC)*-based inverse filtering is based on the assumption of a linear model, such methods do not behave well when pathology is present due to nonlinearities introduced by the pathology itself. Some of the authors concerned have also proposed nonlinear signal processing for the same task [19], [22].

Throughout this research the automatic detection of voice impairments is carried out by means of *Gaussian mixture models* (*GMMs*) using the well-known short-term *mel frequency cepstral coefficients* (*MFCC*) as an alternative to the above mentioned methods. The main advantage of this approach is that it does not exhibit dependency on previous pitch estimations. Since this parameter has demonstrated a good degree of reliability for this task in previous research [23], the main idea in this paper is to reduce the dimensionality of the feature vectors in order to minimize complexity without affecting the recognition

Manuscript received July 8, 2003; revised September 25, 2005. This work was supported in part by the Spanish Ministry of Education under Grant TIC2003-08956-C02-00, Grant TIC2002-0273, and Grant PR2002-0239. *Asterisk indicates corresponding author.*

*J. I. Godino-Llorente is with the Universidad Politécnica de Madrid. EUIT Telecomunicación, Crta. Valencia km 7, 28031 Madrid, Spain (e-mail: igodino@ics.upm.es).

P. Gómez-Vilda is with the Universidad Politécnica de Madrid. Facultad de Informática, Campus de Montegancedo, Boadilla del Monte, 28660 Madrid, Spain.

M. Blanco-Velasco is with the Universidad de Alcalá. Campus Universitario, Alcalá de Henares, 28871 Madrid, Spain.

Digital Object Identifier 10.1109/TBME.2006.871883

performance. For this task, two methods extracted from the classical statistics have been used: the *F-Ratio* and the *Fisher's discriminant ratio*. Both are regarded as feature selection methods in existing literature, providing an idea about the significance of every single feature. The *Fisher* criteria and *F-Ratio* have been used on previous occasions [24] in speech recognition to reduce the vector length in a speech recogniser based on hidden Markov models (*HMM*) and both linear prediction cepstral coefficients (*LPCC*) and *MFCC*. In both cases 12 features and energy were used complemented with the first and second derivatives (up to 38 components), and it was demonstrated that the second order derivatives were the less significant, with some exceptions for the low frequency related features and the differences of energy.

In short-term analysis, the detection is carried out on a frame basis. Each frame is quantified with a vector formed by a feature set. In this approach, each feature set consists of the following features: the energy, L mel-cepstral coefficients, L first temporal derivatives, L second temporal derivatives of these coefficients—known as delta (Δ) mel-cepstrum and delta-delta ($\Delta\Delta$) mel-cepstrum, respectively—and the first and second derivative of the energy (known as delta energy and delta-delta energy). The final decision about the presence or absence of pathology is taken assuming independence between observations by summing the log-probabilities obtained for each frame of a patient's utterance and establishing a threshold to separate both sets. This parameterization approach has been used previously for this task in [23].

II. Database

The company Kay Elemetrics distributes a CD-ROM database of approximately 700 records originally developed by the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Labs [25]. The acoustic samples are the sustained phonation of the vowel /ah/ (1–3 s. long) from patients with normal voices and a wide variety of organic, neurological, traumatic, and psychogenic voice disorders in different stages (from early to mature). The speech samples were collected in a controlled environment and sampled with a 50- or 25-kHz sampling rate and 16 bits of resolution. A downsampling with a previous half band filtering has been carried out over some registers in order to adjust every utterance to the sampling rate of 25 kHz.

Since we are interested in pathologies which affect the vocal folds, the process has been carried out over sustained phonations of the vowel /ah/. This is to ensure that the vocal folds remain in motion throughout the entire utterance.

The database has been segmented according to the criteria explained in [14]. The subset taken from the database contains 53 normal and 173 pathological speakers with a wide range of organic, neurological, traumatic and psychogenic voice disorders (Table IV). The criteria used in [14] ensure that gender and age are uniformly distributed between the two classes (Table I).

III. Methods

The digital speech signal (1–3 s. long) is framed and windowed (a 40 ms. Hamming window was used throughout the various experiments) with no preemphasis [26], [27]. The frames were extracted with a 50% frame shift. The framing was followed by an endpoint detector [28] allowing voiced and unvoiced segments or silences to be separated. The next step was feature extraction, which was necessary in order to reduce the pattern dimensionality and complexity. These feature vectors fed a *Gaussian mixture model*-based detector enabling a final decision about the absence or presence of pathology for every frame to be reached.

TABLE I
Distribution of Normal and Pathological Speakers [14]

	Number		Mean age (years)		Age range (years)		Std. deviation (years)	
	♂	♀	♂	♀	♂	♀	♂	♀
Normal	21	32	38.81	34.16	26-59	22-52	8.49	7.87
Pathol.	70	103	41.71	37.58	26-58	21-51	9.38	8.19

A. Computation of Recognition Features (MFCC)

It is widely recognized that the acoustic signal itself contains information about the vocal tract and the excitation waveform. The basic idea in this research is to use a short-term nonparametric approach capable of modeling the effects of pathologies on both the excitation (vocal folds) and the system (vocal tract), although throughout this paper emphasis has been placed on pathologies which mainly affect the vocal folds.

Mel cepstral coefficients [27] can be estimated by using a parametric approach derived from linear prediction coefficients (*LPC*), or by using a nonparametric *fast Fourier transform* (FFT)-based approach. However, FFT-based *MFCCs* typically encode more information from excitation; whereas *LPC*-based *MFCCs* remove this. This idea is demonstrated in [29], were FFT-based *MFCCs* are found to be more dependent on high-pitched speech resulting from loud or angry speaking styles than *LPC*-based *MFCCs*, which were found to be more sensitive to additive noise in speech recognition tasks. This is the case because *LPC*-based *MFCCs* ignore the pitch-based harmonic structure seen in FFT-based *MFCCs*.

It is well known that pathological voice is induced by an increase of mass, a lack of closure, or a change in the elasticity of the vocal folds. The result is that the movement of the vocal folds is not balanced and an incomplete closure of the vocal folds may appear in some or all glottal cycles. This is the reason why changes appear over the whole harmonic structure, (increasing the inter-harmonic energy and the fundamental frequency perturbation), and it also explains why energy increases at higher energy components due to aerial turbulence induced by an incomplete closure of the glottal clef.

FFT *MFCCs* [29] were considered to be appropriate for our purpose because in the presence of voice disorders they demonstrate an inherent ability to model either an irregular movement of the vocal folds, or a lack of closure induced by an increase of mass or due to a change in the properties of the tissue covering the vocal folds. The alterations related with the mucosal waveform due to an increase of mass are reflected in the low bands of the *MFCC*, whereas the higher bands are able to model the noisy components due to a lack of closure. Both alterations are reflected as noisy components with poor outstanding components and wide band spectrums. The spectral detail given by the *MFCC* can be considered good enough for our purpose.

MFCC parameters [27], [30], [31] are obtained by calculating the discrete cosine transform (*DCT*) over the logarithm of the energy in several frequency bands as shown in

$$c_m = \sum_{k=1}^{M} \log(S_k) \cdot \cos\left[m \cdot (k-0.5) \cdot \frac{\pi}{M}\right] \quad (1)$$

where $m = (1:L)$; L being the order of the *MFCC* coefficients, and S_K given by

$$S_k = \sum_{j=1}^{NFFT} W_k(j) \cdot X(j) \quad (2)$$

where $k = (1:M)$; M being the number of the mel bands in the mel scale, $W_k(j)$ is the triangular weighting function associated with the k_{th} mel band in the mel scale, and $X(j)$ is the NFFT-point magnitude spectrum ($j = 1 : \mathrm{NFTT}$). It is usual to fix $\mathrm{M} = \mathrm{round}\,(3 \cdot \ln(\mathrm{sampling\ frequency}))$.

