Automatic Diagnosis of Pathological Voices

GASTÓN SCHLOTTHAUER, MARÍA EUGENIA TORRES
Universidad Nacional de Entre Ríos
Lab. Señales y Dinámicas no Lineales
CC 47 Suc 3 (3100) Paraná
ARGENTINA
gschlott@bioingenieria.edu.ar, metorres@ceride.gov.ar
CRISTINA JACKSON–MENALDI
Wayne State University
School of Medicine
Detroit, Michigan
USA

Abstract: Spasmodic dysphonia (SD) is a voice disorder characterized by voice breaks. Muscle tension dysphonia (MTD) is a form of voice misuse characterized by excessive muscular effort. While the first pathology is not a psychological condition and has a neurological origin, the last one does not include a neurological disorder and is correctable with voice therapy. Patients with SD are often not identified for treatment. These two pathologies are only correctly differentiated by experts. The importance of a correct diagnosis is directly related with the application of the suitable treatment. Our goal is to provide voice pathologists with a new tool to confirm their diagnosis. In the present work, we present a preliminary approach to this problem, building an automatic classifier using acoustical measurements on registered sustained vowels /*a*/ and pattern recognition tools based on neural networks. As long as we know, there are not previous published works in automatic classification of these two pathologies. However, there are works on automatic classification between normal and pathological voices. Our results overcome the best reported classification between pathological and normal voices, and have a good discrimination between SD and MTD.

Key–Words: spasmodic dysphonia, muscle tension dysphonia, pattern recognition.

1 Introduction

Spasmodic dysphonia (SD) is a term applied to patients with specific voice problems. Symptoms of interruptions in fluency may be misdiagnosed as spasmodic dysphonia. Over the last decade, the term spastic dysphonia has been replaced with the term spasmodic because the phonatory characteristics are more consistent with speech spasm than with spasticity or rigidity. Onset occurs slowly over a period of months or years. SD is divided into three types: *adductor spasmodic dysphonia* (AdSD), *abductor spasmodic dysphonia* (AbSD), and *mixed dystonia*. The AdSD is a most common problem than AbSD. AdSD is characterized by hyperadduction of the vocal folds, irregular interrupted, effortful, strangled, strained, staccato voice. AbSD is characterized by contraction of the posterior cricoarytenoids muscles and produce breathy interruptions. SD, a neurological disorder of muscle tonicity, is a true dystonia, which some times coexists with tremor, usually obvious during sustain vowels. Thus it needs to be differentiated of severe *muscular tension dysphonia* (MTD), also called *muscle tension dysphonia*, a psychogenic dysphonia wich may be misdiagnosed.

SD needs to be diagnosed with basic history, physical examination, voice pathologist assessment, objective voice analysis, MRI of the brain, laryngeal EMG, and laboratory tests. Laryngeal assessment needs to be done with videostrolaryngoscopy with flexible and rigid endoscope. It helps in detecting other neurological problems, that sometimes appear mimicking SD.

SD speech is characterized by intermittent voice offsets in the middle of the vowels and the essential symptom is voice breaks.

MTD is a form of voice abuse characterized by excessive muscular effort, and usually by pressed phonation. A diagnosis of MTD suggests that such patients do not have a neurological motor control disorder, but a functional voice production disorder due to excessive laryngeal tension.

SD is clearly not a psychological condition; however, as most of other voice disorders, stress can make SD worse, and voice therapy can make it better. The underlying condition may easily be confused with MTD, which is correctable with voice therapy. It must be emphasized that SD is actually organic (neurological).

Traditional objective voice measures for patients with SD may not always be helpful in the differential diagnosis for the wide variation of findings across subjects. Acoustical analysis may help to diagnose a SD. Fundamental frequency from conversational sample may be useful in identifying each patient's compensatory strategy for managing his or her vocal spasms with extreme muscle tension [1, 2].

Acoustical measures of vocal function are routinely used in the assessments of disordered voice. They are very appealing due to their noninvasive nature. The extraction of such measures from sustained vowel samples is common because of its simpler acoustic structure. In recent years, the use of these measures, in combination with pattern recognition techniques, has motivated the appearance of several works concerning the automatic discrimination between pathological and normal voices. Recent works deal with the classification between normal and pathological voices [3–7], however in them the pathologies are grouped in the same set. The best classification was obtained by Parma and Jamieson [5] using nine acoustic measures and achieving an accuracy of 96.5 %. As previously stated, it is an interesting challenge to develop an automatic tool in order to help voice pathologists in the diagnosis of SD and MTD. With this in mind, in the present work we attempt to separate normal from pathological voices and additionally to obtain a classification in SD or MTD for the pathological cases. In this preliminary approach, we use only one sample speech of each patient. Eight well–known acoustic parameters are extracted for each voice, conforming a pattern. These patterns are then classified into three categories: Normal, SD, and MTD.

2 Materials and Methods

The analyzed voices were obtained from 89 speakers divided into 36 dysphonics (15 patients with MTD and 21 with SD, more specifically AdSD) and 53 normal speakers. These speech signals are registers of the sustained vowel /*a*/. The subjects were instructed to sustain the vowel, /*a*/, for at least 3 s at a comfortable pitch and loudness. The samples were recorded at a rate of 22 kHz and with a resolution of 16 bits. Acoustic measures are then extracted from the sustained vowels, including short–term perturbations of fundamental frequency and intensity (termed *jitter* and *shimmer*, respectively), and glottal noise measures. A consensus does not exist on the utility of these measures for discriminating between normal and pathological voices [7]. We build an eight–dimensional vector associated with each patient, and we use the nonlinear properties of NN for classification [8, 13]. As will be seen, the inclusion of these measures in combination with nonlinear techniques allow us to attain accurate discriminations. Due to the small number of available data, we have applied the *Leave–One–Out* method in all the cases. This means that the classification space is computed with every case in the database except the one that is being classified. In this way, the classification results are more realistic and close to the true classification rates [8, 9]. We present here the parameters considered in order to construct the vector associated to each patient. The estimation of fundamental frequency (F_0) is of special importance, since the calculation of other parameters is based on its good estimation. Due to its robustness, the Waveform Matching (WM) algorithm is the one of choice for pathological voices or in the presence of moderate levels of background noise [10], thus we adopted it for F_0 extraction.

