Pathological Voice Discrimination using Cepstral Analysis, Vector Quantization and Hidden Markov Models

Silvana C. Costa, Benedito G. Aguiar Neto, *Member*, *IEEE* and Joseana M. Fechine

***Abstract*—Pathological voice discrimination has been made using digital signal processing techniques as a complementary tool to videolaringoscopy exams. This method is non-invasive to patients compared to laringoscopy. This paper aims at analyzing the use of cepstral analysis to discriminate voices affected by vocal fold pathologies. A Vector Quantizer using a distortion measurement followed by a Hidden Markov Model-based classifier is employed. Results obtained show an effective and objective way in analyzing voice disorders caused by a vocal fold pathology.**

I. Introduction

SOME vocal fold pathologies affect the vocal folds causing modifications in the voice. They could appear as a modification of the excitation morphology (the distribution of mass on vocal fold is increased). 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 [1],[2].

The evaluation of a voice quality is usually based on listening to the patient's voice or in the inspection of the vocal folds through laryngoscopy. These first techniques are both subjectives. The second one is more accurate, but is considered invasive and may cause discomfort to the patients. It also requires high cost tools.

Acoustic analysis could be employed as a useful tool in the diagnosis of diseases, as a complementary technique for the direct observation of the vocal folds. It is a non-invasive technique based on the digital processing of voice signal, and it can be used to measure the alterations in the vocal function and the evaluation of the voice. This technique aims mainly at the precocious detection of vocal folds pathologies or the evaluation of the vocal quality of patients subject to surgical or pharmacological processes in the vocal folds.

Some researchers have dedicated efforts for obtaining efficient methods for discriminating normal and pathological voices using acoustic analysis [3]-[8].

These methods have extensively used estimation of glottal noise, feature extraction from time-frequency parameters, linear prediction modeling and measures based on auditory modeling. However, the research for a more detailed and representative acoustic analysis of pathological voice signals is still a promising area.

In this work, techniques of digital signal processing are used to carry out an acoustic analysis of the pathological voice. The study is related to the case of voice disorders caused by vocal fold edemas. For this purpose, a parametric analysis based on linear prediction coding and a non parametric approach were carried out and the following parameters of voice are obtained: cepstral (CEP), delta cesptral (DCEP), weighted cepstral (WCEP), weighted delta cepstral (WDCEP) and mel-cepstral coefficients (MEL).

A vector quantization technique (VQ) associated with a distortion measurement is applied to the cepstral parameters of the speech signal. The VQ was trained with voices affected by the considered pathology and the results will be used to build an effective method basis for detecting pathological voices. Then, a left-to-right Hidden Markov Model is applied to refine the classification process. Results show an effective method in discriminating pathological voices.

II. Cepstral Analysis

A. Cepstral Coefficients

The speech signal can be considered as the result of the convolution of the excitation with vocal tract sample response of a linear model of speech production. By cepstral analysis, it is possible to separate these two components.

Pathological speech presents significant spectral differences of normal voices. The noisy characteristic of pathological voices affected by laringeal pathologies as vocal fold edema, for example, suggests the existence of relevant components in high frequencies of the spectrum [4].

Cepstral coefficients can be calculated recursively from the linear predictor (LP) coefficients, $\alpha(k)$, by means of [9]:

$$\begin{cases} c(1)=-\alpha(1) \\ c_i(n)=-\alpha(n)-\sum\limits_{j=1}^{n-1}(1-\frac{j}{n})\alpha(j)c(n-j) \qquad 1<n\leq p \end{cases} . \tag{1}$$

where p is the number of cepstral coefficients.

Cepstral coefficients obtained by (1) provide a good measure of the difference in the spectral envelope of the speech frames [10].

Manuscript received July 5, 2008. This work was supported in part by the Federal University of Campina Grande (UFCG), the Federal Center of Technological Education of Paraíba and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil.

Silvana C. Costa is currently professor at the Federal Center of Technological Education of Paraíba and a Doctor's student of Federal University of Campina Grande, Paraíba, Brazil (phone: 05583-3310-1107, e-mail: silvana@dee.ufcg.edu.br).

