Discriminative methods for the detection of voice disorders[1]

Juan Ignacio Godino-Llorente[1], Pedro Gómez-Vilda[1], Nicolás Sáenz-Lechón[1], Manuel Blanco-Velasco[2], Fernando Cruz-Roldán[2], and Miguel Angel Ferrer-Ballester[3]

[1] Universidad Politécnica de Madrid, EUIT de Telecomunicación, Ctra. de Valencia km. 7, 28031, Madrid, Spain
igodino@ics.upm.es
[2] Universidad de Alcalá, Escuela Politécnica, Ctra. de Madrid-Barcelona, km. 33,6, 28871, Alcalá de Henares, Madrid, Spain
[3] Universidad de Las Palmas de Gran Canaria, ETSI de Telecomunicación, Campus de Tarifa, 35017, Las Palmas de Gran Canaria, Spain

Abstract. Support Vector Machines (SVMs) have become a popular tool for discriminative classification. An exciting area of recent application of SVMs is in speech processing. In this paper discriminatively trained SVMs have been introduced as a novel approach for the automatic detection of voice impairments. SVMs have a distinctly different modelling strategy in the detection of voice impairments problem, compared to other methods found in the literature (such a Gaussian Mixture or Hidden Markov Models): the SVM models the boundary between the classes instead of modelling the probability density of each class. In this paper it is shown that the scheme proposed fed with short-term cepstral and noise parameters can be applied for the detection of voice impairments with a good performance.

1 Introduction

Voice diseases are increasing dramatically nowadays due mainly to unhealthy social habits and voice abuse. These diseases have to be diagnosed and treated at an early stage, especially larynx cancer. Acoustic analysis is a useful tool to diagnose such diseases; furthermore, it presents two main advantages: it is a non-invasive tool, and provides an objective diagnosis, being a complementary tool to those methods based on the direct observation of the vocal folds using laryngoscopy.

The state of the art in acoustic analysis allows to estimate a large amount of long-term acoustic parameters such the 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

[1] This research was carried out under grants: TIC2003-08956-C02-00 and TIC-2002-0273 from Ministry of Science and Technology of Spain; and PR2002-0239 from the Ministry of Education of Spain.

Index (SPI), Frequency Amplitude Tremor (FATR), Glottal to Noise Excitation (GNE), and many others [1-8], conceived to measure the quality and "degree of normality" of voice records. Former studies [9;10] show that the detection of voice alterations can be carried out by means of the before mentioned long-term estimated acoustic parameters, so each voice frame is quantified by a single vector. However, their reliable estimation is based on an accurate measurement of the fundamental frequency: a difficult task, especially in the presence of certain pathologies.

In the last recent years newer approaches are found using short-time analysis of the speech or electroglottographic (EGG) signal. Some of them, address the automatic detection of voice impairments from the excitation waveform collected with a laryngograph [11] or extracted from the acoustic data by inverse filtering [12]. However, due to the fact that inverse filtering is based on the assumption of a linear model, such methods do not behave well when pathology is present due to non-linearities introduced by pathology in itself.

On the other hand, it is well known that the acoustic signal itself contains information about the vocal tract and the excitation waveform as well. The basic idea for this research is to use a non-parametric approach able of modeling the effects of pathologies on both the excitation (vocal folds) and the system (vocal tract), although through the present research emphasis has been placed in pathologies affecting mainly to the vocal folds.

In this study, a novel approach to detect the presence of pathology from voice records is proposed and discussed by means of short-time parameterization of the speech signal. The automatic detection of voice alterations is addressed by means of Support Vector Machines (SVM) using non-parametric short-term Mel Frequency Cepstral Coefficients (MFCC) [13] complemented with short-term noise measurements. Each voice record is characterized with as many vectors as time frames are produced from each speech sample. The detection is carried out for each frame, and the final decision is taken establishing a threshold over the frame account classified as normal or pathological.

