10th Brazilian Congress on Computational Intelligence (CBIC'2011), November 8 to 11, 2011, Fortaleza, Ceará Brazil

Automatic Perceptual Evaluation of Voice Quality According to the GRBAS using Artificial Neural Networks

Alvaro D. Orjuela[1,2] and Julián D. Arias-Londoño[1]

[1] GIBIO – Facultad de Ingeniería Electrónica y Biomédica
Universidad Antonio Nariño – Bogota D.C. - Colombia
[2] Laboratorio de Processamento de Sinais
Universidade Federal do Rio de Janeiro – RJ - Brasil
dorjuela@ieee.org, julian.arias@uan.edu.co

Abstract – In this work a comparison between two approaches for automatic classification of the GRBAS perceptual protocol of voice signals is performed. A first approach uses a classical parameterization of voice based on noise parameters and Mel frequency cepstral coefficients. In the second approach a set of parameters extracted from a nonlinear analysis of time series is used. Artificial Neural Networks have been chosen to make the classification due to the ability that they have for multi-class problems. The results show values in the fair agreement level of the Kappa index.

Keywords – Neural Networks, GRBAS, Pathological Voices, Nonlinear Analysis.

1 Introduction

In the clinical environment, an objective quantitative evaluation of voice use to be carried out by means of a combination of perceptual evaluations and acoustic parameterizations of the speech trace. The perceptual evaluation of voice consists on a subjective diagnostic technique, based on comparisons with another voice patients or with previous impressions of the same voice. The main problem is that a reliable perceptual analysis requires a standardized ability to avoid inter and intra listener differences in the evaluations [1][2]. Although, the voice assessment based on acoustic parameters has become in an increasingly technique of analysis. The perceptual evaluation is still the most practiced method for the evaluation and clinical management of voice disorders [3]. Moreover, a good enough correlation between acoustic parameters and perceptual evaluation of voices remains unfound [3][4][5].

Perceptual evaluation has been widely criticized because it is subjective. As a result, the reliability of the evaluation is not always adequate and auditory perceptual ratings can be confounded by factors such as the listener's perceptual bias, the listener's experience, the type of rating scale used, the listener's fatigue, the perceptual sensitivity of the listener to a particular voice feature and to the voice sample being evaluated [6]. This situation can be improved using an automatic system, which can provide accurate, reproducible and graded measures of a patient's voice quality, helping the speech and language therapists with the patient's treatment and rehabilitation [7]. However, few efforts have been performed in this way due to lack of standardized protocols and also low correlation with objective acoustical analysis. Currently, the most widely accepted and recommend by The Japanese Society of Logopeadics and Phoniatrics and the European Research Group evaluation protocol is the Grade, Roughness, Breathiness, Aesthenia, Strain (GRBAS) perceptual rating protocol [8]. It has been demonstrated that, on the basis of low intra-rater and inter-rater variances, the GRBAS scale seems to be the most reliable and relevant perceptual voice quality evaluation [2].

Therefore in this work a system for the automatic perceptual evaluation of voice quality is presented. The system is based on a pattern recognition approach where the set of parameters used includes noise measures, mel-cepstral coefficients and complexity measures, in order to take into account as much information as possible about the physical phenomena involved in the voice production process which could be useful for the perceptual analysis [9].

Since each parameter of the GRBAS scale can take one of four different values (classes), rating a voice according to it, can be seen as a multiclass problem, i.e. classifying each parameter of the GRBAS scale is a 4-class problem. In this sense, the classification in this work is performed using Artificial Neural Networks (ANN) which can be used for multiclass-problems in a direct way. Five different ANNs were trained (one per each GRBAS parameter) in order to provide the final decision about the quality of the voice signal. Additionally, ANNs have been previously used for the detection of pathological voices and for the automatic evaluation of voice quality with successful results [7][9]

2 GRBAS Scale and Parameterization of Voice

The GRBAS protocol comprises five qualitative scales: Grade of dysphony (G), Roughness (R), Breathiness (B), Asthenicity (A), and Strainess (S). For each one, a value in the range 0-3 is considered, where 0 corresponds to healthy voice, 1 to light disease, 2 to moderate and 3 to severe. Despite of some limitations, GRBAS is simple and fast, and has a good correlation with some acoustic parameters [9].