Each band in the frequency domain is bandwidth dependant on the central frequency of the filter. The higher the frequency is, the wider is bandwidth is. This method is based on the human perception system, establishing a logarithmic relationship between the real frequency scale (Hz) and the perceptual frequency scale (mels). The suggested formula that models this relationship is as follows [27]:

$$F_{\mathrm{mel}} = 2595 \cdot \log_{10}\left(1 + \frac{F_{\mathrm{Hz}}}{700}\right). \quad (3)$$

A representation providing an improved view of the dynamic behavior of speech can be obtained by extending the analysis to include information about the frame energy (quadratic sum of the amplitude divided by window length), and temporal derivatives of the parameters among neighbor frames [26], [27]. Both first (Δ) and second derivatives ($\Delta\Delta$) have been used in the present study. To introduce temporal order into the parameter representation, let us denote the m_{th} coefficient at time t by $c_m(t)$ (4))

$$\frac{\partial c_m(t)}{\partial t} = \Delta c_m(t) \approx \mu \cdot \sum_{i=-N}^{N} i \cdot c_m(t+i) \quad (4)$$

where μ is an appropriate normalization constant and $(2N+1)$ is the number of frames over which the computation is performed.

The first and second derivatives provide information about the dynamics of the time-variation in *MFCC* parameters. *A priori*, these features have been considered significant because, due to the presence of disorders there is a lower stability in the speech signal; therefore, larger time variations of the parameters may be expected in pathological speech relative to normal speech.

For each time frame t, the result of the analysis is a vector of L cepstral coefficients, L delta cepstral coefficients, L delta-delta coefficients, the energy, one delta energy and one delta-delta energy. This vector has $3 \cdot L + 3$ components and the dimensionality of the feature space is $D = 3 \cdot L + 3$, as follows:

$$\begin{aligned} o(t) = &(E(t), c_1(t), c_2(t), \ldots, c_L(t), \\ &\Delta E(t), \Delta c_1(t), \Delta c_2(t), \ldots, \Delta c_L(t), \\ &\Delta\Delta E(t), \Delta\Delta c_1(t), \Delta\Delta c_2(t), \ldots, \Delta\Delta c_L(t)) \end{aligned} \quad (5)$$

where $o(t)$ is a feature vector with D elements.

The calculation of delta (Δ) and delta-delta ($\Delta\Delta$) was carried out by means of anti-symmetric moving-average *finite impulse response* (*FIR*) filters [Fig. 1(a) and (b)] to avoid phase distortion of the temporal sequence (length 9 for Δ, and 3 for $\Delta\Delta$).

The length of the *MFCC* vector used in the present study ranges from 10 to 26 in order to find the optimal dimensionality for the declared purposes.

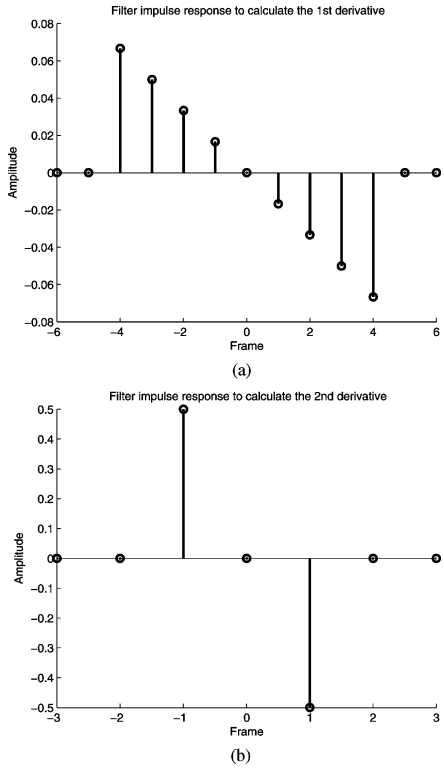

Fig. 1. Filter impulse responses used to calculate (a) the first derivative of the *MFCC* temporal sequence and (b) the second derivative of the *MFCC* temporal sequence. Both are anti-symmetric FIR filters.

B. GMM-Based Voice Pathology Detector

Let $x \in \Re^n$ be a random vector that has an arbitrary distribution. The distribution density of x is modeled as a Gaussian mixture density, a mixture of Q component densities, given by [32]

$$p\left(\frac{x}{\lambda}\right) = \sum_{i=1}^{Q} c_i \cdot p_i(x), \quad \sum_{i=1}^{Q} c_i = 1, \quad c_i \geq 0 \quad (6)$$

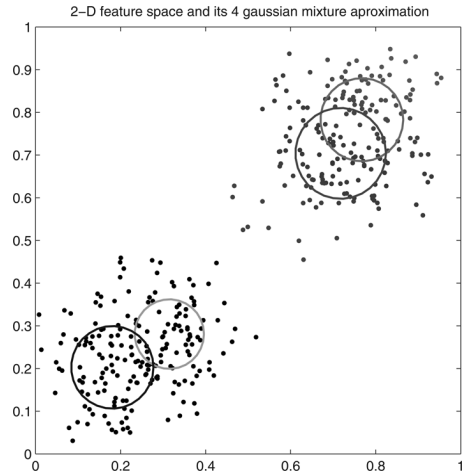

Fig. 2. Scatter plot of a two-dimensional (2-D) cepstral vector and its approximation by means of a 2-D Gaussian mixture.

where $p_i(x)$, $i = 1, \ldots, Q$ are the component densities, and c_i, $i = 1, \ldots, Q$ are the component weights. Each component density is an n-variate Gaussian function of the form

$$p_i(x) = \frac{1}{(2\pi)^{n/2}|C_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T C_i^{-1}(x-\mu_i)\right] \tag{7}$$

with μ_i the $n \times 1$ mean vector and C_i the $n \times n$ covariance matrix.

As the Gaussian components are acting together to model the overall probability density function (*pdf*), full covariance matrices are not necessary. For our purpose, the linear combination of diagonal covariance Gaussians has the capacity to model the correlation among feature elements. With this assumption, the parameters of the mixture can be represented by

$$\lambda = \{c_i, \mu_i, C_i\}, \qquad i = 1, \ldots, Q \tag{8}$$

where C_i is the covariance matrix.

The main motive for using the *GMM* as a representation of the acoustic space is that it has been demonstrated that a linear combination of Gaussian basis functions has a capacity to represent a large class of sample distributions [33], [34]. Fig. 2 shows how a Gaussian mixture is able to approximate the histogram of a given sequence.

The detector used is based on a Gaussian mixture model trained with an *expectation-maximization* (*EM*) algorithm [35]–[37]. The *EM* algorithm is an iterative method for learning maximum-likelihood parameters of a generative model where some of the random variables are observed, and some are hidden. The hidden random variables may represent quantities which we consider to be the underlying causes of the observables.

C. Dimensionality Reduction

There are a number of methods outlined in existing pattern recognition literature for reducing the dimensionality of a feature space. Several of these have been used in speech and speaker recognition with good results [24]. In this paper we study two feature selection methods: the *F-Ratio,* and *Fisher's discriminant ratio*.

1) The F*-Ratio:* The *F-Ratio* [24], [38], [39] has been widely used as the figure of merit for feature selection in speaker recognition applications. It is defined as the ratio of the between-class variance and the within-class variance. In the context of feature selection for pattern classification, this ratio selects the features which maximize the scatter between classes. The following assumptions must be enforced: a) the feature vectors within each class must have Gaussian distribution; b) features should be uncorrelated; c) the variances within each class must be equal. Since the variances within each class are generally not equal, the pooled within-class variance is used to define the *F-Ratio*. To simplify, if the number of training patterns in each of the K classes is assumed to be the same (N), the *F-Ratio* is defined as

$$F_i = \frac{B_i}{W_i} \tag{9}$$

where B_i is the between-class variance and W_i the pooled within-class variance, both given by

$$B_i = \frac{1}{K}\sum_{k=1}^{K}(\mu_{ik} - \mu_i)^2 \tag{10}$$

$$W_i = \frac{1}{K}\sum_{k=1}^{K} W_{ik} \tag{11}$$

where μ_{ik} and W_{ik} are the mean and variance of the ith feature for the kth class, and μ_i is the overall mean for the ith feature. These are given by

$$\mu_{ik} = \frac{1}{N}\sum_{k=1}^{K} x_{ikn} \tag{12}$$

$$W_{ik} = \frac{1}{N}\sum_{n=1}^{N}(x_{ikn} - \mu_{ik})^2 \tag{13}$$

$$\mu_i = \frac{1}{K}\sum_{k=1}^{K}\mu_{ik} \tag{14}$$

where x_{ikn} is the ith feature of the nth training pattern from the kth class.

