AUTOMATIC GRBAS ASSESSMENT USING COMPLEXITY MEASURES AND A MULTICLASS GMM-BASED DETECTOR

J. D. Arias-Londoño¹, J. I. Godino-Llorente ², N. Sáenz-Lechón², V. Osma-Ruiz², J.M. Gutiérrez-Arriola²

¹Bioinstrumentation Research Group, Universidad Antonio Nariño, Bogotá, Colombia.
²Department of Circuits & Systems Engineering, E.U.I.T de Telecomunicación, Universidad Politécnica de Madrid, Spain.

Corresponding author: J.D. Arias-Londoño (julian.arias@uan.edu.co)

Abstract: This paper presents a system for the automatic assessment of pathological voice quality according to the GRBAS protocol, which uses a short time scheme and a characterization based on 9 complexity measures, including conventional nonlinear statistics and 7 entropy based features. The classification is carried out using three different multiclass classification strategies all of them based on Gaussian Mixture Models. The performance of the system is measured in terms of efficiency and a statistical agreement index. The results show that the complexity analysis provides relevant information for the automatic assessment of voice quality according to the GRBAS protocol.

Keywords: Automatic GRBAS assessment, Complexity measures, multiclass classification.

I. INTRODUCTION

In the clinical environment, evaluation of voice is usually carried out by means of a combination of perceptual evaluations and acoustic parameterizations of the speech trace. Perceptual evaluation consists on a subjective diagnosis of voice quality, based on comparisons with other voices, 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]. Although the assessment based on acoustic parameters is becoming a usual technique of analysis, perceptual evaluation is still the most practiced method for the evaluation and clinical management of voice disorders [2]. Unfortunately, a good correlation between acoustic parameters and perceptual evaluation of voices remains unfound [3].

Perceptual evaluation has been widely criticized because its subjectivity. 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 [4]. This situation can be improved using an automatic system, which should provide accurate, reproducible and graded measures of a patient’s voice quality, helping speech and language therapists with the patient’s treatment and rehabilitation [5]. 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 [6]. It has been demonstrated that, on the basis of low intra-rater and inter-rater variances, the GRBAS protocol seems to be the most reliable and relevant perceptual voice quality evaluation [1].

On the other hand, the complexity analysis of pathological voices seeks to quantify the effects of nonlinear phenomena involved in the voice production process, due to changes in the dynamic properties of the vocal cords and laryngeal tissues because of the presence of pathology. This kind of analysis has demonstrated to provide more stable results than conventional acoustical analysis when the voice signals do not present a quasiperiodic structure [7]. Additionally, the information obtained using complexity analysis has demonstrated to be relevant for the evaluation of different types of laryngeal pathologies [7], and also complementary to the one obtained using conventional methods of characterization (such as noise measures and cepstral coefficients) for the automatic detection of pathological voices [8].

In this sense, this work explores the discrimination capabilities of nine complexity measures for the perceptual evaluation of pathological voices according to the GRBAS protocol and their use in an automatic system for the assessment of voice quality. Since each scale of the GRBAS protocol can take one of four different values (classes), rating a voice according to it, can be seen as a multiclass problem Therefore, the classification is carried out using three different multiclass strategies based on binary Gaussian Mixture Models (GMM) classifiers: One vs All, All vs All and Hierarchical (like tree) classification.

The results are shown in terms of efficiency and statistical agreement index. The subset of complexity
measures along with the classification strategy which provided the best result for each scale of the GRBAS protocol are also reported.

II. METHODS

In the first stage of the process, the speech signal is framed and windowed in order to perform a short-time analysis. This approach is well established in speech processing tasks, including speech recognition, or speaker identification and verification. Nevertheless, the nonlinear analysis of speech signals on a frame basis is a recent approach [8]. This analysis is supported by the fact that changes in the dynamics of a pathological voice can be presented during long periods of time or suddenly. Slow changes in the speech signal are related to the biological processes in which the properties of the tissues evolve. On the opposite side, sudden changes can be explained by the presence of extra masses or changes in the biomechanical properties of the tissues of the vocal folds, modifying the dynamic behavior during the voice production process, and producing abrupt variations in the vibration regime of the vocal fold that can be understood as bifurcations [9]. These phenomena can be better detected and characterized using a short-time scheme. The analysis was carried out using frames of 55 ms with a 50% frame shift according to previous results [8], in which the frame length was selected taking into account criteria related to the minimum signal length that must be used for a good estimation of nonlinear features and also a minimum number of pitch periods for a good characterization of the signal stability. In the following section each of the complexity measures used in the characterization stage will be exposed.

A. Parameterization

First of all, 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 [10]. 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 9 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.

Approximate Entropy (A_E): In the field of nonlinear dynamics, complexity measures often quantify statistically the evolution of the trajectory in the embedded phase space. 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 [8]. The most employed measure in this context is A_E, which is a measure of the average conditional information generated by diverging points of the trajectory [8]. 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 [11], which cannot be asserted for voice signals.

There are several modifications of A_E published in the literature. Among them the most important is the Sample Entropy (S_E), developed with the aim of obtaining a more independent measure than A_E with respect to the signal length.