Benedito G. Aguiar Neto, B. G., is a full professor of the Department of Electrical Engineering, Federal University of Campina Grande, Brazil (phone: 05583-3310-1107, e-mail: bganeto@ dee.ufcg.edu.br).

Joseana M. Fechine is a professor at the Computer Science Department of Federal University of Campina Grande, Paraíba, Brazil. (phone: 05583-3310-1122, e-mail: joseana@dsc.ufcg.edu.br.

B. Delta Cepstral Coefficients (DCEP)

The first derivatives of the cepstral coefficients (Delta Cepstral Coefficients) is given by [10]:

$$\frac{\Delta c(n,t)}{\Delta t} = \Delta c_i(n) \approx \phi \sum_{k=-K}^{K} kc(n,\ t+k), \qquad (2)$$

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

These coefficients are used in order to observe the information of voice transitions in pathological speech signal versus normal speech.

In this work, the delta cepstral coefficients are obtained as a simplified version of (2), as it was proposed by [10]:

$$\Delta c_i(n) = [\sum_{q=-K}^{K} kc_{i-q}(n)]G, \quad 1 \le n \le p, \qquad (3)$$

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 [12].

C. Weighted Cepstral Coefficients (WCEP)

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 [10].

The weighted cepstral coefficients, $cw_i(n)$, are obtained by [10]-[12]:

$$cw_i(n) = c_i(n) \cdot w(n). \qquad (4)$$

The bandpass liftering (BPL) window was applied, given by [10]:

$$w(n) = \begin{cases} 1 + \frac{L}{2} sin(\frac{n\pi}{L}), & n = 1,2,...,L \\ 0, & \text{otherwise.} \end{cases}, \qquad (5)$$

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

D. Weighted Delta Cepstral Coefficients (WDCEP)

Weighted Delta Cepstral coefficients are obtained replacing (4) in (5), resulting on

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

Using (6), the characteristics of weighted cepstral and delta cepstral are associated.

E. Mel-cepstral Coefficients (MEL)

Mel-cepstral analysis is based on the human auditory perception system, which incorporates some aspects of audition. This method provides a logarithmic relationship between the real and the perceived frequency scales (mels). Mel-frequency cepstral coefficients, $c_{mel}(n)$, are calculated by means of [13]:

$$c_{mel}(n) = \sum_{k=1}^{M} \log[S(k)].\cos[n(k - \frac{1}{2}).\frac{\pi}{M}] \quad n = 0,1,....,M. \ , \qquad (7)$$

where M is the number of mel bands in the mel scale and *S*(*k*), given by

$$S(k) = \sum_{j=1}^{NFFT} W_k(j).X(j) \quad k = 1,..., M. \ , \qquad (8)$$

where $W_k(j)$ is the triangular weighting windows associated with the mel-scales, and *X*(*j*) is the NFFT-point magnitude spectrum [3].

The approximate formula to compute the mels for a given frequency *f* in Hz is given by [13]

$$f_{\text{mel}}\text{=2595.log10(1+}f\text{(Hz)/700)} . \qquad (9)$$

III. DATABASE AND METHODOLOGY

A. Database

The database used was developed by the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab [14]. The following cases were selected: 44 patients presenting vocal fold edema; 53 patients with normal voices and 23 patients affected by nodules (07), cysts (08) and paralysis (08). The speech signals are a sustained vowel /a/.

B. Methodology

Figure 1 and Figure 2 show diagram blocks of the methodology employed.

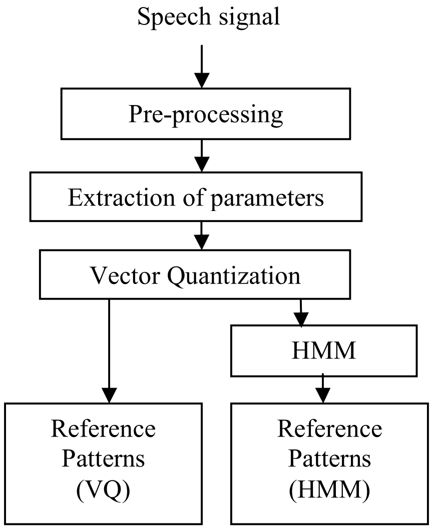

Fig. 1 – Training Phase.