2) Fisher's Discriminant Ratio: This ratio [24], [38], [39] represents the relationship between within-class and inter-class variances under the same assumptions as the *F-Ratio*. The following assumptions must be enforced: a) the feature vectors within each class must have Gaussian distribution; b) features should be uncorrelated; and, c) the variances within each class must be equal. Given a set of classes $w_k, k = 1, 2, \ldots, K$, twice scatter measurements can be defined as follows.

a) *Within-class scatter (Sw):* is a measurement of the scattering of the samples that belong to a class wk around their respective means

$$S_w = \frac{1}{K}\sum_{k=1}^{K} E\{(x-\mu_k)(x-\mu_k)^t\} \quad x \in w_k. \tag{15}$$

b) *Interclass scatter (Sb):* measures the scattering of each class mean around the overall mean

$$S_b = \frac{1}{K}\sum_{k=1}^{K}(\mu_k - \mu_0)(\mu_k - \mu_0)^t \tag{16}$$

where μ_k is the mean of class w_k and μ_0 is the mean value of the whole dataset without considering the class segmentation.

There exist several ways to quantify the discriminative power. The interclass separation measurement can be calculated comparing the relationship between within-class and inter-class scattering. The following is a standard computation:

$$J = |S_w^{-1} S_b|. \tag{17}$$

A feature vector is said to be optimum if the inter-class separation is maximized. If computing these measurements is carried out for every single feature i alone, such measurements are known as *Fisher's discriminant ratio* F_i. The higher the value of F_i, the more important the feature is: this means that feature i has a low variance with respect to inter-class variance, and this is the reason why it is desirable to discriminate between them. These criteria adopt a special form in the one-dimensional, two-class problem, quantifying the separability capabilities of individual features

$$F_i = \frac{(\mu_{ic} - \mu_{i\overline{c}})^2}{\sigma_{ic}^2 - \sigma_{i\overline{c}}^2} \tag{18}$$

where subindex C corresponds to normal voices and $\overline{C}$ to pathological.

D. Performance Measurements

The likelihood ratio test is applied to determine whether the voice is to be classified as normal or not. Both models, λ_c and $\lambda_{\overline{c}}$, (normal and pathological) are obtained. For an utterance $X = \{\vec{x}_1, \vec{x}_3, \ldots, \vec{x}_T\}$ and a model λ_c, the likelihood ratio (applying Bayes' rule and disregarding the constant prior probabilities) in the log domain is

$$\Lambda(X) = \log\left[p\left(\frac{X}{\lambda_C}\right)\right] - \log\left[p\left(\frac{X}{\lambda_{\overline{C}}}\right)\right]. \tag{19}$$

The log-likelihood ratio is compared with a threshold θ and the voice is said to be pathological if $\Lambda(X) > \theta$ and normal if $\Lambda(X) < \theta$. The decision threshold is then set to adjust the tradeoff between rejecting pathological voices (false rejection) or accepting normal voices (false acceptance). Fig. 3(a) shows the histogram of the log-likelihood ratio for the false acceptance and false rejection of a typical case. Fig. 3(b) shows false acceptance and false rejection plots versus threshold θ. Both lines cross over the *equal error rate* point (*ERR*).

The log-likelihood of the utterance given the normal or pathological model is computed assuming independence between observations by summing the probabilities obtained for each frame over a patient's utterance

$$\log\left[p\left(\frac{X}{\lambda_C}\right)\right] = \frac{1}{T}\sum_{i=1}^{T}\log\left[p\left(\frac{x_i}{\lambda_{\overline{C}}}\right)\right]. \tag{20}$$

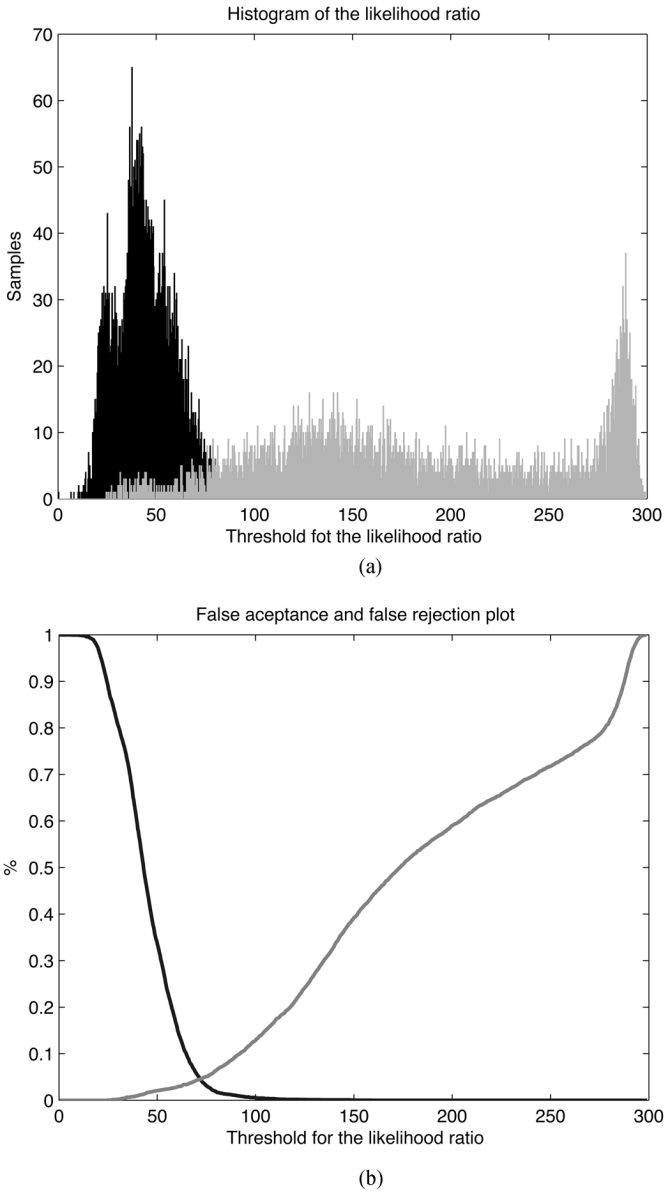

Fig. 3. (a) Histogram of the log-likelihood ratio for false acceptance and false rejection (frame accuracy). (b) Cumulative false acceptance (right) and false rejection (left).Both lines cross over the equal error rate point (ERR). The unit of analysis of the figure is frames.

The $1/\mathrm{T}$ scale is used to normalize the likelihood for utterance duration (i.e., number of frames).

A block diagram of the detector is shown in Fig. 4. For a given input test utterance X, the choice is between s (X is a normal voice record) and n (X is a pathological voice record). Possible decisions are: S, the record is classified as normal; N, the record is classified as pathological.

E. Performance Evaluation

The k-fold cross-validation scheme [40] was used for estimating the classifier performance. The variance of the performance estimates was decreased by averaging results from multiple runs of cross validation where a different random split of the training data into folds is used for each run. In this study nine repetitions were used to estimate classifier performance figures.