The present study is focused on those organic pathologies resulting in an affection of the vocal folds, which are due most of the times to vocal misuse, and reveal themselves as a modification of the excitation organ morphology (i.e. vocal folds), which may result in the increment of mass or rigidity of certain organs, thus resulting in a different pattern of vibration altering the periodicity (bimodal vibration), reducing higher modes of vibration (mucosal wave), and introducing more turbulent components in the voice record. Within this group the following pathologies can be enumerated among others: polyps, nodules, paralysis, cysts, sulcus, edemas, carcinomas, etc…

2 Methodology

Each instance in the training set contains one "target value" (class label) and several "attributes" (features). The features are calculated from short-time windows extracted form the speech utterances. The window length was selected to contain at least two consecutive pitch periods ($2 \cdot T_0$) [14]. In order to ensure a window size of at least $2 \cdot T_0$

for the lowest fundamental frequency, feature extraction was performed using a 40 ms. Hamming windows with an overlap of 50% between adjacent frames. The frame rate obtained is 50 frames/s. Fig. 1 shows a block diagram describing the process set up for the detection of voice alterations.

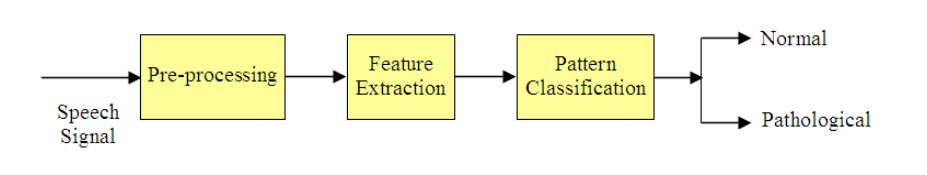


Fig. 1. Block diagram of the speech pathology detector: preprocessing front-end, feature extraction and detection module.

Results and comparisons in terms of frame accuracy are based on the calculation of the confusion matrix (as expressed in Table I). Results are calculated over the set of simulation frames. The final decision about the presence or absence of pathology is obtained by means of a threshold over the number of normal or pathological frames. The threshold was selected to be the 80% of the frame error (i.e. if the 80% of the frames are classified as normal then the register is taken as normal; otherwise is taken as pathological).

2.1 Database

Tests have been carried out using the database developed by the Massachusetts Eye and Ear Infirmary Voice and Speech Labs [15]. The speech samples were collected in a controlled environment and sampled at a 16-bit resolution. A downsampling with a previous half band filtering has been done to adjust every utterance to the sampling rate of 25 kHz. The acoustic samples are sustained phonations (1~3 s. long) of vowel /ah/ from patients (males and females) with normal voices and a wide variety of organic, neurological, traumatic, and psychogenic voice disorders.

The k-fold cross-validation scheme 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. 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 has been calculated averaging the results obtained from each data set.

For each set, data files have been 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 in a file basis (not in a frame basis) in order to check and prevent the system to learn speaker-related features. Both male and female voices have been mixed altogether in the training and validation sets.

The number of voice samples randomly selected from the database to build each set for cross-validation was 140 (53 normal and 77 pathological voices). The asymmetry is due to the fact that normal voice records are, more or less, 3 seconds long; whereas pathological voice records are shorter, because people with voice disorders have many problems to sustain a vowel during more than 2 or 3 s.

As the pre-processing front-end divides the speech signal into overlapping frames, one input vector per frame will be used to train the classifier. The total amount of vectors used to train the system is around 10.000, each corresponding to a framed window. Around 48% of them correspond to normal voices, and the remaining 52% to pathological ones.

2.2 Parameterization

Through this approach the detection of voice disorders is conducted by means of short-time features. For each frame was extracted: a) 11 MFCCs; b) 3 noise measurements: Harmonics to Noise Ratio (HNR), Normalized Noise Energy (NNE), and Glottal to Noise Excitation Ratio (GNE); c) the energy of the frame; d) and the first temporal derivatives (Δ) extracted from each enumerated parameter. The final feature vector was dimension 30 (11 MFCCs, 3 Noise features, Energy, and 15 Δ). A brief description of these parameters is given next.