In order to maintain the stationarity, the speech signals are pre-processed. The signals are multiplied by a 20 (ms) Hamming window with an overlap of 50%. A filter of preemphasis (0.95) was also used (pre-processing). Then each parameter is calculated using an LP filter of *p*=12 coefficients.

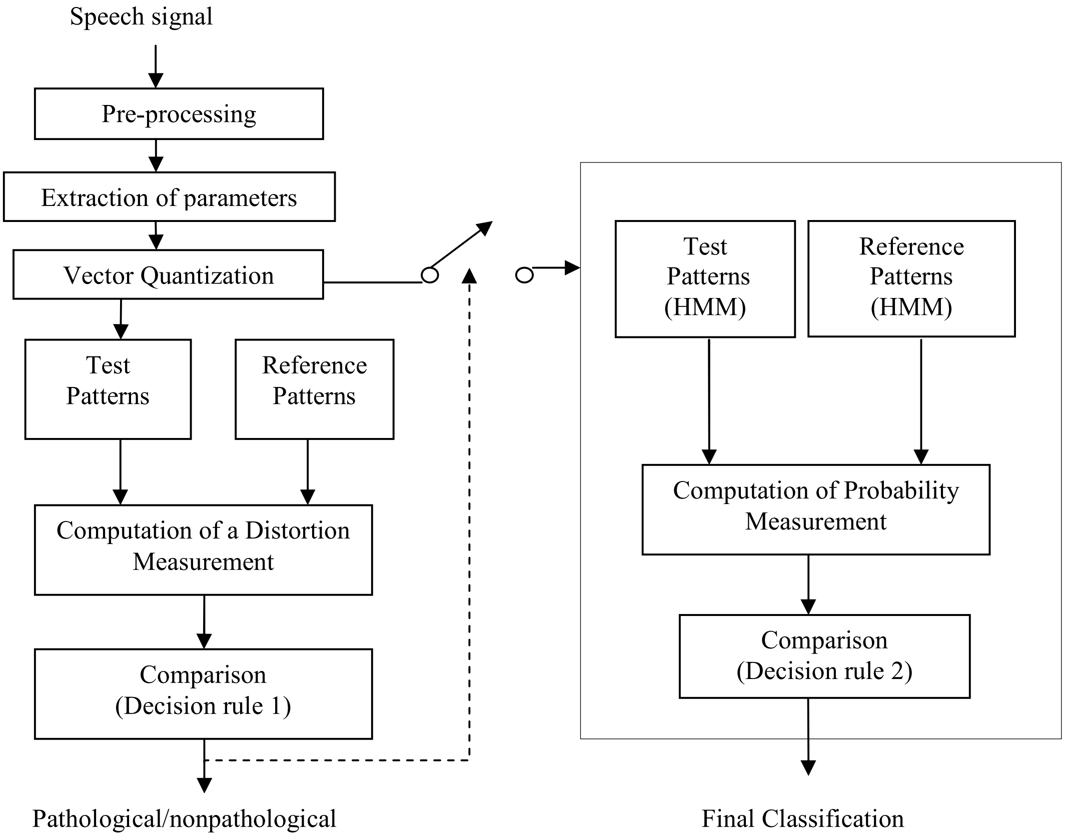

Fig. 2 – Classification Phase.

A training phase is carried out and the reference patterns are obtained for each parameter using just voices under vocal fold edema. A different classifier is used to each parameter computed as described in Section II.

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 /a/. It was used 50% of vocal fold edema cases in the training phase. To the test phase were used the other 50% of voices signals under vocal fold edema, and all the normal (53) and 23 voices under nodules (07), cysts (08) and paralysis (08). After the feature extraction, a codebook is generated using the Euclidean distortion measurement and the nearest neighbor rule was used to find the codevector. It consists of the N discrete level generation that each input vector could assume.