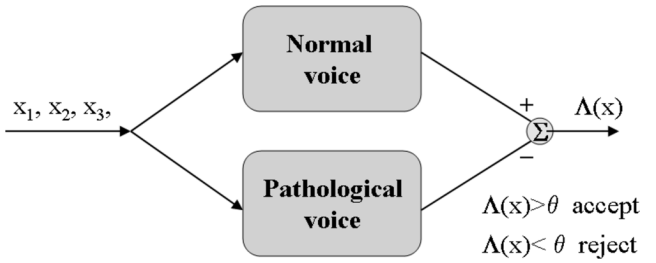

Fig. 4. Scheme of the detector. From each frame, the likelihood is calculated over both models.

TABLE II
CONFUSSION MATRIX

EVENT			
DECISION		ABSENT	PRESENT
		Efficiency (%)	
	ABSENT	TN	FN
	PRESENT	FP	TP

For each run of k-fold cross validation the total normal population and a randomly selected group of abnormals equal in size to the normal population was utilized. The performance was calculated by averaging the results obtained from each data set.

For each set, data files were split randomly into two subsets: the first for training (70%), and the second (30%) to simulate and validate results, keeping the same proportion for each class. The division into training and evaluation datasets was carried out on a file basis (not on a frame basis) in order to check and prevent the system from learning speaker-related features. Both male and female voices were mixed altogether in the training and validation sets. It is very important to use different speakers for training and validation in order to ensure that the system does not learn speaker dependent characteristics.

The training sets were used not only to extract the mixture parameters but also the value of θ. The threshold θ was calculated from the false acceptance and false rejection curves, averaging all the runs of the k-folds cross validation. According to the central limit theorem, the average of the log-likelihood ratio for false acceptance and false rejection follow a Gaussian distribution.

The number of speakers randomly selected from the subset of the database (segmented according to [14]) to build each set for cross-validation was 200 (53 normal and 147 pathological voices). This asymmetry is due to the fact that normal voice records last for approximately 3 s, whereas pathological voice records tend to be shorter as people with voice disorders have great difficulty holding a vowel over 2 or 3 s. Nine different sets were constructed for cross validation.

In order to evaluate the performance of the detector, and to enable comparisons to be made, several measurements (TP, TN, FN, and FP), ratios (SE, and SP) and curves [*relative operating characteristic (ROC)* and *Detection Error Tradeoff (DET)*] were taken into account (Table II).

1) True negative (TN): the detector found no event (normal voice) when indeed none was present.
2) True positive (TP): the detector found an event (pathological voice) when one was present
3) False rejection (FN): the classifier missed an event. Also called false negative

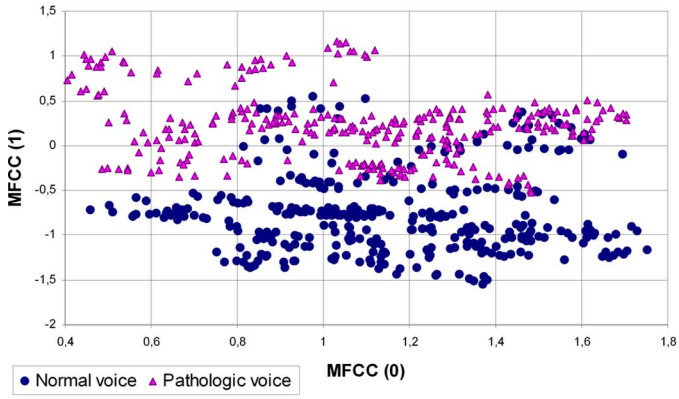

Fig. 5. The scatter plot represents the first *MFCC* versus the second *MFCC* coefficient for normal and pathological voices. (Color version available online at http://ieeexplore.ieee.org).

4) False acceptance (FP): the detector found an event when none was present. Also called false positive
5) Sensitivity (SE): likelihood that an event will be detected given that it is present

$$\mathrm{SE} = 100 \cdot \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{21}$$

6) Specificity (SP): likelihood that the absence of an event will be detected given that it is absent

$$\mathrm{SP} = 100 \cdot \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{22}$$

7) Efficiency (E): likelihood that the classification is correct

$$E\mathit{T} = 100 \cdot \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP}}. \tag{23}$$

1) Relative Operating Characteristic (ROC) or Receiver Operating Characteristic: The ROC is a popular tool in medical decision-making [41], [42]. It reveals diagnostic accuracy expressed in terms of sensitivity (or true positive rate) against (1-specificity) (or false acceptance rate) at all possible threshold values in a convenient way. The area underlying the *ROC* can be used as an estimation of the probability that the perceived abnormality detected will allow a correct identification. This index varies from 0.5 (no apparent accuracy) to 1.0 (perfect accuracy) as the *ROC* curve moves towards the left and top boundaries.

2) Detection Error Tradeoff (DET): The *DET* [43], [43] has been used widely for the assessment of detection performance in speaker identification tasks. The *DET* curve plots error rates on both axes, giving uniform treatment to both types of error, and it uses a scale for both axes which spreads out the plot and better distinguishes different well performing systems and usually produces plots that are close to linear.

IV. RESULTS

Fig. 5 represents the scatter plot of the first *MFCC* versus the second *MFCC* parameter. Although there is a clear overlapping,

TABLE III
AREA UNDER *ROC* CURVE (AUC) AND CONFIDENCE MATRIX FOR SOME OF THE TEST PERFORMED USING $MFCC + \mathrm{E} + \Delta + \Delta\Delta$ (FILE ACCURACY)

	GMM 6 mixtures				GMM 8 mixtures				GMM 12 mixtures			
	AUC	Confussion Matrix		Efficiency (%)	AUC	Confussion Matrix		Efficiency (%)	AUC	Confussion Matrix		Efficiency (%)
MFCC16	0.9976	86.4	13.5	93.2±3,8	0.9986	84.2	15.7	92.08±4.3	0.9957	84.2	15.7	92.10±4.5
		0	100			0.01	99.9			0	100	
MFCC22	0.9970	84.4	15.6	92.7±2.31	0.9970	82.8	17.2	92.1±5.33	0.9994	85.6	14.4	91.33±5.71
		0	100			0	100			0	100	
MFCC24	0.9997	88.2	11.8	94.07±3.28	0.9978	81	19	90.48±5.47	0.9885	72.5	27.5	85.96±5.19
		0	100			0	100			0.5	99.5	

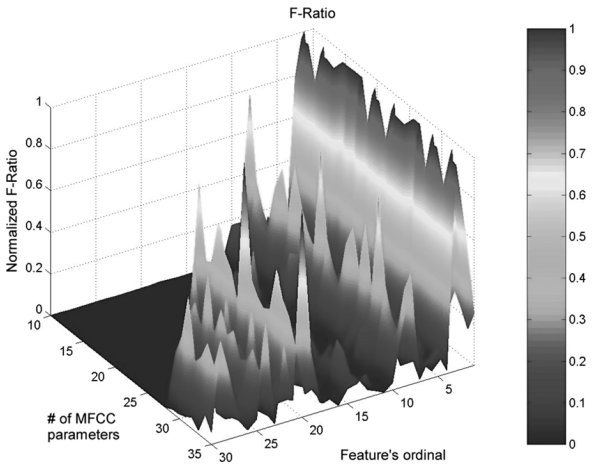

Fig. 6. F-Ratio for every test performed. F-Ratio is represented versus the number of *MFCC* parameters and the position of each feature in the feature vector, being E the first, *MFCC*(1) the second, *MFCC*(2) the third, ΔE the $(\mathrm{L}+2)$th, $\Delta MFCC(1)$ the $(\mathrm{L}+3)$th, $\Delta\Delta$E the $(2\mathrm{L}+3)$th, $\Delta\Delta MFCC(1)$ the $(2\mathrm{L}+4)$th, and so on.