Calculation of the MFCC parameters: MFCCs have been calculated following a non-parametric modeling method, which is based on the human auditory perception system. The term mel refers to a kind of estimate related to the perceived frequency. The mapping between the real frequency scale (Hz) and the perceived frequency scale (mels) is approximately linear below 1 kHz and logarithmic for higher frequencies. The bandwidth of the critical band varies accordingly to the perceived frequency [13]. Such mapping converts real into perceived frequency and matches with the idea that a well trained speech therapist is able, most of the times, to detect the presence of a disorder just listening the speech.

MFCCs can be estimated using a parametric approach derived from Linear Prediction Coefficients (LPC), or using a non-parametric FFT-based approach. However, FFT-based MFCCs typically encode more information from excitation, while LPC-based MFCCs remove the excitation. Such an idea is demonstrated in [16], 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, witch were found more sensitive to additive noise in speech recognition tasks. This is so because LPC-based MFCCs ignore the pitch-based harmonic structure seen in FFT-based MFCCs.

FFT-based MFCC parameters are obtained calculating the Discrete Cosine Transform (DCT) over the logarithm of the energy in several frequency bands as in ec. 1:

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

where $1 \le m \le L$; L being the order, and S_k given by ec. 2.

$$S_k = \sum_{j=0}^{\frac{k}{2}-1} W_k(j) \cdot X(j) \quad \textbf{(2)}$$

where $1 \le k \le M$; M being the band number in mel scale; $W_k(j)$ is the triangular weighting function associated with the *kth* mel band in mel scale.

Each band in the frequency domain is bandwidth dependant of the filter central frequency. The higher the frequency is, the wider the bandwidth is.

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.

Noise features. MFCCs have been complemented with three classical short-term measurements that were specially developed to measure the degree of noise present due to disorders. These features are: Harmonics to Noise Ratio (HNR), Normalized Noise Energy (NNE), and Glottal to Noise Excitation Ratio (GNE). The aim of these features is to separate the contribution of the excitation and the noise present, that is much higher in pathological conditions.

Harmonics to Noise Ratio (HNR). This parameter [3] is a measurement of the voice pureness. It is based on calculating the ratio of the energy of the harmonics related to the noise energy present in the voice (both measured in dB). Such measurement is carried out from the speech cepstrum, removing by liftering the energy present at the rahmonics. Fourier transformed the resulting liftered cepstrum to provide a noise spectrum which is subtracted from the original log spectrum. This results in, what is termed here, a source related spectrum. After performing a baseline correction procedure on this spectrum, the modified noise spectrum is subtracted from the original log spectrum in order to provide the HNR ratio estimate.

Normalized Noise Energy (NNE). This parameter [4] is a measurement of the noise present in the voice respect to the total energy (i.e. NNE is the ratio between the energy of noise and total energy of the signal -both measured in dB). Such measurement is carried out from the speech spectrum, separating by comb filtering the contribution of the harmonics in the frequency domain, from the valleys (noise). Between the harmonics, the noise energy is directly obtained from the spectrum. The noise energy is assumed to be the mean value of both adjacent minima in the spectrum.

Glottal to Noise Excitation Ratio (GNE). This parameter [8] is based on the correlation between Hilbert envelopes of different frequency channels extracted from the inverse filtering of the speech signal. The bandwidth of envelops is 1 kHz, and frequency bands are separated 500 Hz. Triggered by a single glottis closure, all the frequency channels are simultaneously excited, so that the envelopes in all channels share the same shape, leading to high correlation between the envelopes. The shape of each excitation pulse is practically independent of preceding or following pulses. In

case of turbulent signals (noise, whisper) a narrowband noise is excited in each frequency channel. These narrow band noises are uncorrelated (if the windows that define adjacent frequency channels do not overlap too much). The GNE is calculated picking the maximum of each correlation functions between adjacent frequency bands. The parameter indicates whether a given voice signal originates from vibrations of the vocal folds or from turbulent noise generated in the vocal tract.