Recurrence and fractal scaling analysis: Considering that there is a combination of both deterministic and stochastic components in the voice signal during phonation [11], the deterministic component can be
characterized by a measure called *Recurrence period density entropy* (RPDE) and the stochastic component by means of a *Detrended 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 [11]. 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 [8]. 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 ($H_{MC}$), and two empirical estimations of the DHMM entropy: Shannon entropy ($H_{ES}$) and Renyi entropy ($H_{E2}$).

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.

**B. Classification**

As previously commented, each scale of the GRBAS protocol can take one of four different values (classes), therefore the classification of a voice according to such protocol can be seen as a multiclass classification problem. In this sense, the classification in this work is performed using three different multiclass strategies based on binary classifiers, namely *One-vs All*, *All-vs-All* and *Hierarchical*, all of them employing GMM as the core of the pattern classifier stage.

- **One vs all:** In this approach, a binary classifier discriminates between a given class and the other $n_c-1$ classes. For this approach, the number of binary classifiers required is $N=n_c$, where the $k$-th classifier is trained with positive examples belonging to class $k$ and negative examples belonging to the other $n_c-1$ classes. When testing an unknown pattern, the classifier that provides the maximum output is considered the winner, and the label of this class is assigned to that pattern.

- **All vs all:** In this approach, a binary classifier is built to discriminate between every possible pair of classes, while discarding the rest of the classes. This requires $N=n_c(n_c-1)/2$ binary classifiers. When testing a new example, a voting is performed among the classifiers and the class with the maximum number of votes wins.

- **Hierarchical:** Another way to address the multiclass classification problem is to perform a hierarchical division of the output space, i.e. arranging the classes like a tree. The tree is created in such a way that the classes at each parent node are divided into a number of clusters, one for each child node. The process continues until the leaf nodes contain only a single class. At each node of the tree, a simple classifier, usually a binary classifier, makes the discrimination between the different child class clusters. Following a path from the root node to a leaf node leads to a classification of a new pattern. This method uses $N=n_c-1$ binary classifiers for an $n_c$-class problem.

**C. 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. The validation was performed using a leave-one-out crossvalidation strategy due to the small number of voice recorders belonging to some of the classes.

**III. Results**

Table I shows the set of complexity measures which perform better for each scale of the GRBAS protocol. From these results it is possible to know which characteristics provide the largest contribution to the automatic evaluation of each scale of the GRBAS protocol. The best sets of features were determined based on a brute-force search and discriminative criteria. Table
I also shows the efficiency achieved for every scale. It is also worth to note that, in most of the cases, the best performance was obtained by using a “One-vs-All” multi-class classification strategy.

Table I. Best sets of complexity measures and efficiency obtained for each scale of the GRBAS protocol.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Set of features</th>
<th>Efficiency [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>CD, Ar, Sx, DFA, HFC</td>
<td>56.44</td>
</tr>
<tr>
<td>R</td>
<td>LLE, CD, Ar, Sx</td>
<td>55.11</td>
</tr>
<tr>
<td>B</td>
<td>LLE, Sx, DFA, HFC, HFB</td>
<td>57.18</td>
</tr>
<tr>
<td>A</td>
<td>LLE</td>
<td>66.67</td>
</tr>
<tr>
<td>S</td>
<td>HFC</td>
<td>46.67</td>
</tr>
</tbody>
</table>

Table II shows estimations of the statistical agreement index, Kappa, for each scale of the GRBAS protocol. The Kappa index measures the agreement between the classification provided by the system and the evaluation supplied by the specialist who labeled the database. For the sake of comparison, table II also shows the agreement obtained between to experienced medical specialist for the assessment of pathological voices according to the GRBAS protocol reported in [1]. The results show that although the agreement obtained by the system is still lower than the one obtained by two specialists, the results are comparable and therefore the information obtained from the complexity analysis and the classification methodology employed in this work can be useful for improving the automatic assessment of voice quality according to the GRBAS protocol.

Table II. Kappa indexes obtained by the system for each scale of the GRBAS protocol.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Kappa</th>
<th>Kappa in [1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0.40</td>
<td>0.51</td>
</tr>
<tr>
<td>R</td>
<td>0.40</td>
<td>0.46</td>
</tr>
<tr>
<td>B</td>
<td>0.37</td>
<td>0.43</td>
</tr>
<tr>
<td>A</td>
<td>0.32</td>
<td>0.41</td>
</tr>
<tr>
<td>S</td>
<td>0.24</td>
<td>0.34</td>
</tr>
</tbody>
</table>

IV. DISCUSSION AND CONCLUSIONS

The analysis of agreement between the automatic system and the rater who labeled the database showed that the performance of the system is a bit lower compared to the agreement obtained by two experienced specialist reported by [1]. Nevertheless, the results show that complexity analysis provides relevant information for this task. It is worthy to note, that the nonlinear analysis of speech signals is proposed as a complement of the analysis based on classical acoustic parameters. Therefore, it is very likely that, similar results to the obtained in [8] for the detection of pathological voices, can be reached by means of the combination of conventional and nonlinear analysis, improving the automatic rate of voices according to perceptual criteria. The multiclass classification strategy showed interesting results, however, it remains open the problem of dealing with a small number of samples in some classes.

ACKNOWLEDGMENTS

This research work has been financed by the Universidad Antonio Nariño, Colombia, under grant 2010238 - PI/UAN-2011-476bit and Spanish government through the project grant TEC2009-14123-C04-02.

REFERENCES