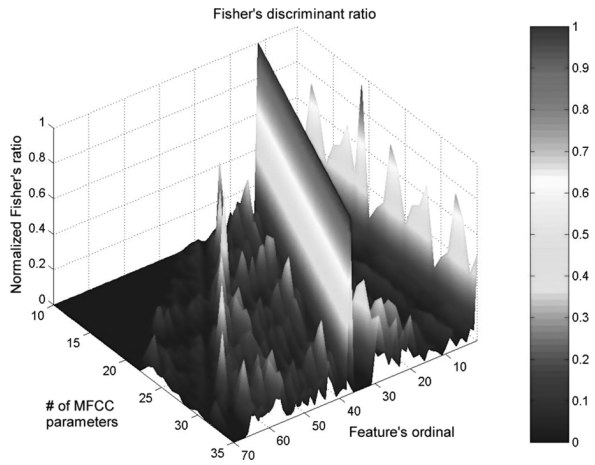

Fig. 7. Fisher's discriminant ratio for every test performed, represented versus the number of parameters and the position of each feature in the feature vector, being E the first, *MFCC*(1) the second, *MFCC*(2) the third, ΔE the $(\mathrm{L}+2)$th, $\Delta MFCC(1)$ the $(\mathrm{L}+3)$th, $\Delta\Delta$E the $(2\mathrm{L}+3)$th, $\Delta\Delta MFCC(1)$ the $(2\mathrm{L}+4)$th, and so on.

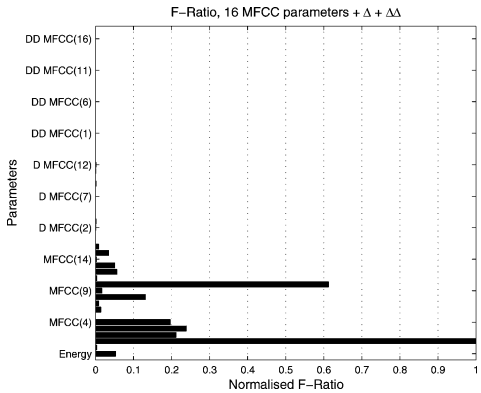

Fig. 8. Normalized F-Ratio plot using 16 $MFCC$ features $+ \Delta + \Delta\Delta$.

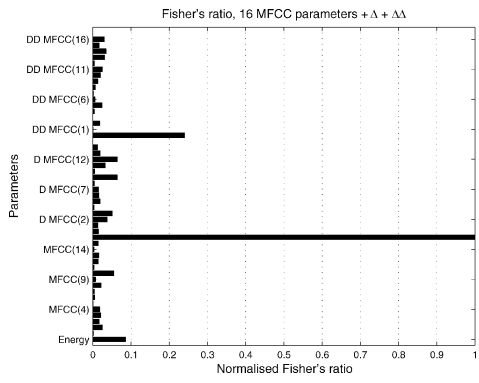

Fig. 9. Normalized Fisher's discriminant ratio plot using 16 $MFCC$ features $+ \Delta + \Delta\Delta$.

it is easy to appreciate the ability of this family of parameters to separate normal and pathological sets.

Training was carried out, with a number of *MFCC* static parameters ranging from 10 to 26. The energy (E) of the frame and dynamic features (Δ and $\Delta\Delta$) complemented every feature vector. *GMM* models were trained using 4, 6, 8, 10, 12, and 14 mixtures. The best performance was obtained using 16 *MFCC* features and a *GMM* with 6 centers. Independently of the combination of static and dynamic parameters used ($\mathrm{E}+MFCC+\Delta+\Delta\Delta$, $\mathrm{E}+MFCC+\Delta$, or $\mathrm{E}+MFCC$), the results are fairly similar. Table III represents some results for the most interesting working points: 16, 22, and 24 *MFCC* parameters, trained with a *GMM* of 6, 8, and 12 mixtures.

Fig. 6 shows the normalized F-*Ratio* for every test performed. The F-*Ratio* is represented in a three-dimensional (3-D) plot versus the number of *MFCC* parameters and each feature's position in the feature vector (its ordinal), being E

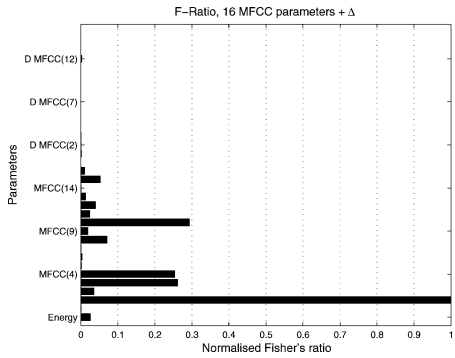

Fig. 10. Normalized F-Ratio plot using 16 $MFCC$ features + Δ.

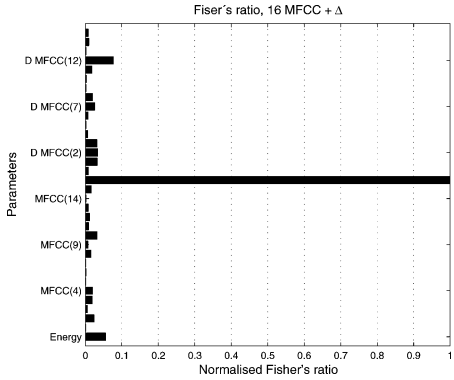

Fig. 11. Normalized Fisher's discriminant ratio plot using 16 $MFCC$ features + Δ.

the first, *MFCC*(1) the second, *MFCC*(2) the third, ΔE the $(L+2)$th, $\Delta MFCC(1)$ the $(L+3)$th, $\Delta\Delta E$ the $(2L+3)$th, $\Delta\Delta MFCC(1)$ the $(2L+4)$th, and so on. It is easy to observe that the most important features following this criterion are concentrated in the first third of the feature vector (*MFCC* static features). For simplicity reasons it was plotted only for the first 30 features. These results are shown better in Figs. 8 and 10 where a slice (16 *MFCC* parameters) of the previous plot is represented. It is easy to observe that the first and second derivatives seem to be not significant under the assumptions of this criterion.

Under the assumption of Gaussian and uncorrelated feature vectors [24], Fig. 7 shows the normalized *Fisher's discriminant ratio* for every test performed. The *Fisher's discriminant ratio* is again represented in a 3-D plot versus the number of *MFCC* parameters and each feature's position in the feature vector. It can be observed that the most important features following this criterion are concentrated in the first and second third of the feature vector. These results are shown better in Figs. 9 and 11 where a slice (16 *MFCC* parameters) of the previous plot is represented. It is easy to observe that the second derivative seems to be not significant under the assumptions of this criterion, due mainly to the similarity among normal and pathological voices; however, first derivatives remain important.

In order to show the ability of the Δ and $\Delta\Delta$ features, Figs. 12 and 13 show, respectively, five different *ROCs* and *DETs* corresponding to different detectors developed from 16 $MFCC + \mathrm{E}$. These detectors were developed by averaging the different runs of the k-folds cross validation scheme. The first detector was developed by feeding with $\mathrm{E} + MFCC$ features alone, leaving aside Δ and $\Delta\Delta$ parameters. The second detector was developed with input vectors made up of ΔE and $\Delta MFCC$ features alone. The third was formed with $\Delta\Delta E$ and $\Delta\Delta MFCC$ features alone. The fourth with $MFCC + \mathrm{E}$ and their Δs. And the fifth was constructed with $MFCC + \mathrm{E}$, their Δs, and their $\Delta\Delta s$. For this purpose a 8 mixtures *GMM* was trained for every vector combination. The *ROCs* and *DETs* show that Δ and $\Delta\Delta$ features alone have no discrimination ability at all by themselves, whereas *MFCC* features alone achieve a good performance, slightly improved when complemented with Δ but with no statistical significance. It is clear that the classification abilities of the Δ and $\Delta\Delta$ alone are quite poor compared with *MFCC*, but combined performance slightly improved the efficiency.