Temporal derivatives. A representation better showing the dynamic behavior of speech can be obtained by extending the analysis to include the temporal derivatives of the parameters among neighbor frames. First (Δ) derivative has been used in the present study. To introduce temporal order into the parameter representation, let's denote the m_{th} coefficient at time t by $c_m(t)$ [13]:

$$\frac{\partial c_m(t)}{\partial t} = \Delta c_m(t) \approx \mu \cdot \sum_{k=-K}^{K} k \cdot c_m(t+k) \qquad (3)$$

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

For each frame t, the result of the analysis is a vector of L coefficients, to which another L-dimensional vector giving the first time derivative is appended; that is:

$$o(t) = (c_1(t), c_2(t),..., c_L(t), \Delta c_1(t), \Delta c_2(t),..., \Delta c_L(t)) \qquad (4)$$

where $o(t)$ is a feature vector with $2 \cdot L$ elements.

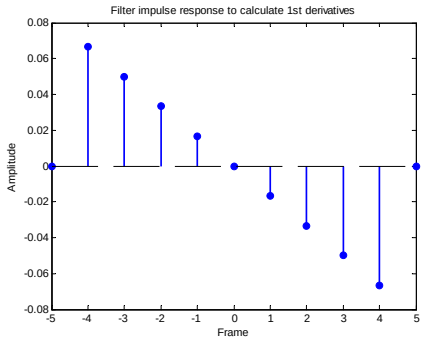


Fig. 2: Filter impulse response to calculate the Δs of the temporal sequence of parameters**.**

The Δ provides information about the dynamics of the time-variation in the parameters, providing relevant information on short-time variability. A priori, these features have been considered significant because, due to the presence of disorders a lower degree of stationarity may be expected in the speech signal [11], therefore larger temporal variations of the parameters may be expected. Another reason to complement the feature vectors with speed is that SVMs do not consider any temporal dependence by themselves as Hidden Markov Models (HMM) do. The calculation

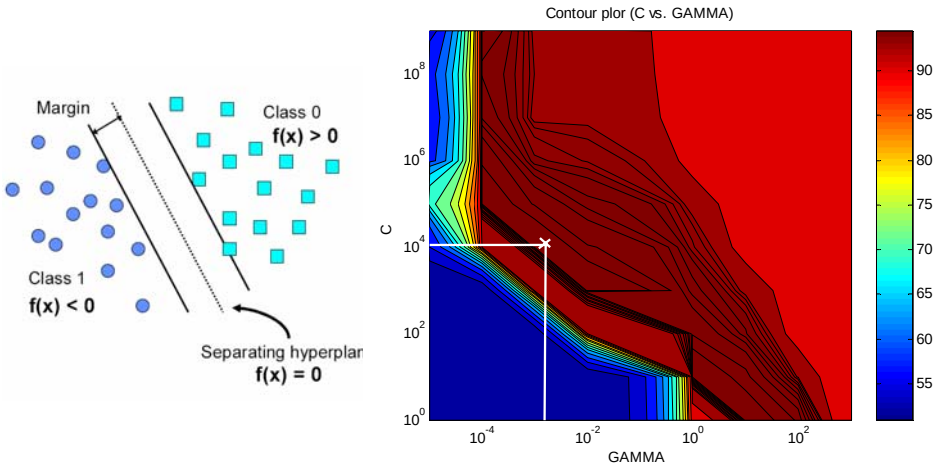
of Δ has been achieved by means of anti-symmetric Finite Impulse Response (FIR) filters to avoid phase distortion of the temporal sequence (Fig. 2).

3 An overview of the SVM detector

A support vector machine (SVM) [17] is a two-class classifier constructed from sums of a kernel function $K(\cdot,\cdot)$:

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(x, x_i) + b \tag{5}$$

where the t_i are the target values, $\sum_{i=1}^{N} \alpha_i t_i = 0$ and α_i >0. The vectors x_i are support vectors and obtained from the training set by an optimization process [17]. The target values are either 1 or -1 depending upon whether the corresponding support vector is in class 0 or class 1. For classification, a class decision is based upon whether the value, $f(x)$, is above or below a threshold.



a) b)

Fig. 3. a) Basis of the Support Vector Machine; b) Contour plot (penalty parameter C vs. γ) to show the cell where the detector performs better. The grid selected is $(C, \gamma)=(104, 10\text{-}3)$.