Thus, an N-level vector quantizer can be defined as a mapping Q of a K-dimensional Euclidean space R^K into a finite subset W of R^K. Thus,

$$Q : R^K \rightarrow W \qquad (10)$$

where the codebook $W=\{w_i ; i=1, 2, ...N\}$ is the set of codevectors, K is the dimension of the quantizer and N is the number of codevectors in W [15].

The mapping Q assigns to a K-dimensional real-valued input vector x a K-dimensional codevector $w_i=Q(x)$.

VQ defines a partitioning of the K-dimensional Euclidean space into non-intercepting cells,

$$S_i = \{x : Q(x) = w_i\}, i = 1, 2, ..., N \qquad (11)$$

As the Voronoi cell S_i collects together all input vector mapping to the i-th codevector, the codevector w_i may be viewed as a pattern-class label of the input patterns belonging to S_i.

The mapping of the input vector x to a codevector w_i occurs if

$$d(x,w_I) < d(x,w_i), \forall_i \neq I, \qquad (12)$$

where $d(.)$ is a distortion function. It follows the nearest neighbor rule to find the codevector that presents the greatest similarity to x. Here, LBG algorithm and the least mean square distance were used [16],[17].

Here, it was used N=64 and K=12.

In the classification process, a pre-classification is made using a decision rule, based on least mean square distortion between test and reference patterns. A distortion threshold is applied and a decision is taken: pathological or nonpathological. In the case of the distortion obtained to a determined signal is higher to the threshold, a re-estimation using Hidden Markov Models (HMM) (Figure 3) will be make to refine the classification process.

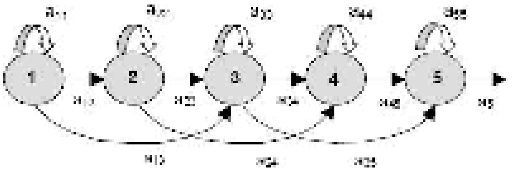

Fig. 3 – left-to-right HMM with 5 states [18].

HMM is specified by a five-tuple (S, O, π, A, B) [18].

1) $S = \{1, 2, \ldots, N\}$

Set of hidden states.

N: the number of states , s_t : the state at time t

2) $O = \{o_1, o_2, \ldots, o_M\}$

Set of observation symbols.

M: the number of observation symbols.

3) $\pi = \{\pi_i\} \qquad \pi_i = P(s_o = i), \mathbf{1} \le i \le N$

The initial state distribution.

4) $A = \{a_{ij}\} \qquad a_{ij} = P(s_t = j \mid s_{t-\mathbf{1}} = i), \mathbf{1} \le i, j \le N$

State transition probability distribution.

5) $B = \{b_j(k)\}$ e

$$b_j(k) = P(X_t = o_k \mid s_t = j), \mathbf{1} \le j \le N, \mathbf{1} \le k \le M$$

Observation symbol probability distribution in state.

Given a HMM, $\lambda = (A, B, \pi)$, and a sequence of observations, $O = \{o_1, o_2, \ldots, o_M\}$, the objective in the final classification phase is calculate the probability $P(O \mid \lambda)$ (the probability of the model that generates the observations) [18].

Discrete left-to-right HMMs of five states are used to represent each parameter. Reference patterns, $\lambda = (A, B, \pi)$, are obtained using Baum-Welch algorithm and test patterns are obtained using a probability measurement [18]. A second decision rule, using a threshold is applied and a final decision is taken.

IV. RESULTS

Three classes are considered here: 1) Edema (voices under vocal fold edema); 2) Normal (normal voices, without any pathology on vocal folds); and 3) Other Pathologies (voices affected by nodules, cysts and paralysis).

For simplicity, the tests were divided in three cases:

- *Case 1*: the system was trained with the class Edema and tested with Edema and Normal voices (Edema x Normal);
- *Case 2*: the system was trained with Edema and tested with Edema and Other Pathologies, considered as different classes (Edema x Other Pathologies);
- *Case 3*: the system was trained with Edema and tested with Edema plus Other Pathologies in the same class and Normal voices as another class ((Edema + Other Pathologies) x Normal).

