Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

Optimizing laryngeal pathology detection by using combined cepstral features

Raissa Tavares (1), Nathália Monteiro (1), Suzete Correia (1), Silvana C. Costa (1), Benedito G. Aguiar Neto (2) and Joseana Macêdo Fechine (2)

(1) Federal Institute of Education, Science and Technology, João Pessoa, Brazil
(2) Federal University of Campina Grande, Campina Grande, Brazil

PACS: 43.72.Ar , 43.70.Dn .

ABSTRACT

There are several diseases that affect the human voice quality which can be organic or neurological**.** Acoustic analysis of voice features can be used as a complementary and noninvasive tool for the diagnosis of laryngeal pathologies. The degree of reliability and effectiveness of the discriminating process depends on the appropriate acoustic feature extraction. This work presents a parametric method based on cepstral features to discriminate pathological voices of speakers affected by vocal fold edema and paralysis from healthy voices. Cepstral, weighted cepstral, delta cepstral, and weighted delta cepstral coefficients are obtained from speech signals. A Vector Quantization is carried out individually for each feature in the classification process, associated with a distortion measurement. The goal is to evaluate a performance of a classifier based on the individual and combined cepstral features. The average, the product and the weighted average are the different combination strategies applied yielding a multiple classifier that is more efficient than each individual technique. To assess the accuracy of the system, 153 speech files of sustained vowel /ah/ (53 healthy, 44 vocal fold edema and 56 paralysis) of the Disordered Voice Database from Massachusetts Eye and Ear Infirmary (MEEI) are used. Results show that the employed parameters are complementary and they can be used to detect vocal disorders caused by the presence of vocal fold pathologies.

INTRODUCTION

Voice production is a complex process that involves muscle movements, respiration, and the brain control as well as hearing sensory system feedback [1]. Origins of voice disorders include structural, medical, and neurological alterations of the respiratory, laryngeal, and vocal tract mechanisms. Some pathologies are originated from maladaptive or inappropriate voice use. Other voice disorders are developed in direct response to psychogenic factors. These various physical, voice use, and psychological influences indicate that many voice disorders and laringeal pathologies are provoked from more than one origin. For example, inappropriate vocal behaviors or excessive vocal demands may incite structural changes in the vocal mucosa [2]. In the presence of vocal fold pathologies, significant changes appear in the voice caused by a modification of the excitation morphology (the distribution of mass on vocal fold and its stiffeness are increased). Pathologies of the vocal fold include those that cause any alteration in its histological structure. These are classified as organic pathologies as nodules, polyps, cysts and edemas. Voice disorders can also be caused by other pathologies which are provoked by neuro-degenerative diseases such as paralysis, Parkinson's disease and multiple sclerosis [3], [4].

Early detection of laryngeal pathologies, significantly increases the effectiveness of treatment. The diagnosis is usually made by laryngoscopical exams, which are considered invasive, causing discomfort to patients. Digital signal processing techniques, performing an acoustic analysis for vocal quality assessment are a simple and noninvasive measurement procedure. These techniques provide an objective diagnosis of pathological voices, and may be used as complementary tool in laryngoscopical exams [3]5, [4]6.

The main task of acoustic evaluation of pathological voices is related to feature extraction. Specific statistical parameters based on the linear model of speech production can be used as significant acoustic features. It is known that the voice signal is produced as a result of glottal pulses or a signal varying randomly, like noise excitation filtered by the vocal tract [4], [6], [7].

Pathology, such as Reinke's edema, polyps and paralysis affect the vocal fold or other components of the vibratory system, producing a more irregular vibration. Reikes's edema cause a excessive swelling that affects the entire length of the vocal folds and therefore the glottis closure usually is complete [2]. A vocal fold polyp interferes in glottis closure and vocal fold vibration, and depends on the type and its location. The Paralysis provokes an inadequate vocal fold closure, due to the altered resting position of the paralyzed fold. In fact, it is widely known that pathological vocal folds can present variation in the cycle of the vibratory movement because of changes in the vocal folds elasticity. For other hand, by pathology, such as vocal nodules, during vibration, the mass and stiffness of the vocal fold cover are increased, but the mechanical properties of the transition and body may not be affected [8].