The kernel $K(\cdot,\cdot)$ is constrained to have certain properties (the Mercer condition), so that $K(\cdot,\cdot)$ can be expressed as:

$$K(x, y) = b(x)^t b(y) \tag{6}$$

where $b(x)$ is a mapping from the input space to a possibly infinite dimensional space. In this paper, a Radial Basis Function (RBF) kernel (ec. 7) has been used.

$$K(x, y) = e^{-\gamma \|x-y\|^2}, \quad \gamma > 0 \tag{7}$$

The optimization condition relies upon a maximum margin concept (Fig. 3a). For a separable data set, the system places a hyperplane in a high dimensional space so that the hyperplane has maximum margin. The data points from the training set lying on the boundaries are the support vectors in ec. 5. For the RBF kernel, the number of centers, the centers themselves x_i, the weights α_i, and the threshold b are all calculated automatically by the SVM training by an optimization procedure. The training imply adjusting the parameter of the kernel, γ, and a penalty parameter, C, of the error term (a larger C value corresponds to assign a higher penalty to errors). The goal is to identify good (C, γ) pairs, so that the classifier can accurately predict unknown data.

Data were normalized into the interval [-1, 1] before feeding the net. The parameters (C, γ), were chosen by cross-validation to find the optimum accuracy. At each (C, γ) grid, sequentially eight folds are used as the training set while one fold as the validation set. The grid finally selected is $(C, \gamma)=(10^4, 10^{-3})$ (Fig. 3b).

6 Results

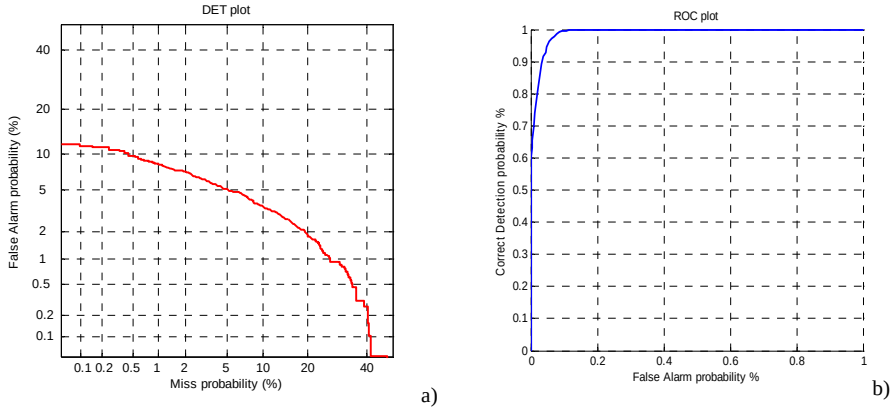
The Detection Error Tradeoff (DET) [18] and Receiver Operating Characteristic (ROC) [19] curves have been used for the assessment of detection performance (Fig. 4). ROC displays the diagnostic accuracy expressed in terms of sensitivity against (1-specificity) at all possible threshold values in a convenient way. In the DET curve we plot error rates on both axes, giving uniform treatment to both types of error, and use 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.

As shown in Fig. 4 the detector has been developed to minimize the miss probability, because in this context it is better to obtain a false alarm than missing a detection. Fig. 4a reveals that the Equal Error Rate (EER) is around 5%, however, a bias has been introduced to ensure correct detections. The results shown are significantly better than those obtained using other pattern recognition techniques such as Multilayer Perceptron (MLP) [20], and it must be remarked that the convergence rates shown by the present technique compare also significantly better against the mentioned techniques. Table 1 shows the performance of the detector in terms of frame accuracy.

The proposed detection scheme may be used for laryngeal pathology detection. In speech, as in pattern recognition, the objective is not to obtain extremely representative models, but to eliminate recognition errors. The SVM algorithm constructs a set of reference vectors that minimizes the number of misclassifications. This methodology requires a shorter time for training than other approaches such as MLP.