As a summary, Fig. 14 shows the efficiency of the test using an 8 mixture *GMM* model. Once again no relevant differences are observed when including either the first or second derivatives of the *MFCC* parameters. In view of this plot, no assumptions about the influence of Δ or $\Delta\Delta$ can be made due to the overlapped confidence intervals.

V. Discussion

In speech, as in pattern recognition, the objective is not to obtain extremely representative models, but to eliminate recognition errors. The *GMM* provides a good estimation of the probability density function that corresponds to the hyperspace formed by the n-dimensional *MFCC* vectors, minimizing the number of misclassifications in the training data. It is well established that one of the main advantages of *GMM* is its ability to classify results correctly even when classes are similar. This methodology requires a shorter time for training than other approaches such *multilayer perceptron* (*MLP*) or *learning vector quantization* (*LVQ*) [44]. Furthermore, the *GMM* approach displays comparable accuracy with respect to *LVQ*, or *MLP*.

Looking at the *F-Ratio* plots it can be observed that the most important features are the independent parameters. The differences between the most and least important features are very clear. In view of the results, independent *MFCC* parameters can be considered with a higher significance than E, Δ, and $\Delta\Delta$. The first and second derivatives seem to be lacking in interest. Such a conclusion is reflected in the efficiency of the test where it is observed that the variations in efficiency are not significant when eliminating the energy, Δ and $\Delta\Delta$ (Figs. 12–14).

Fisher's criterion partially supports these conclusions. Independent parameters and Δ have a higher relevance, whereas $\Delta\Delta$ showed a lower significance.

VI. Conclusion

The proposed scheme may be used for laryngeal pathology detection. With respect to accuracy, it can be shown that each test yielded a false acceptance ratio lower than a false rejection ratio with an efficiency around $94 \pm 3.2\%$. In order to ensure that the features in the reduced space have physical meaning,

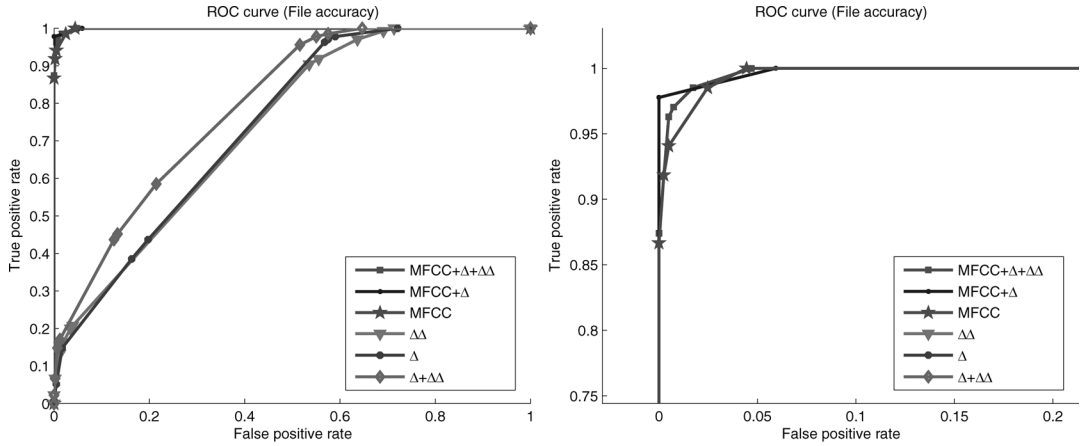

Fig. 12. (a) *ROCs* showing the true and false positive rate using 16 *MFCC* features alone ($\mathrm{AUC} = 0{,}996$), Δ alone ($\mathrm{AUC} = 0{,}727$), $\Delta\Delta$ alone ($\mathrm{AUC} = 0{,}721$), $\Delta + \Delta\Delta$ ($\mathrm{AUC} = 0{,}797$), $MFCC + \Delta$ ($\mathrm{AUC} = 0{,}998$), and $MFCC + \Delta + \Delta\Delta$ ($\mathrm{AUC} = 0{,}998$) and an 8 mixtures *GMM*. (b) Close-up of (a) plot close to $\mathrm{TP} = 1$.

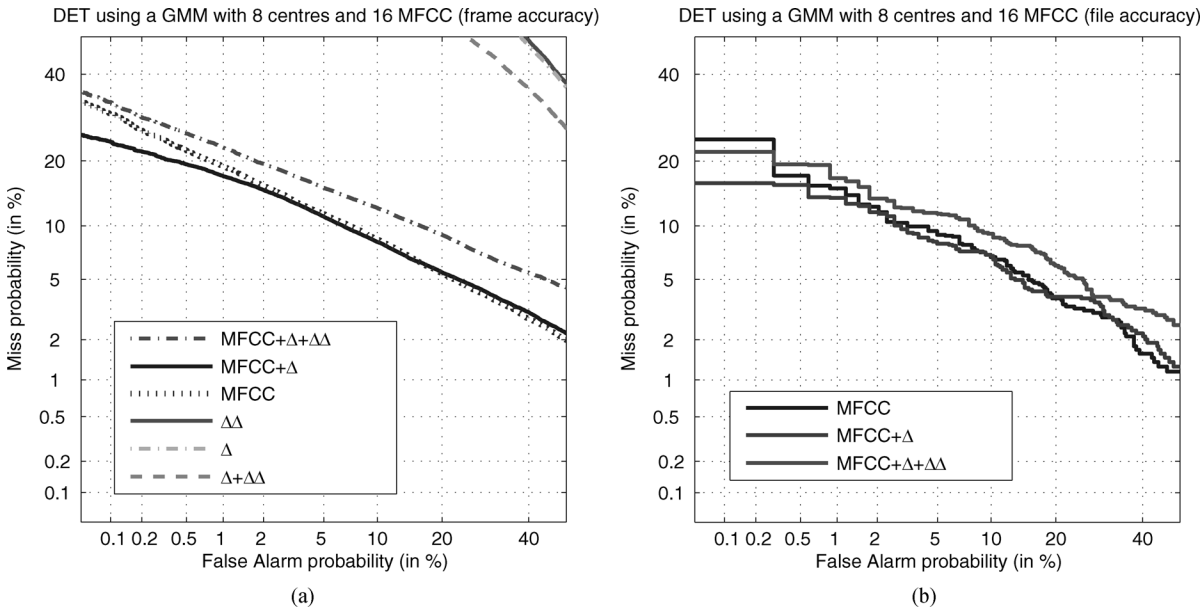

Fig. 13. *DET* plot calculated over 16 *MFCC* features alone, Δ alone, $\Delta\Delta$ alone, $MFCC + \Delta$, and $MFCC + \Delta + \Delta\Delta$ using an 8 mixtures *GMM*. The closer the plot is to the origin, the better the performance of the classifier is (a) frame accuracy and (b) file accuracy.

cepstral parameters complemented with their fist derivatives are considered the best solution for our purpose. We can conclude that the combination of the second derivatives do not show relevant influence on the results.

As in speaker and speech recognition [24], second temporal derivatives of *MFCC* parameters do not provide complementary information to the pattern recognition task of the automatic detection of voice impairments, increasing the computational complexity.

Regarding further future research, this scheme has to be tested using running speech. Preliminary studies have revealed that this approach could be used with text-dependent running speech maintaining performance. Only a small tuning would be required in the endpoint detector [28] to avoid not only silences, but also unvoiced frames.

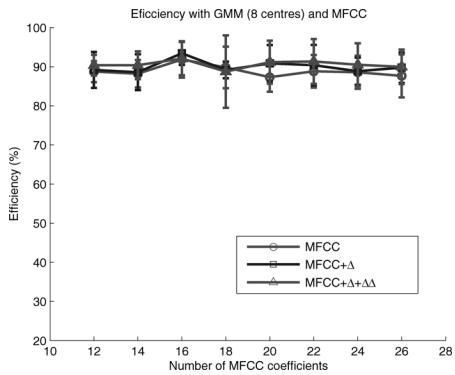

Fig. 14. Efficiency using a *GMM*-based detector with 8 mixtures and *MFCC* static parameters combined with Δ, with Δ and $\Delta\Delta$, and with energy Δ and $\Delta\Delta$.