To evaluate the performance of the methods, the following measurements were used [3]:

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

The Efficiency is computed to each case representing the correct classification of a given class when that is present, given by

$$E(\%) = (CR+CA)/(CR+CA+FA+FR) \text{ x } 100. \qquad (13)$$

- *Case 1 – Edema x Normal*

The efficiency of methods in discriminating pathological voices from normal voices is presented in Table I.

TABLE I
PERFORMANCE EVALUATION – EFFICIENCY TO THE CASE OF VOCAL FOLD EDEMA AND NORMAL VOICES.

Parameter	QV	HMM
Cepstral	90 %	97 %
Weighted Cepstral	90 %	99 %
Delta Cepstral	92 %	99 %
Weighted Delta Cepstral	87 %	97 %
Mel Cepstral	97 %	99 %

The superiority of MEL (E =97%) using only QV is shown in Table I. To the other parameters, however, HMM improves the results considerably. All efficiency values are higher than 95%.

- *Case 2 – Edema x Other Pathologies*

Table II shows results obtained when comparing voices under vocal fold edema and Other Pathologies (nodules, cysts and paralysis) in different classes. Cepstral and Weighted Cepstral methods have a good performance, but it is lower than case 1. The methods employed using QV were not efficient in discriminating Other Pathologies from Edema as well as discriminating Edema from Normal voices.

TABLE II
PERFORMANCE EVALUATION - VOCAL FOLD EDEMA AND OTHER PATHOLOGIES IN DIFFERENT CLASSES.

Parameter	QV	HMM
Cepstral	80 %	95 %
Weighted Cepstral	80 %	94 %
Delta Cepstral	69 %	78 %
Weighted Delta Cepstral	72 %	94 %
Mel Cepstral	61 %	73 %

When using HMM, the performance is higher then using only QV. Efficiency about 95% is obtained (Cepstral). There is a great improvement in efficiency using HMM.

The methods with HMM using Cepstral, Weighted Cepstral and Weighted Delta Cepstral coefficients have a very good performance (Effciency higher than 90%). Delta Cepstral, however, presents an efficiency lower than 80%. MEL method, however has the lower performance than all the others both in QV and HMM classification.. Perceptual aspects of mel coefficients does not give a good discrimination among perceptual aspects from the pathologies in analysis. All of them are vocal fold pathologies and their perception aspects are very similar to human perception auditory. To make a good distinction among these pathologies, a more accurate method has to be employed. As cited in [20], an accurate diagnose of nodule, polyp and laryngeal edema is very difficult, requiring some information about microscopic aspects.

- *Case 3 – (Edema + Other Pathologies) x Normal*

In Table III is presented the values of Efficiency obtained to this case, in which Edema and Other Pathologies are considered in the same class.

TABLE III
PERFORMANCE EVALUATION - VOCAL FOLD EDEMA AND OTHER PATHOLOGIES IN THE SAME CLASSES.

Parameter	QV	HMM
Cepstral	92 %	96 %
Weighted Cepstral	87 %	99 %
Delta Cepstral	89 %	99 %
Weighted Delta Cepstral	83 %	92 %
Mel Cepstral	95 %	98 %

In Table III, Mel cepstral parameter have best performance, using QV and HMM. The results in Case 3 are better than in the Case 2. However, all the methods are their best performance in discriminating Normal from Edema.

The performance using HMM in the classification process is also higher in this case to all parameters, when considering the pathologies nodules, cysts, paralysis and edema in the same class.

V. CONCLUSION

The efficiency of the cepstral analysis is shown observing results on the three cases considered here and shown in the previous section. Cepstral analysis is efficient in tracking the variability in speech given by the edema pathology. The increase of mass in vocal folds affects their vibration producing irregular patterns. Disordered voice signals were analyzed using LPC-based cepstral coefficients and its derivatives (weighted cepstral, delta cepstral, weighted cepstral and weighted delta cepstral). Mel-cepstral coefficients were also applied to discriminating pathological voices. The behavior of cepstral coefficients and their derivatives represents the result of the incomplete closure of vocal folds, because of the pathology during a sound production. Results show LPC-based Cepstral coefficients are very representative of changes in vocal tract by the edema pathology.