Table 1. Confusion matrix to show the performance of the classifier in terms of frame accuracy; a) True negative (TN): the detector found no event (normal voice) when none was present; b) True positive (TP): the detector found an event (pathological voice) when one was present; c) False negative (FN) or false rejection: the detector found no event (normal) when present (pathological); d) False positive (FP) or false acceptance: the detector found an event (pathological) when none was present (normal); e) Sensitivity: possibility for an event to be detected given that it is present; f) Specificity: possibility for the absence of an event to be detected given that it is absent.

		Event		
Decision		**Present**	**Absent**	***Efficiency (%):* 95,0±1,8**
	Present	TP: 91.01 (%)	FP: 0.77 (%)	***Sensitivity (%): 0,99***
	Absent	FN: 9.99 (%)	TN: 99.23 (%)	***Specificity (%): 0,91***



a) b)

Fig. 4. a) DET plot to show the False Alarm vs. Miss probability; b) ROC plot to show the False Alarm vs. Correct Detection probability.

References

1. Baken, R. J. and Orlikoff, R. 2000. Clinical measurement of speech and voice, 2nd ed. Singular Publishing Group.
2. Feijoo, S. and Hernández, C. 1990. Short-term stability measures for the evaluation of vocal quality. Journal of Speech and Hearing Research 33:324-334.
3. de Krom, G. 1993. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. Journal of Speech and Hearing Research 36:254-266.
4. Kasuya, H., Ogawa, S., Mashima, K., and Ebihara, S. 1986. Normalized noise energy as an acoustic measure to evaluate pathologic voice. Journal of the Acoustical Society of America 80:1329-1334.
5. Winholtz, W. 1992. Vocal tremor analysis with the vocal demodulator. Journal of Speech and Hearing Research562-563.

6. Boyanov, B. and Hadjitodorov, S. 1997. Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases. IEEE Engineering in Medicine & Biology Magazine 16:74-82.
7. Deliyski, D. Acoustic model and evaluation of pathological voice production. 1969-1972. 1993. Berlin, Germany. Proceedings of Eurospeech '93.
8. Michaelis, D., Gramss, T., and Strube, H. W. 1997. Glottal-to-Noise Excitation ratio - a new measure for describing pathological voices. Acustica/Acta acustica 83:700-706.
9. Yumoto, E., Sasaki, Y., and Okamura, H. 1984. Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness. Journal of Speech and Hearing Research 27:2-6.
10. Hadjitodorov, S., Boyanov, B., and Teston, B. 2000. Laryngeal pathology detection by means of class-specific neural maps. IEEE Transactions on Information Technology in Biomedicine 4:68-73.
11. Childers, D. G. and Sung-Bae, K. 1992. Detection of laryngeal function using speech and electroglottographic data. IEEE Transactions on Biomedical Engineering 39:19-25.
12. Gavidia-Ceballos, L. and Hansen, J. H. L. 1996. Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection. IEEE Transactions on Biomedical Engineering 43:373-383.
13. Deller, J. R., Proakis, J. G., and Hansen, J. H. L. 1993. Discrete-time processing of speech signals Macmillan Series for Prentice Hall, New York.
14. Manfredi, C., D'Aniello, M., Bruscaglioni, P., and Ismaelli, A. 2000. A comparative analysis of fundamental frequency estimation methods with application to pathological voices. Medical Engineering and Physics 22:135-147.
15. Kay Elemetrics Corp. Disordered Voice Database. Version 1.03. 1994. Lincoln Park, NJ, Kay Elemetrics Corp.
16. Bou-Ghazale, S. E. and Hansen, J. H. L. 2000. A comparative study of traditional and newly proposed features for recognition of speech under stress. IEEE Transactions on Speech and Audio Processing 8:429-442.
17. Vapnik, V. 1999. An overview of statistical learning theory. IEEE Transactions on Neural Networks 10:988-1000.
18. Martin, A., Doddington, G. R, Kamm, T., Ordowski, M., and Przybocki, M. The DET curve in assessment of detection task performance. IV, 1895-1898. 1997. Rhodes, Crete. Proceedings of Eurospeech '97.
19. Hanley, J. A. and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29-36.
20. Godino-Llorente, J. I. and Gómez-Vilda, P. 2004. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. IEEE Transactions on Biomedical Engineering 51:380-384.