The severity of hoarseness is quantified under the parameter G (Grade) integrating all deviant components. Two main components of hoarseness can be identified: Breathiness (B), which is the audible impression of turbulent air leakage through

an insufficient glottal closure, and it may include short aphonic moments (unvoiced segment); and Roughness (R), which is an audible impression of irregular glottic pulses, abnormal fluctuations in f_0, separately perceived acoustic impulses (as in vocal fry), and includes diplophonia and register breaks.

As pointed out before, in this work two different characterization schemes are used. The first approach is based on acoustic parameters, while the second approach uses a nonlinear analysis. The Figure 1 shows how were implemented the schemes. Next, every approach is explained in detail.

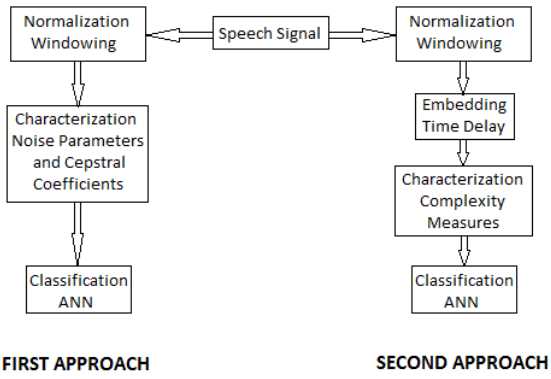

Figure 1 – Schematics of the two different systems for the automatic assessment of GRBAS protocol evaluated in this work.

First Approach

Much of studies are based on the use of the acoustic parameters: amplitude and frequency perturbation parameters, noise parameters and mel-frequency cepstral parameters (MFCC) [9]. In the present work a first approach is based on the use of noise parameters as the harmonics to noise ratio (HNR), normalized noise energy (NNE), glottal to noise excitation ratio (GNE) and mel-frequency cepstral coefficients (MFCC). This set of features has been extensively used for evaluation of voice quality. The characterization is performed using a short-time scheme where the voice signal is windowed using 40 ms Hamming windows with a 50% frame shift. Each frame is compoused by 15 features, which consists of the noise parameters (HNR, NNE, GNE) and 12 MFCC coefficients calculated based on the Fast Fourier Transform. As the study is making a voice analysis by person, the features extracted are stored by person to implement the classification.

Second Approach

A complexity analysis of biomedical signals requires a previous reconstruction of the state space of the underlying system to be characterized. Such reconstruction is carried out using a mathematical procedure called embedding, which typically is based on the time-delay embedding theorem [9]. The embedding theorem establishes that, when there is only a single sampled quantity from a dynamical system, it is possible to reconstruct a state space that is equivalent to the original (but unknown) state space composed of all the dynamical variables. The points in the state-space form trajectories, and the set of trajectories from a time series is known as attractor. From each speech frame an attractor is reconstructed and subsequently a set of 11 complexity measures are estimated.

Largest Lyapunov Exponent (LLE): LLE is a measure of the separation rate of infinitesimally close trajectories of the attractor [10]. In other words, LLE measures the sensibility to the initial conditions of the underlying system, since one of the main characteristics of nonlinear systems is the possibility that two trajectories in the state space begin very close and diverge through time, which is a consequence of the unpredictability and inherent instability of the solutions in the state space. Theoretically, a positive value of LLE means an exponential divergence of nearby trajectories and consequently a more complex dynamic behavior in the attractor.

Correlation dimension (CD): CD is a measure of the dimensionality of the space occupied by a set of random points or its geometry. Moreover, it characterizes the scaling properties of a distribution of points in an m-dimensional space (being m the dimension of the embedded attractor). The CD is the fractal dimension that has received more attention in the literature. This is mainly because its estimation is easier than others. Besides, it provides a good measure of the complexity of the dynamics, i.e. it measures the number of active degrees of freedom [11].

Approximate Entropy (AE): In the field of nonlinear dynamics, complexity measures often quantify statistically the evolution of the trajectory in the embedded phase space [12]. However, if a signal is considered as the output of a dynamical system in a specific time period, it is regarded as a source of information about the underlying dynamics; therefore, the amount of information about the state of the system that can be obtained from the signal can also be considered as a kind of complexity. The fundamental idea to measure the "amount of information" comes from the information theory, and is termed Entropy.