TABLE IV
FILES EXTRACTED FROM THE DATABASE

Normal	Pathological				
1. AXH1NAL.NSP	1. ALB18AN.NSP	54. JEG29AN.NSP	107. MFC20AN.NSP	160. TAB21AN.NSP	
2. BJB1NAL.NSP	2. AMC14AN.NSP	55. JFG08AN.NSP	108. MPB23AN.NSP	161. TDH12AN.NSP	
3. BJV1NAL.NSP	3. AOS21AN.NSP	56. JFN21AN.NSP	109. MPF25AN.NSP	162. TLP13AN.NSP	
4. CAD1NAL.NSP	4. AXD19AN.NSP	57. JHW29AN.NSP	110. MPS09AN.NSP	163. TLS09AN.NSP	
5. CEB1NAL.NSP	5. AXT13AN.NSP	58. JLD24AN.NSP	111. MRB11AN.NSP	164. TPP24AN.NSP	
6. DAJ1NAL.NSP	6. BAH13AN.NSP	59. JLS11AN.NSP	112. MRC20AN.NSP	165. TPS16AN.NSP	
7. DFP1NAL.NSP	7. BEF05AN.NSP	60. JMC18AN.NSP	113. MWD28AN.NSP	166. VAW07AN.NSP	
8. DJG1NAL.NSP	8. BKB13AN.NSP	61. JPP27AN.NSP	114. MXC10AN.NSP	167. WCB24AN.NSP	
9. DMA1NAL.NSP	9. BLB03AN.NSP	62. JRF30AN.NSP	115. MXN24AN.NSP	168. WFC07AN.NS	
10. DWS1NAL.NSP	10. BPF03AN.NSP	63. JTM05AN.NSP	116. NFG08AN.NSP	169. WJB06AN.NSP	
11. EDC1NAL.NSP	11. BSG13AN.NSP	64. JXC21AN.NSP	117. NJS06AN.NSP	170. WJP20AN.NSP	
12. EJC1NAL.NSP	12. CAC10AN.NSP	65. JXD30AN.NSP	118. NKR03AN.NSP	171. WPB30AN.NSP	
13. FMB1NAL.NSP	13. CAK25AN.NSP	66. JXF11AN.NSP	119. NLC08AN.NSP	172. WST20AN.NS	
14. GPC1NAL.NSP	14. CLS31AN.NSP	67. KAB03AN.NSP	120. NMB28AN.NSP	173. WXE04AN.NSP	
15. GZZ1NAL.NSP	15. CMA06AN.NSP	68. KAC07AN.NSP	121. NMC22AN.NSP		
16. HBL1NAL.NSP	16. CMR06AN.NSP	69. KCG23AN.NSP	122. NML15AN.NSP		
17. JAF1NAL.NSP	17. CRM12AN.NSP	70. KCG25AN.NSP	123. NMV07AN.NSP		
18. JAN1NAL.NSP	18. CTB30AN.NSP	71. KDB23AN.NSP	124. OAB28AN.NSP		
19. JAP1NAL.NSP	19. DAP17AN.NSP	72. KJB19AN.NSP	125. PAT10AN.NSP		
20. JEG1NAL.NSP	20. DAS30AN.NSP	73. KLC06AN.NSP	126. PDO11AN.NSP		
21. JKR1NAL.NSP	21. DBF18AN.NSP	74. KLC09AN.NSP	127. PGB16AN.NSP		
22. JMC1NAL.NSP	22. DJP04AN.NSP	75. KLD26AN.NSP	128. PLW14AN.NSP		
23. JTH1NAL.NSP	23. DMC03AN.NSP	76. KMC22AN.NSP	129. PMC26AN.NSP		
24. JXC1NAL.NSP	24. DMP04AN.NSP	77. KMS29AN.NSP	130. PMD25AN.NSP		
25. KAN1NAL.NSP	25. DRC15AN.NSP	78. KMW05AN.NSP	131. PMF03AN.NSP		
26. LAD1NAL.NSP	26. DSC25AN.NSP	79. KPS25AN.NSP	132. RCC11AN.NSP		
27. LDP1NAL.NSP	27. DSW14AN.NSP	80. KTJ26AN.NSP	133. RHP12AN.NSP		
28. LLA1NAL.NSP	28. DVD19AN.NSP	81. KXB17AN.NSP	134. RJF22AN.NSP		
29. LMV1NAL.NSP	29. DWK04AN.NSP	82. KXH30AN.NSP	135. RJL28AN.NSP		
30. LMW1NAL.NSP	30. EAB27AN.NSP	83. LAC02AN.NSP	136. RJR15AN.NSP		
31. MAM1NAL.NSP	31. EAS11AN.NSP	84. LAD13AN.NSP	137. RJZ16AN.NSP		
32. MAS1NAL.NSP	32. EAS15AN.NSP	85. LAI04AN.NSP	138. RMB07AN.NSP		
33. MCB1NAL.NSP	33. EEC04AN.NSP	86. LAP05AN.NSP	139. RPJ15AN.NSP		
34. MFM1NAL.NSP	34. EED07AN.NSP	87. LBA15AN.NSP	140. RPQ20AN.NSP		
35. MJU1NAL.NSP	35. EJH24AN.NSP	88. LBA24AN.NSP	141. RTL17AN.NSP		
36. MXB1NAL.NSP	36. EMP27AN.NSP	89. LES15AN.NSP	142. RWC23AN.NSP		
37. MXZ1NAL.NSP	37. ESS05AN.NSP	90. LGM01AN.NSP	143. RXM15AN.NSP		
38. NJS1NAL.NSP	38. EWW05AN.NSP	91. LJH06AN.NSP	144. RXP02AN.NSP		
39. OVK1NAL.NSP	39. FMR17AN.NSP	92. LJS31AN.NSP	145. SAC10AN.NSP		
40. PBD1NAL.NSP	40. FXC12AN.NSP	93. LLM22AN.NSP	146. SAE01AN.NSP		
41. PCA1NAL.NSP	41. GDR15AN.NSP	94. LNC11AN.NSP	147. SAV18AN.NSP		
42. RHG1NAL.NSP	42. GMM09AN.NSP	95. LRD21AN.NSP	148. SBF11AN.NSP		
43. RHM1NAL.NSP	43. GMS05AN.NSP	96. LVD28AN.NSP	149. SCC15AN.NSP		
44. RJS1NAL.NSP	44. GSB11AN.NSP	97. LWR18AN.NSP	150. SEC02AN.NSP		
45. SCK1NAL.NSP	45. GXL21AN.NSP	98. LXC01AN.NSP	151. SEF10AN.NSP		
46. SCT1NAL.NSP	46. GXT10AN.NSP	99. LXC06AN.NSP	152. SEG18AN.NSP		
47. SEB1NAL.NSP	47. HJH07AN.NSP	100. LXR15AN.NSP	153. SEK06AN.NSP		
48. SIS1NAL.NSP	48. HLM24AN.NSP	101. MAB06AN.NSP	154. SHD04AN.NSP		
49. SLC1NAL.NSP	49. HXI29AN.NSP	102. MAM08AN.NSP	155. SJD28AN.NSP		
50. SXV1NAL.NSP	50. HXL58AN.NSP	103. MCA07AN.NSP	156. SLC23AN.NSP		
51. TXN1NAL.NSP	51. JAP02AN.NSP	104. MCW21AN.NSP	157. SLG05AN.NSP		
52. VMC1NAL.NSP	52. JCC10AN.NSP	105. MEC06AN.NSP	158. SMA08AN.NSP		
53. WDK1NAL.NSP	53. JCR01AN.NSP	106. MEC28AN.NSP	159. SWS04AN.NSP		

APPENDIX

Table IV shows the files included according the criteria used in [14] to segment the Kay Elemetrics Voice Disorders Database.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their assistance in the evaluation of the paper.