Acoustic measures provide indirect observation of the voice problem and can help to identify the specific pathologies and its severity. There are a large number of acoustic measures, most of which are based on direct extraction of acoustic features.

Essentially, two parametric methods based on the linear model for the human speech production mechanism approaches have been considered on the literature so far. The first one is obtained from Linear Predictive Coding (LPC) analysis. The second parametric approach is an LPC-based cepstral analysis [9]-[12].

Cepstral analysis is applied to obtain a linear relationship between the excitation energy of the signal and the filter used. It can be very useful for the study of laryngeal disorders, as it allows processing the signal of the glottis (excitation) separately from the effects of vocal tract resonance, which facilitates the understanding of the changes that occur in the vocal folds. It is expected that any vocal disorders caused by morphological changes in vocal fold caused by a laryngeal pathology can be captured by Cepstral Coefficients [5].

In this paper we will use a parametric method based on cepstral analysis to discriminate pathological voices originating from vocal fold edema and paralysis from healthy voices. Cepstral (CEP), weighted cepstral (WCEP) delta cepstral (DCEP), and weighted delta cepstral (WDCEP) parameters are used as features to detect the irregularities of the pathological voices in comparison with the normal voice. A vector quantization technique (VQ) was associated with a distortion measurement to classify the speech signal by each parameter. The VQ was trained with voices affected by the considered pathologies individually and the results will be used to build an effective method basis for detecting Reinke's edema from normal or paralysis from normal. The vocal impairments observed for each pathology are different. While by Reinke's edema the glottis closure is usually complete it is irregular and incomplete by vocal fold paralysis.

To improve the performance of the cepstral classifiers, an approach based on multiple features classifiers is evaluated. This solution is based on the principle that by combining complementary information from distinct features classifiers a performance can be achieved which is better than that of any individual classifier. For that, three combination rules are considered: the combination by average, by product and by the weighted average, which are modifications of the strategies used in [13].

CEPSTRAL ANALYSIS OVERVIEW

The Linear Predictive Coding (LPC) estimates each speech sample based on a linear combination of the p previous samples; a larger p enables a more accurate model. It provides a set of speech parameters that represent the vocal tract [6]. It is expected that any change in the anatomical structure of the vocal tract, because of pathology, affects the LPC coefficients. Considering a healthy vocal tract, the speech disorders observed in LPC coefficients are provided by the changes in the vocal folds. A linear predictor with prediction coefficients, $\alpha(k)$, is defined as a system whose output is

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha(k) s(n-k), \qquad (1)$$

where p is the predictor order, and n-th sample of s(n).

Considering that speech signal is the result of convolving excitation with vocal tract sample response by cepstral analysis, it is possible to separate the two components. One step in cepstral deconvolution transforms a product of two spectra into a sum of two signals. In practice, the complex cepstrum is not needed. The real cepstrum is obtained by [14]:

$$c(i) = \frac{1}{N} \sum_{k=0}^{N-1} \log[X(k)] e^{j2\pi ki/N} \quad n = 0,1,....,\ N\text{-}1 \qquad (2)$$

Where $X(k)$ is equivalent to sampling the Fourier transform of $x(n)$ (windowed version of $s(n)$ at N equally spaced frequencies from w=0 to 2π and c(i) is the i-th cesptral coefficient of x(n).

Cepstral coefficients can be computed recursively from the linear predictor coefficients, $\alpha(i)$, by means of [14]:

$$\begin{cases} c(1) = -\alpha(1) \\ c(i) = -\alpha(i) - \sum_{k=1}^{i-1} (1 - \frac{k}{i}) \alpha(k) c(i-k) \quad 1 < i \le p \end{cases} \qquad (3)$$

The first derivative of the cepstral coefficients (Delta Cepstral Coefficients - DCEP) is given by [14],[15]:

$$\frac{\Delta c(n,t)}{\Delta t} = \Delta c_i(n) \approx \phi \sum_{k=-K}^{K} k c(n,\ t+k), \qquad (4)$$

where $c(n,t)$ is the *n-th* LP coefficient at time t, ϕ is a normalization constant and $2K+1$ is the number of frames over which the computation is performed.