Entropy is a measure of the uncertainty of a random variable [9]. The most employed measure in this context is AE, which is a measure of the average conditional information generated by diverging points of the trajectory [9]. The advantage of using entropy based measures is that they measure the complexity of the signal without making assumptions about the nature of the process (deterministic or stochastic), whilst conventional nonlinear statistics such as LLE and CD assume that this nature is entirely deterministic [13], which cannot be asserted for voice signals. There are several modifications of AE published in the literature. Among them the most important is the Sample Entropy (SE), developed with the aim of obtaining a more independent measure than AE with respect to the signal length. Another two measures derived from AE are the Gaussian Kernel Approximate Entropy (GAE) and the Gaussian kernel Sample Entropy (GSE) which use a Gaussian function to measure the distance between points in the state space instead of a Heaviside function (as in the case of AE and SE) for determining the diverging points on the trajectories. In this way, nearby points have greater weight than the distant ones.

Recurrence and fractal scaling analysis: Considering that there is a combination of both deterministic and stochastic components in the voice signal during phonation [13], the deterministic component can be characterized by a measure called Recurrence period density entropy (RPDE) and the stochastic component by means of a Detrendend fluctuation analysis (DFA). RPDE quantifies any ambiguity that might exist in the fundamental frequency; the level of ambiguity is often an indicative of vocal dysfunction [13]. On the other hand, DFA characterizes the changing details of aeroacoustic breath noise in the voice and therefore it is sensitive to similar features in voice as Noise to Harmonic Ratio (NHR), but instead of NHR, DFA does not depend on a previous pitch estimation which is a difficult task for pathologic signals.

Hidden Markov entropy measurements: Most of the complexity measures used in the state of the art to characterize pathological voices, are based on multiple comparisons of the points in the attractor to establish the neighborhood of each point according to a particular distance measure. From such comparisons, the diverging points of the attractor are determined. The neighborhood of a particular vector in the state space is then understood as a region of the space in which the distance between that vector and the others is lower than a certain value (r). However, the temporal information of the points in the attractor is not taken into account. Since the points in the attractor should follow an ordered path –at least with normal stable voices–, the Hidden Markov entropy measurements were formulated to quantify the amount of information about the state of the system, taking into account the dynamic information of the points in the attractor [9]. The dynamic of the points in the attractor is modeled as a hidden Markov process (HMP) throughout a discrete hidden Markov model (DHMM), which can also be seen as an estimation of the probability density function of the process; from this model three different entropy measures are estimated: the entropy of the Markov chain (HMC), and two empirical estimations of the DHMM entropy: Shannon entropy (HES) and Renyi entropy (HER). All the complexity measures described in this section have already been used for the characterization of voice diseases and also for the automatic detection of pathological speech signals [7][8][11], showing relevant results.

The frames used to extract these parameters were of 55 ms long with an overlapping of 50%, using rectangular windows instead of more complex ones, since complexity measures lack of the spectral leakage problems presented in FFT-based parameters [9].

3 Classification using Neural Networks

ANN have demonstrated be an alternative solution in multiclass classification problem [15]. The used training can be done in a supervised way when all labels of the classes are known. In this case Multi-Layer Perceptrons (MLP) are used with feedforward connections, due to the utility that these networks have demonstrated to solve classification problems [16].

The architecture of the MLP consists of an input, an output layer and hidden layers. Most of works present only one hidden layer as the best solution to pattern classification [15]. The input is composed by the set of features used to make the classification. The output layer has one unit per class in the most of cases. This gives to the ANN versatility in the multiclass classification.

The main problem of MLP neural networks consists on finding the number of the units in the hidden layer to perform a good classification. This issue has been solve in a heuristic way, where it is developed a search altering this number in the training, and after it is chosen the number of units when the performances is the best [16].

4 Dataset and Experimental Setup

Testing was carried out using a subset of the database developed by The Massachusetts Eye and Ear Infirmary Voice & Speech Laboratory. All available 226 voices (173 pathological and 53 normal) were presented to an experienced voice therapist in a randomized order and without providing any information about the diagnosis. For each speaker, both recordings (sustained vowel and running text) were made available to him and he was asked to provide a perceptual rating for each speaker according to the GRBAS protocol. Figure 2 shows the distribution of dataset for the different scales.