REFERENCES

[1] B. Boyanov and S. Hadjitodorov, "Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases," *IEEE Eng. Med. Biol. Mag.*, vol. 16, no. 4, pp. 74–82, Jul./Aug. 1997.

[2] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *J. Acoust. Soc. Am.*, vol. 80, no. 5, pp. 1329–1334, Nov. 1986.

[3] H. Kasuya, S. Ogawa, Y. Kikuchi, and S. Ebihara, "An acoustic analysis of pathological voice and its application to the evaluation of laryngeal pathology," *Speech Commun.*, vol. 5, pp. 171–181, 1986.

[4] C. Manfredi, "Adaptive noise energy estimation in pathological speech signals," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 11, pp. 1538–1543, Nov. 2000.

[5] Y. Qi and R. E. Hillman, "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," *J. Acoust. Soc. Am.*, vol. 102, no. 1, pp. 537–543, 1997.
[6] E. Yumoto, W. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc. Am.*, vol. 71, no. 6, pp. 1544–1550, 1982.
[7] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio – A new measure for describing pathological voices," *Acustica/Acta Acustica*, vol. 83, pp. 700–706, 1997.
[8] G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *J. Speech, Hearing Res.*, vol. 36, no. 2, pp. 254–266, Apr. 1993.
[9] S. Feijoo and C. Hernández, "Short-term stability measures for the evaluation of vocal quality," *J. Speech, Hearing Res.*, vol. 33, pp. 324–334, Jun. 1990.
[10] W. Winholtz, "Vocal tremor analysis with the vocal demodulator," *J. Speech, Hearing Res.*, no. 35, pp. 562–563, 1992.
[11] R. J. Baken and R. Orlikoff, *Clinical Measurement of Speech and Voice*, 2nd ed. San Diego, CA: Singular, 2000.
[12] S. Hadjitodorov, B. Boyanov, and B. Teston, "Laryngeal pathology detection by means of class-specific neural maps," *IEEE Trans. Inform. Technol. Biomed.*, vol. 4, pp. 68–73, Mar. 2000.
[13] E. Yumoto, Y. Sasaki, and H. Okamura, "Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness," *J. Speech, Hearing Res.*, vol. 27, no. 1, pp. 2–6, Mar. 1984.
[14] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *J. Speech, Language, Hearing Res.*, vol. 43, no. 2, pp. 469–485, Apr. 2000.
[15] B. Boyanov, T. Ivanov, S. Hadjitodorov, and G. Chollet, "Robust hybrid pitch detector," *Electron. Lett.*, vol. 29, no. 22, pp. 1924–1926, 1993.
[16] C. Manfredi, L. Pierazzi, and P. Bruscaglioni, "Pitch estimation for noise retrieval in time and frequency domain," *Med. Biol. Eng. Comput.*, vol. 37, no. 2, I, pp. 532–533, 1999.
[17] D. Cairns, J. H. Hansen, and J. Riski, "A noninvasive technique for detecting hypernasal speech using a nonlinear operator," *IEEE Trans. Biomed. Eng.*, vol. 43, pp. 33–45, 1996.
[18] L. Gavidia-Ceballos and J. H. L. Hansen, "Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection," *IEEE Trans. Biomed. Eng.*, vol. 43, no. 4, pp. 373–383, Apr. 1996.
[19] J. H. L. Hansen, L. Gavidia-Ceballos, and J. F. Kaiser, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 3, pp. 300–313, Mar. 1998.
[20] T. Ritchings, M. McGillion, and C. Moore, "Pathological voice quality assessment using artificial neural networks," *Med. Eng. Phys.*, vol. 24, no. 8, pp. 561–564, 2002.
[21] D. G. Childers and K. Sung-Bae, "Detection of laryngeal function using speech and electroglottographic data," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 1, pp. 19–25, Jan. 1992.
[22] M. Oliveira Rosa, J. C. Pereira, and M. Grellet, "Adaptive estimation of residue signal for voice pathology diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 1, pp. 96–104, Jan. 2000.
[23] J. I. Godino-Llorente and P. Gómez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network-based detectors," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 380–384, Feb. 2004.
[24] K. Paliwal, "Dimensionality reduction of the enhanced feature set for the HMM-based speech recognizer," *Digital Signal Process.*, vol. 2, pp. 157–173, 1992.
[25] Kay Elemetrics Corp. Disordered Voice Database1.03 ed. 1994.
[26] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, ser. Macmillan. Upper Saddle River: Prentice-Hall, 1993.
[27] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[28] J. A. Freeman and D. M. Skapura, *Neural Network, Algorithms, Applications, and Programming Techniques*. Reading, MA: Addison-Wesley, 1993.
[29] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 429–442, Jul. 2000.
[30] J. R. Deller, J. G. Proakis, and J. Hansen, *Discrete-Time Processing of Speech Signals*. New York: McMillan, 1993.
[31] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.
[32] R. J. Schalkoff, *Pattern Recognition: Statistical, Structural and Neural Approaches*. New York: Wiley, 1991.
[33] B. Fritzell, "Inverse filtering," *J. Voice*, vol. 6, no. 111, p. 114, 1992.
[34] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Acoust., Speech, Signal Process. Mag.*, vol. 29, no. 2, pp. 254–272, APR. 1981.
[35] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, Nov. 1996.
[36] D. A. Reynolds and R. C. 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.
[37] D. A. Reynolds, "Speaker identification using Gaussian mixture speaker mode," *Speech Commun.*, vol. 17, pp. 91–108, 1995.
[38] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[39] K. Fukunaga, *Statistical Pattern Recognition*, 2 ed. San Diego, CA: Academic, 1990.
[40] S. M. Weiss and C. A. Kulikowski, *Computer Systems That Learn*. San Mateo, CA: Morgan Kaufmann, 1991.
[41] J. A. Hanley and B. McNeil, "The meaning and use of the area under the receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, 1982.
[42] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristics curves derived from the same cases," *Radiology*, vol. 148, no. 3, pp. 839–843, Sep. 1983.
[43] A. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech '97*, Rhodes, Crete, 1997, vol. IV, pp. 1895–1898.
[44] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE Acoust. Speec, Signal Process Mag.*, pp. 4–21, Apr. 1987.

Juan Ignacio Godino-Llorente (M'04) was born in Madrid, Spain. He received the M.Sc. degree in communications engineering in 1996 and the Ph.D. degree in computer science from the Universidad Politécnica de Madrid, Madrid, Spain, in 2002

He is Associate Professor in the Circuits and Systems Engineering Department, at the Universidad Politécnica de Madrid. His main research is in the field of biomedical signal processing.

Pedro Gómez-Vilda (M'81) was born in Burgo de Osma, Spain. He received the M.Sc. degree in communications engineering in 1978 and the Ph.D. degree in computer science from the Universidad Politécnica de Madrid, Madrid, Spain, in 1983.

He is Professor in the Computer Science and Engineering Department, at Universidad Politécnica de Madrid since 1988. His current research interests are biomedical signal processing, speaker identification, robust speech recognition, and genetic signal processing.

Dr. Gómez Vilda is a member of the ISCA and EURASIP and the LSSA-TC Biochip Group.

Manuel Blanco-Velasco (M'05) was born in Saint Maur des Fossés, France, in 1967. He received the engineering degree from the Universidad de Alcalá, Madrid, Spain, in 1990, the M.Sc. degree in communications engineering from the Universidad Politécnica de Madrid in 1999, and the Ph.D. degree from the Universidad de Alcalá, Madrid, in 2004.

From 1992 to 2002, he has been with the Circuits and Systems Department at the Universidad Politécnica de Madrid as Assistant Professor. In April 2002, he joined the Signal Theory and Communications Department of the Universidad de Alcalá where he is now working as Associate Professor. His main research interest is biomedical signal processing.