The delta cepstral coefficients are obtained as a simplified version of (3), as it was proposed by:

$$\Delta c_i(n) = [\sum_{q=-K}^{K} k c_{i-q}(n)] G, \quad 1 \le n \le p, \qquad (5)$$

where G is a gain term (for example, 0.375), p is the number of delta cepstral coefficients, K=2, n the coefficient index and i the frame of analysis [16].

In order to account for the sensitivity of the low-order cepstral coefficients to overall spectral slope and the sensitivity of the high-order cepstral coefficients to noise, cepstral weighting (liftering) is employed.

The weighted cepstral coefficients (WCEP), $cw_i(n)$, are obtained by [14]-[17]:

$$cw_i(n) = c_i(n) \cdot w(n). \qquad (6)$$

The type of window used in this work was the band pass liftering (BPL), given by [15]:

$$w(n) = \begin{cases} 1 + \frac{L}{2} sin(\frac{n\pi}{L}), & n = 1,2,...,L \\ 0, & \text{otherwise.} \end{cases}, \qquad (7)$$

where L is the size of the window. The BPL weighs a cepstral sequence by (6) so that the lower- and higher-order components are de-emphasized.

Weighted Delta Cepstral coefficients (WDCEP) associates the characteristics of weighted cepstral and delta cepstral by [14],[15]:

$$\Delta cw_i(n) = \Delta c_i(n) \cdot w(n). \quad (8)$$

DATABASE AND METHODS

The speech signals used were extracted from the Disordered Voice Database, model 4337, recorded by the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab [18]. It includes more than 1,400 voice samples (i.e., sustained /ah/ and first 12 seconds of Rainbow Passage) from approximately 700 subjects. In this work, the analysis is applied in the sustained vowel /ah/ from 152 subjects. The selected cases are: 44 patients presenting vocal fold edema - 33 female (17 to 85 years old) and 11 male (23 to 63 years old), most of them (32) with bilateral edema; 55 cases of paralysis – 30 female (19 to 80 years old) and 25 male (15 to 77 years old) and 53 patients with normal voices which are composed of 21 male (26 to 59 years old), and 32 female (22 to 52 years old).

The discriminating process of voices, using individual features, is made in two steps: training and test/classification (Fig. 1). First, the signals are pre-processed: speech signals are multiplied by a 20 ms Hamming window with an overlap of 50% and a filter of pre-emphasis (0.95) is also used. Then each cepstral parameters is calculated after LP coefficients (p=12).

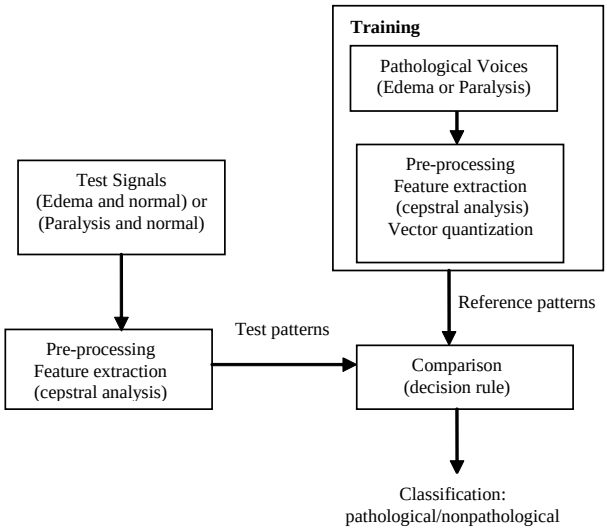

Figure 1. The discrimination process based on cepstral analysis.

To each feature, two vector quantizers are employed: one is trained by using voices affected by vocal fold edema (Edema VQ Classifier) and the other one is trained by using voices affected by vocal fold paralysis (Paralysis VQ Classifier) [20].

The VQ-classifiers are applied to static feature vectors, which are computed for every analysis frame of the speech samples over a dynamic input sustained vowel /ah/. It is used 50% of pathological voices in the training phase. After the feature extraction, a codebook is generated using the Euclidean distortion measurement and the nearest neighbour rule is used to find the codevector. LBG algorithm to quantization and the least mean square distance for classification process are used [21].

In the test/classification phase, the other 50% of the pathological (edema or paralysis) and all normal voices are pre-processed and after the feature extraction (test patterns), they are compared to the reference patterns obtained in the training phase. A distortion measurement (least square mean error) is associated to a threshold that gives the best separation between the classes (pathological/nonpathological).