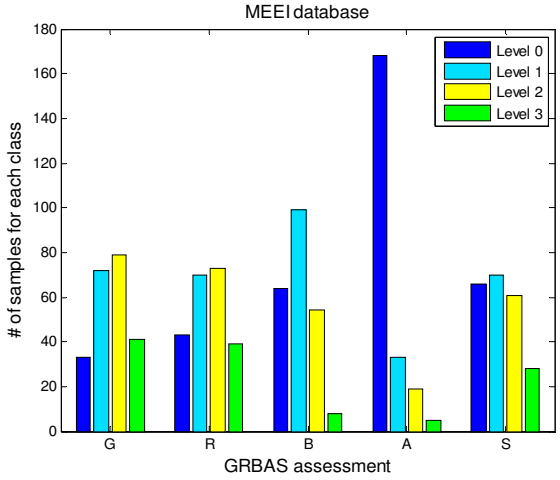

Figure 2 – Distribution of Dataset

This database was randomly divided in three sets: training with 60% of data, validation and test with 20% each one, this to implement the early stop learning technique in order to avoid loss of generalization [15], the validation set makes this task. For ANN training, it was used the Resilient Backpropagation (*Rprop*) [15] algorithm over the training set and accuracy is calculated using the test set. As was shown in the Figure 2 the number of samples per class (disease's level) has different distribution for each scale, in some cases this difference is notable between the classes.. This can be observed in the B and A scales, which have few samples for the fourth class (green bar in the Figure 2).

In order to attempt the difference between class samples in a special manner two trials were developed for each approach. Experiments were realized without and with repetition of samples in the training set, the smaller populations were repeated to obtain the same number of samples of the biggest class and so balance all classes. This guarantees that the classification do not be disturbed by the populations size. Both of cases were realized 5 times for each architecture and for each scale (G, R, B, A and S), this allow the behavior of the ANN with different initial values, and then was choose the ANN with the best performance.

As mentioned before, two approaches were performed one of them with the acoustic parameteres (noise and Mel coefficients) and other with the nonlinear analysis parameters. Both of cases the features were normalized before they were presented to the ANN, subtracting the mean and dividing by the standard deviation.

In the first approach the size of the input of ANN was composed by 15 variables: 12 mel-frequency cepstral coefficients and three noise measures. The ANN architecture is composed by one hidden layer and one output layer, the number of units in the hidden layer is found out heuristically, this is doing modifying the architecture using an even number of units from two to twenty. A higher number of units was not treated, in order to avoid a specialized network with over fitting in the training set. The units of the hidden layer all use the hyperbolic tangent as activation function and the cost function adopted was the mean square error (MSE). The output layer was implemented with four units, one for each level of the GRBAS scales. Every unit of this layer calculated the output through the logistic function, which was used as activation function. This function gives values between [0 - 1] similar to a probability function, then the unit with the higher value was the winner, associating the unit with the level of the studied parameter (the class more likely).

A second approach has as input a vector with 11 features extracted from a nonlinear analysis as mentioned previously. The procedure to implement the ANN architecture and the training was realized as in the first approach. The number of units was changed two by two units and twenty at the end. Also, 5 simulations were run to observe the initialization effect. The ANN with the best performance on the test set was chosen.

In both of cases, the accucacy was calculated with the trained ANN and information by person. This was made, taking the addition of the output for all frames extracted for one person. The output with the highest value was considered as the winning class.

5 Results

Neural networks were trained with the training set for every approach and each scales. The acoustic parameters approach have networks trained with repetition and no repetition in the training set, in each case five networks were implemented one per each scale (G, R, B, A and S). Table 1 shows the results for the accuracy for each scale. Two cases with and no repetition are presented for this approach. The number of units in the hidden layer is shown in the same table.

In the second approach based on nonlinear analysis were earned ten networks similar to the first approach.Table 2 includes the results for the second approach. The results were calculated adding the diagonal over the confusion matrix obtained by the network with the best performance. The number of units in the hidden layer for each network is included in the same table.

Table 1 – Confusion matrices of the classification parameters for the first approach

Parameter	Without Repetition		With Repetition	
Parameter	**Accuracy**	**Units in hidden layer**	**Accuracy**	**Units in hidden layer**
G	57.78 %	16	51.11 %	2
R	56.52 %	16	54.35 %	18
B	63.04 %	6	63.04 %	2
A	80.43 %	6	82.61 %	6
S	46.67 %	8	55.56 %	10

Table 2 – Confusion matrices of the classification parameters for the second approach

Parameter	Without Repetition		With Repetition	
Parameter	**Accuracy**	**Units in hidden layer**	**Accuracy**	**Units in hidden layer**
G	48.89 %	4	48.89%	4
R	47.83 %	4	54.35%	8
B	58.70 %	4	58.70%	8
A	78.26 %	8	86.96%	20
S	46.67 %	18	48.89%	2

6 Discussions

As seen in Table 1 for the first approach, the A scale has the best performance with an accuracy of 80.43%. This result is improved by the repetition in the training set because the accuracy increase becomes 82.61%. About B scale the result is less than A scale, the repetition does not cause difference on performance. In this case the accuracy is 63.04% using the network with the best performance. The G and R scales have accuracy with 56.52% and 57.78%, when the repetition was implemented the accuracy decreased. Finally, the S scale has the worse result with 46.67% of accuracy. In this case the repetition cause an increment in the result, prove of this is the 55.56% of accuracy shown in the Table 1.