REFERENCES

[1] S. B. Davis, “Acoustic Characteristics of Normal and Pathological Voices”, *Speech and Language: Advances in Basic Research and Practice,* Vol. 1, pp. 271–335, 1979.

[2] F. Quek, M. Harper, Y. Haciahmetoglou, L.Chen, and L. O. Raming, “Speech pauses and gestural holds in Parkinson´s disease”, *Proc. of International Conference on Spoken Language Processing*, pp. 2485-2488, 2002.

[3] 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, October, 2006.

[4] Shama, A. Krishna, and N. U. Cholayya, “Study of Harmonics-to-Noise Ratio and Critical-Band Energy Spectrum of Speech as Acoustic Indicators of Laryngeal and Voice Pathology”, *EURASIP Journal on Advances in Signal Processing*, Vol. 2007, 2007.

[5] P. J. Murphy and Olatunji O. Akande, “Noise Estimation in Voice Signals Using Short-term Cepstral”, *J. of the Acoust. Society of America*, pp. 1679-1690, Vol. 121, No. 3, March, 2007.

[6] A. A. Dibazar, T.W. Berger, and S. S. Narayanan, “Pathological Voice Assessment”. *Proc. of the 28th IEEE EMBS Annual International Conference*, NY, USA, Aug., 2006.

[7] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, “Discrimination of Pathological Voices Using a Time-Frequency Approach”. *IEEE Trans. on Biomedical Engineering*, Vol. 52, No. 3, March, 2005.

[8] M. Bahoura and C. Pelletier, “Respiratory Sounds Classification using Analysis and Gaussian Mixture Models”, *Proceedings of the 26th Annual Conference of the IEEE EMBS*, September, 2004.

[9] T. Yoh’ichi, “A Weighted Cepstral Distance Measure for Speech Recognition”. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 35, No. 10, pp.1414-1422, October, 1987.

[10] J. R. Mammone, X., Zhang, and R. P. Ramachandran, “Robust Speaker Recognition - A Feature-Based Approach”, *IEEE Signal Processing Magazine*, Vol. 13, No. 5, pages 58-71, September, 1996.

[11] S. Furui, “Cepstral Analysis Technique for Automatic Speaker Verification”, *IEEE Trans. on ASSP*, Vol. 29, No. 2, pp 254-272, April, 1981.

[12] J. M. Fechine, “Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística”, *Doctor’s Thesis*, Electrical Engineering, Federal University of Paraíba, Brazil, 2000.

[13] Douglas O’Shaughnessy, *Speech Communications: Human and Machine,* 2nd Edition, NY, IEEE Press, 2000.

[14] Kay Elemetrics, *Kay Elemetrics Corp. Disordered Voice Database,* Model 4337, 03 Ed, 1994.

[15] J. Makhoul, S. Roucos, and H. Gish, “Vector Quantization in Speech Coding”, *Proceedings of the IEEE, Vol. 73, No. 11, November*, pp. 1551-1588, 1985.

[16] Y. Linde, A. Buzo, and R. M. Gray, “An Algorithm for Vector Quantizer Design”, *IEEE Transaction on Communications*, Vol. COM-28, No. 1, pp 84-95, January, 1980.

[17] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocky. “The DET Curve in Assessment of Detection Task Performance”. *Proceedings of Eurospeech*, Vol. 4, pages 1895-1898, 1997.

[18] L. R. Rabiner, "A Tutorial on Hidden Markov Models, and Selected Applications in Speech Recognition". *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257--286, Feb. 1989.

[19] B. M. J. Neves, J. G. Neto, P. Pontes, “Histopathological and Immunohistochemical Differentiation of Epithelial Alterations in Vocal Nodule Comparing to Polyps and to Laryngeal Edema”, *Rev. Bras. Otorrinolaringologia*, Vol.70, n.4, 439-48, jul./ago. 2004.