After obtaining the individual feature classifier results for each case (edema or paralysis), they are combined by the three rules: average, product and weighted average. It is expected an improvement in the classification rates when comparing to individual results.

FEATURE COMBINATION

The individual classification distortion values are combined 2-by-2, 3-by-3 and 4-by-4 (Fig. 2) for each combining rule (average, product and weighted average).

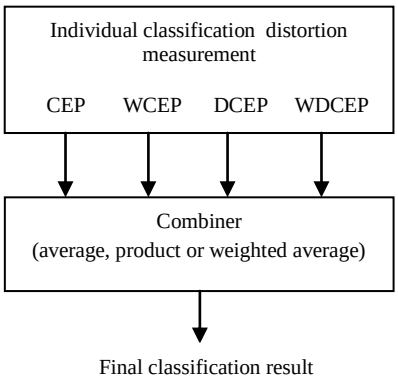

Figure 2. Overview of the Combined Feature Classifier.

To evaluate the combined features, it is necessary to make the assumption that a speech signal must be assigned to one of the M possible classes and assume that L classifiers are available. The distortion measurement used by the ith QV-classifier is denoted as d_i. Three combination rules have been employed:

- Combination by **A**verage: The value assigned to each class is the normalized distortions average of the VQ classifier outputs.

.
$$D = \frac{1}{M}\sum_{i=1}^{L} d_i, \quad (9)$$

- Combination by Product: The value assigned to each class is the normalized distortions product of the VQ classifier outputs.

$$D = \prod_{i=1}^{L} d_i, \quad (10)$$

- Combination by Weighted Average: The value assigned to each class is the weighted normalized distortions average of the VQ classifier outputs:

$$D = \frac{1}{M}\sum_{i=1}^{L} \lambda_i d_i, \quad (11)$$

where D denotes the distortion obtained after combination and λ_i are the weights for each VQ classifier distortion. For the weighted average rule, the optimum weights are obtained by an exhaustive search procedure.

RESULTS

The evaluation of performance is made by using the Efficiency rate (E) measurement, which represents the correct classification of a given class when that is present, given by (Godino-Llorente et al, 2006):

$$E(\%)= 100 .(CR+CA)/(CR+CA+FA+FR) \quad (12)$$

where:

- *CA*: Correct **a**cceptance - The presence of the pathology is detected when that is really present;
- (*CR*): Correct rejection - It is detected the correct absence of the pathology;
- (*FA*): False acceptance - It detects the presence of the pathology when it is not present; and
- (*FR*): False rejection - The presence of the pathology is rejected when, in fact, it is present.

The results are divided in two cases for analysis: Edema x Normal and Paralysis x Normal.

Edema x Normal

Table I presents the individual cepstral features evaluation performance for Edema x Normal voices. The best result is obtained to delta cepstral parameter.

The results for average, product and weighted average combinations are presented in Table II. In the average rule, an improvement of 5% related to the best individual case is obtained when CEP, WDCEP and WCEP features are combined. Efficiency about 96% is obtained when combining CEP and WCEP. However, in this case, the weighted average combination did not improve the average combination results. The best performance in discriminating normal voices form voices affected by vocal fold edema was obtained by the product rule with CEP and WDCEP combination.

Table I- Individual cepstral classifier – Edema x Normal

Feature	CR(%)	CA(%)	FA(%)	FR(%)	E(%)
CEP	89	91	11	9	90
WCEP	94	86	6	14	90
DCEP	98	86	2	14	92
WDCEP	91	82	9	18	87

Table II - Performance evaluation for the combined feature classifier (Edema x Normal)

Combined Features	Combination rules		
	A (%)	P (%)	WA (%)
CEP, DCEP	94	97	94
CEP, WDCEP	95	98	95
CEP, WCEP	96	95	96
DCEP, WDCEP	90	91	90
DCEP, WCEP	94	95	94
WDCEP, WCEP	94	95	94
CEP, DCEP, WDCEP	93	95	93
CEP, DCEP, WCEP	96	95	96
CEP, WDCEP, WCEP	97	96	97
DCEP, WDCEP, WCEP	94	94	94
CEP, DCEP, WDCEP, WCEP	96	86	96

Figure 3 shows the distortion measurement distributions for normal voices and voices affected by edema for CEP and WDCEP parameters. As the VQ was trained with edema, the distortion distribution medians are higher to normal voices. The medians differences are clear in this figure, showing the ability of parameters in separating the classes In despite of the best classification rate using the parameters individually was obtained by delta cepstral coefficients (DCEP), the best performance in the combinatios was obtained for CEP and WDCEP, by product rule. Figure 4 shows the distribution data of the distortion measurements for the product combination for CEP and WDCEP.