The second approach (Table 2), the A scales has the highest accuracy with 78.26%. When the repetition is implemented in the training set the result reaches 86.96% of accuracy. The units in the hidden layer in both of cases are 8 and 20 respectively. B scale bears a performance of 58.70% with 4 units in the hidden layer. The repetition does not cause any effect in the result, the accuracy keeps the same. The G, R and S scales have 48.89%, 47.83% and 46.67% respectively. The repetition increase the accuracy for the scales in the R and S parameters with 54.35% and 48.89% of accuracy. The G scale is not influenced by the repetition effect. Comparing the results of two approaches, the accuracy per scale has similar results taking account an 10% interval.

The Kappa coefficient is used to measure the inter-observer variability. This is made when there are differences between studies or approaches to specific problem. The coefficient measures the proportion of agreement when a multiclass classification is implemented [17]. The Kappa coefficient can be interpreted with values in an 0 to 1 interval, when Kappa is between 0.01 to 0.20 the classification has a slight agreement. The interval 0.21 to 0.40 means that the agreement is fair. A moderate agreement is possible when Kappa value is between 0.41 to 0.60. Substantial agreement and almost perfect agreement are in the intervals from 0.61 to 0.80 and from 0.81 to 0.99.

It was calculated the Kappa coefficient for all analyzed classifications, the Table 3 illustrates these Kappa values. The two last columns have the Kappa values for results in previous studies, in order to validate the present work. First of them uses complexity measures based on nonlinear analysis, the classifier is developed using gaussian models mixture (GMM) [19]. The second study is performed with a classification aided by human expert [5].

It is possible to note that the Kappa value has been increase when the repetition in the populations of the training set was implemented for both approaches. This happens with the B, A and S scales in the classification of the first approach, the G and R scales decreased the Kappa value. In the second approach, the R, B, A and S scales increase the Kappa values and the G scale keeps the same value..

Comparing the two approaches, the Kappa coefficients are similar for the R, B and S scales. The G scale maintains its Kappa value for the second approach, but in the first approach changes in a high level when the repetition is implemented. In both approaches the A scale enhances with the repetition and states the best Kappa value of the GRBAS scales.

In a general way, the results are comparable with the work realized using GMMs [19], the B, A and S scales are better when it was used neural networks and complexity measures. Comparing the first approach with [19] the R, B and A scales have results slighty above.

About the classification aided by human expert, the first approximation is better just for B scale when no repetition is implemented. For the second approximation the A scale overcomes the human expert result when the repetition is used in the training set.

Table 3 – Kappa coefficients for the different scales

Scale	Kappa Coefficients					
	Acoustic Parameters (Without Repetition)	**Acoustic Parameters (With Repetition)**	**Nonlinear Analysis (Without Repetition)**	**Nonlinear Analysis (With Repetition)**	**Classification using GMM [19]**	**Human Expert Classification [5]**
G	0.40	0.26	0.33	0.33	0.40	0.51
R	0.43	0.41	0.32	0.40	0.40	0.46
B	0.46	0.49	0.41	0.43	0.37	0.43
A	0.39	0.59	0.24	0.64	0.32	0.41
S	0.23	0.39	0.25	0.27	0.24	0.34

7 Conclusions

When it is analyzed the results of the Kappa coefficients, this value is increased with the repetition of populations in the training sets. This is very notable for the A scale, which changed the Kappa level from fair agreement to substantial agreement in the second approach. For R, B and S scales happened the same effect, just in a minor scale.

The study has comparable results with the human expert classification. This happens for the R and B scales, which state a moderate agreement in the Kappa interpretation. The A scale can be automated due to the results are better than human expert classification using nonlinear analysis features. The G and S scales have distant results compared with the obtained values using human classification.