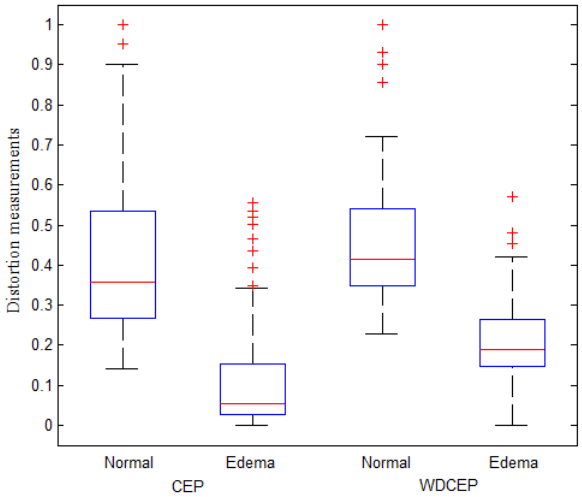

Figure 3. Distortion measurements distribution**:** normal voices and voices affected by vocal fold edema for CEP and WDCEP parameters.

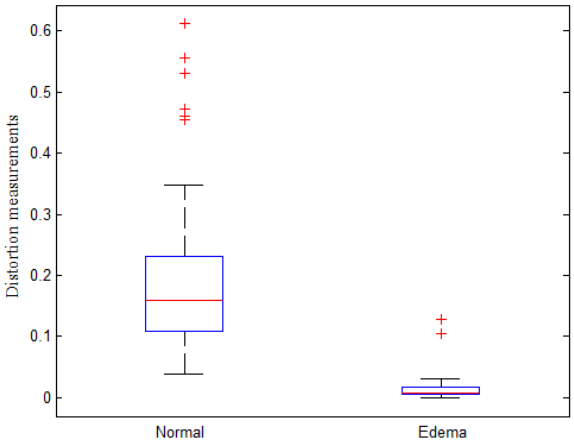

Figure 4. Distortion measurements distribution of normal voices and voices affected by vocal fold edema when combining CEP and WDCEP by product rule.

Paralysis x Normal

Individual results obtained by cepstral analysis for Paralysis VQ Classifier is presented in Table III. The cepstral coefficients (CEP) give the best classification rate (84%). In Edema's case, the best rate is given by delta cepstral coefficients (92%). It is observed that individual classifiers based on cepstral analysis presented higher efficiency rates to Edema x Normal than to the case of paralysis x normal. This suggests an assumption that LPC-based cepstral analysis should be better to track the changes in voices caused by the organic

pathology (edema) than to vocal fold pathologies caused by neurologic diseases as vocal fold paralysis.

Table III - Individual cepstral classifier – Paralysis x Normal

Feature	CR(%)	CA(%)	FA(%)	FR(%)	E(%)
CEP	92	75	8	25	84
WCEP	88	74	12	26	81
DCEP	90	63	10	37	77
WDCEP	90	60	10	40	75

The results obtained to the discrimination between voice affected by vocal fold paralysis and normal voices for average, product and weighted average combinations are presented in Table IV.

Table IV - Performance evaluation for the combined feature classifier (Paralysis x Normal)

Combined Features	Combination rules		
	A (%)	P (%)	WA (%)
CEP, DCEP	87	81	94
CEP, WDCEP	79	81	94
CEP, WCEP	83	84	96
DCEP, WDCEP	77	77	89
DCEP, WCEP	79	80	93
WDCEP, WCEP	79	79	94
CEP, DCEP, WDCEP	78	81	90
CEP, DCEP, WCEP	81	81	89
CEP, WDCEP, WCEP	81	80	91
DCEP, WDCEP, WCEP	78	79	84
CEP, DCEP, WDCEP, WCEP	80	81	93

To the average rule, it is observed an improvement of 6% (CEP and DCEP) related to the best individual case (WCEP). No improvement is given in relation to the individual cases when applying the product rule. The best result for weighted average rule is obtained to the CEP and WCEP combination, increasing 12% in efficiency rate to the best individual case.

Figure 5 shows the differences between the medians of normal voices and voices affected by paralysis.

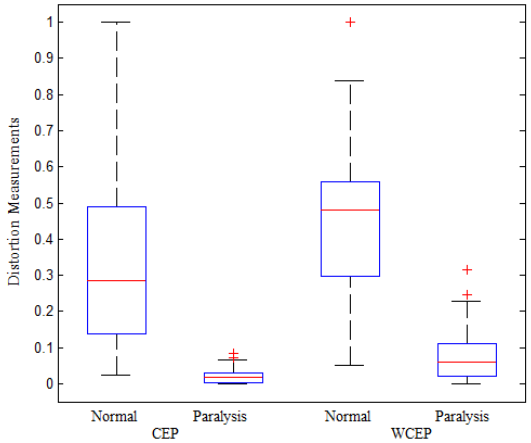

Figure 5. Distortion measurements distribution normal voices and voices affected by paralysis for CEP and WCEP parameters.

As the VQ was trained with paralysis, the distortion medians are higher to normal. It is observed in Fig. 5 that CEP is really better in separating the classes than WCEP. The medians differences are higher to CEP than to the WCEP parameter. The combination of them increases the effciciency rate and distortion distribution of the best performance (weighted average rule) is evaluated in Fig. 6.

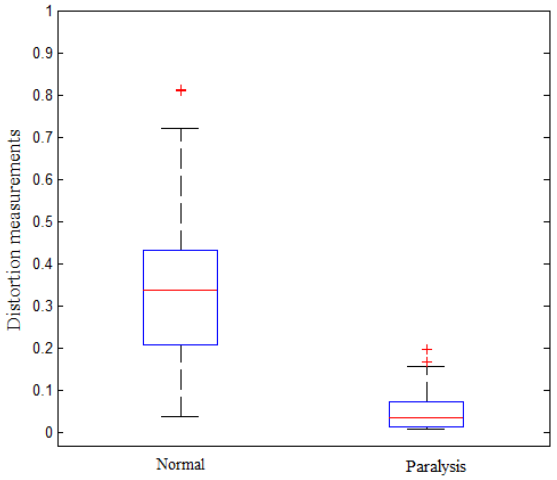

Figure 6. Distortion measurements distribution of normal voices and voices affected by paralysis when combining CEP and WCEP by weighted average rule.

CONCLUSION

In this paper a parametric method based on cepstral, weighted cepstral, delta cepstral, and weighted delta cepstral coefficients is applied to discriminate pathological voices of speakers affected by vocal fold edema and paralysis from healthy voices. The efficiency parameters were evaluated individually and combined for the pathological voice detection problem. Results show that combination of these classifiers yields a significant performance improvement related to individual ones. This means that the parameters employed are complementary and can be used to detect vocal disorders caused by the presence of vocal fold pathologies.

The combination rules presented different behaviour for each pathology considered. For edema, the product combination gives an efficiency rate of 98%, while to paralysis this rule did not improve the result of the best individual case. For paralysis, the weighted average rule is better than the other combinations, yielding 96% of efficiency. However, for edema, this rule did not have any improvement related to the average.

Future works will focus on the application of these techniques while constructing a classification system to discriminate healthy voices from pathological voices as well as among different pathologies. The method is able to discriminating the differences of produced excitations in each case. Furthermore, it is intended to use other classifiers, such as Neural Network and/or Hidden Markov Models.

ACKNOWLEDGEMENTS

The authors acknowledge the National Counsel of Technological and Scientific Development (CNPq) and the PIBITI program for the support and scholarship. We also thank the Federal Institute of Education, Science and Technology (IFPB) for financial support and Federal University of Campina Grande (UFCG) for the voice database.