It is important to note that a next step involves the use of two approaches in one classifier. In that case it is necessary analyzes different strategies for the fusion of classifier due to the fact that the optimum frame size for both characterization approaches is different. Experiments with the acoustic and nonlinear analysis features as input in classifiers based on neural networks are implemented to compared with the techniques used to this moment.

Analysis over the features and study its relevance in the classification is necessary to discard the no relevant variables and in this way improve the classification.

8 Acknowledgments

This work was supported under grant: 2010238 - PI/UAN-2011-473bit from Universidad Antonio Nariño, Colombia.

9 References

[1] Velsvik, I., "Reliability in perceptual analysis of voice qualiy", *Journal of Voice,* vol. 19, no. 4, pp 555-573, 2005.
[2] Dejonckere, P., Obbens, C., de Moore, G.M., and Wienke, G., "Perceptual evaluation of disphonia: reliability and relevance", *Journal of Voice*, vol. 45, no. 2, pp 76–83, 1993.
[3] Hu, Y. and Loizou, P. C., "Evaluation of objective quality measures for speech enhancement", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp 229–238, 2008.
[4] Bhuta, T., Patrick, L. and Garnett, J.D., "Perceptual evaluation of voice Quality and its correlation with acoustic measurements", *Journal of Voice*, vol. 18, no. 3, pp. 299-304, 2004.
[5] Dejonckere, P.H., Remacle, M., Fresnel-Elbaz, E., Wolsard, V., Crevier-Buchman, L.and Millet, B., "Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements", *Rev. Laryngol. Otol. Rhinol,* vol. 117, no. 3, pp 219-224, 1996.
[6] Oates, J., "Auditory-perceptual evaluation of disordered voice quality". Folia Phoniatrica et Logopaedica, vol. 61, no. 1, pp 49–56, 2009.
[7] Ritchings, R., McGillion, M., and Moore, C. "Pathological voice quality assessment using artificial neaural networks", Medical Engineering & Physics, vol. 24, no. 8, pp 561–564, 2002.
[8] Hirano, M., *Clinical Examination of Voice*. Springer-Verlag, New York, USA, 1981.
[9] Arias-Londoño, J.D., Godino-Llorente, J.I., Sáenz-Lechón, N, Osma-Ruiz, V., Castellanos-Domínguez, G, "Automatic detection of pathological voices using complexity measures. noise parameters and mel-cepstral coefficients", IEEE Transactions on Biomedical Engineering, vol. 58, no. 2, pp 370-379, 2011.
[10] M. Costa, A. Goldberger, and C. Peng, "Multiscale entropy analysis of biological signals" Phys. Rev. E, vol 71, pp. 021906-1-021906-18, 2005
[11] Kantz, H. and Schreiber, T., *Nonlinear time series analysis,* 2nd ed. Cambridge, U.K. Cambridge University Press, 2004
[12] B.S. Aghazadeh, H. Khadivi, andM. Nikkhab-Bahbrami.,"Nonlinear analysis and classification of vocal disorders." In *Proc 29th Int. IEEE EMBS Conf.,* 2007, pp. 6199-6202.
[13] Little, M.A., McSharry, P. E., Roberts, S. J., Costello, D., and Moroz, I. M. "Exploring nonlinear recurrence and fractal scaling properties for voice disorder detection". *Biomedical Engineering Online,* vol. 6, no. 23, 2007
[14] J.J: Jiang, Y. Zhanbg, and C. McGilligan, "Chaos in voice, from modeling to measurement", Journal of Voice, vol. 20, no. 1, pp 2-17, 2006
[15] Haykin, S. (1998) 'Neural Networks: A Comprehensive Foundation', Prentice Hall.
[16] Lawrence, S., Giles, C.L., "Overfitting and neural networks: conjugate gradient and backpropagation", *Proceedings of the IEEE-INNS-ENNS International Joint Conference,* vol. 1, pp 114-119, 2000.
[17] Riedmiller, M. (1994) 'Rprop – Description and Implementation Details', Technical Report, University of Karlsruhe
[18] Viera, J. Anthony, and Garret Joanne M., "Understanding Interobserver Agreement: The Kappa Statistic", Family Medicine, Research Series, pp 360-364, May 2005.
[19] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, J. M. Gutierrez-Arriola., "Automatic GRBAS assessment using complexity measures and a multiclass GMM based detector", *Proceedings of the 7th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, MAVEBA 2011,* Florence, Italy, August 25-27, 2011.