REFERENCES

1. W. Chen, C. Peng, X. Zhu, B. Wan, and D. Wei, "SVM-based identification of pathological voices", *Proceedings of the 29th Annual International Conference of the IEEE EMBS*, pp. 3786-3789 (2007) ,
2. C. J. Stemple., L. Glaze and B. Klaben, *Clinical Voice Pathology, Theory and Management*, Plural Publishing, 4th Edition, (2010).
3. L. Salhi, M. Talbi, and A. Cherif, "Voice disorders identification using hybrid approach: wavelet analysis and multilayer neural networks", *World Academy of Science, Engineering and Technolog*, 45, pp. 330-339 (2008).
4. S. C Costa, B. G. Aguiar Neto, J. M. Fechine, S. Correia, "Parametric Cepstral Analysis for Pathological Voice Assessment", *Proceedings of The 23rd ACM SAC'2008*, pp. 1410-1414 (2008) .
5. J. I. Godino-Llorente, P. Gomes-Vilda and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters", *IEEE Trans. on Biom. Engineering*, Vol. 53, No. 10, pp. 1943-1953 (2006).
6. Douglas O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd Edition, NY, IEEE Press, (2000).
7. L. R. Rabiner and R. W Schafer, *Digital Processing of Speech Signals* , New Jersey: Prentice-Hall (1978).
8. L. Wallis, C. Jackson-Menaldi, W. Holland and A. Giraldo "Vocal Fold Nodule vs. Vocal Fold Polyp", Journal of Voice, *18* (1), pp. 125-129, (2004).
9. B. G. Aguiar Neto, *Signal Aufbereitung in Digitale Sprachübertragungssystemen*, Doctor-Thesis, Technische Universität Berlin, Germany (1987).
10. M. Marinaki, C. Contropoulos, I. Pitas, and N Maglaveras, "Automatic Detection of Vocal Fold. Paralysis and Edema", *Proceedings of the 8th Conf. Spoken Language Processing (Interspeech 2004)*, October, pp. 537-540 (2004).
11. 9. V. Parsa, and D. G. Jamieson, "Acoustic Discrimination of Pathological Voice: Sustained Vowels versus Continuous Speech", *Journal of Speech, Language, and Hearing Research,* Vol. 44, April, pp. 327–339 (2001).
12. M. O. Rosa, J. C. Pereira, and M. Grellet , "Adaptive estimation of Residue Signal for Voice Pathology Diagnosis", *IEEE Transactions on Biomedical Engineering*, Vol. 47, No. 1, January, pp. 96-104 (2000).
13. Gavidia-Ceballos, Liliana and Hansen, John H. L. 'Direct Speech Feature Estimation Using an Interactive EM Algorithm for Vocal Fold Pathology Detection', *IEEE Transactions on Biomedical Engineering*, Vol. 43, No. 4, April (1996).
14. J. J. de Oliveira Júnior, M. N. Kapp, C. O. A. Freitas, J. M. Carvalho, R. Sabourin. "Handwritten Recognition with Multiple Classifiers for Restricted Lexicon". *Proceedings of the 17th Brazilian Symp. on Computer Graphics and Image Processing*, vol. 1, pp. 82-89 (2004).
15. L. R. Rabiner and B. H. Huang, *Fundamentals of Speech Recognition*. Englewood Ciffs, New Jersey: Prentice Hall (1993).
16. J. R. Mammone, X., Zhang, and R. P. Ramachandran, "Robust Speaker Recognition - A Feature-Based Approach", *IEEE Signal Processing Magazine*, Vol. 13, No. 5, pp. 58-71 (1996).
17. S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Trans. on ASSP*, Vol. 29, No. 2, pp 254-272 (1981).
18. Joseana M. Fechine, *Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística*, Doctor's Thesis, Electrical Engineering, Federal University of Paraíba, Brazil (2000).
19. Kay Elemetrics Corp. Disordered Voice Database, Model 4337 (1994).
20. Makhoul, J., Roucos, S. and Gish, H. "Vector Quantization in Speech Coding", *Proceedings of the IEEE*, Vol. 73, No. 11, November, pp. 1551-1588 (1985).
21. Y Linde,. A., Buzo, and R. M Gray, "An Algorithm for Vector Quantizer Design", *IEEE Transaction on Communications*, Vol. COM-28, No. 1, January, pp 84-95 (1980).