1. Degree of Voice Breaks or Unvoiceness [3]. This parameter is the total duration of the breaks between the voiced parts of the signal, divided by the total duration of the analyzed part of the signal. Silences at the beginning and at the end of the signal are not considered breaks.

2. Period Perturbation Quotient (*Jitter*) [11].

   (a) Jitter Ratio (Local) or *jitt.* This is the simplest form of F_0 adjusted perturbation index. By definition:

$$jitt = 1000\,\frac{\frac{1}{n-1}\sum_{i=1}^{n-1}|P_i - P_{i+1}|}{\frac{1}{n}\sum_{i=1}^{n} P_i}, \quad (1)$$

   where P_i is the period of the i^{th} cycle, in ms, and n is the number of periods in the sample.

   (b) Relative Average Perturbation (*RAP)*.

$$RAP = \frac{\frac{1}{n-2}\sum_{i=2}^{n-1}\left|\frac{P_{i-1}+P_i+P_{i+1}}{3} - P_i\right|}{\frac{1}{n}\sum_{i=1}^{n} P_i}. \quad (2)$$

(c) Five–point Period Perturbation Quotient (*ppq5*).

$$ppq5 = \frac{\frac{1}{n-4}\sum_{i=3}^{n-2}\left|\frac{\sum_{j=-2}^{2}P_{i+j}}{5} - P_i\right|}{\frac{1}{n}\sum_{i=1}^{n}P_i}. \tag{3}$$

3. Amplitude Perturbation Quotient (*Shimmer*) [11].

(a) Shimmer (*shimm*).

$$shimm = \frac{\frac{1}{n-1}\sum_{i=1}^{n-1}|A_i - A_{i+1}|}{\frac{1}{n}\sum_{i=1}^{n}A_i}. \tag{4}$$

where A_i is the amplitude of the i^{th} cycle, in ms, and n is the number of periods in the sample.

(b) Three–point Amplitude Perturbation Quotient (*apq3*).

$$apq3 = \frac{\frac{1}{n-2}\sum_{i=2}^{n-1}\left|\frac{A_{i-1}+A_i+A_{i+1}}{3} - A_i\right|}{\frac{1}{n}\sum_{i=1}^{n}A_i}. \tag{5}$$

(c) Eleven–point Amplitude Perturbation Quotient (*apq11*).

$$apq11 = \frac{\frac{1}{n-10}\sum_{i=6}^{n-5}\left|\frac{\sum_{j=-5}^{5}A_{i+j}}{11} - A_i\right|}{\frac{1}{n}\sum_{i=1}^{n}A_i}. \tag{6}$$

4. Harmonics–to–Noise Ratio (*HNR*).
This parameter quantifies the amount of glottal noise in the vowel waveform. In contrast to perturbation measures, it attempts to resolve the vowel waveform into signal and noise components, computing their energies ratio. In order to obtain this measure, we adopted the algorithm described in [12].

2.1 Neural Networks

Neural networks (NN) are broadly used in the field of pattern recognition [8, 13]. In the present work Multilayer Perceptron (MLP) has been applied. The resilient backpropagation algorithm was chosen for training, due to its excellent performance for pattern recognition problems [14].

In order to reduce the number of input units in the neural network, we performed a principal component analysis (PCA). In our case, six components contributed with 99.5% of the variance in the data set, meaning that the size of the NN input vectors was reduced from eight to six.

We used hyperbolic tangent activation function in the hidden layer and in the output layer. In order to select the best number of neurons at the hidden layer, its size has been varied from 8 to 34, running 100 experiments in each case. The output layer had three neurons, one for each class (SD, MTD and Normal). The “winner” output was the neuron with the highest value, and the input vector was classified with the class associated with that output.

3 Results

In this section we present the results obtained with proposed approach. By means of a statistical test, we quantify the results obtained for different number of neurons at the hidden layer.

For the purpose of comparing the performance of different network sizes we ran 100 experiments with each configuration registering the mean classification error for each network size. The mean percentage of misclassifications and the corresponding standard deviation for each number of hidden units is depicted in Fig. 1. As it can be seen, the error seems to be stabilized after 14 hidden units.

In order to contrast these mean errors and to obtain information about which mean errors pairs are significantly different and which ones are not, Tukey test with a significance level $\alpha = 0.05$ was performed. It is recommended as multiple comparison test for the family of all pairwise comparisons [15].

In Table 1 we show the results obtained with Tukey test. Each group is labeled using the number of hidden units of the corresponding NN. The first column corresponds to the group. The mean error (mean of the percentage of misclassifications) is presented in the second column. The third column shows the groups that are not significantly different from the one indicated in the first one, with a significance level of $\alpha = 0.05$. From the analysis of this table, we can conclude that when the number of hidden units in the NN is increased above 14, the error is not significantly reduced, in agreement with the results observed in Fig. 1.

In Tables 2, 3, and 4 we show, as an example, the best confusion matrices obtained for each one of the three different numbers of hidden units: 14, 16, and 22 neurons, respectively. We can appreciate that in the three cases the best results reach a 93.26% of correct classifications, and 100% of correct discrimination between pathological and normal voices, even if some pathological ones have been confused. Although the total percentage of discrimination obtained

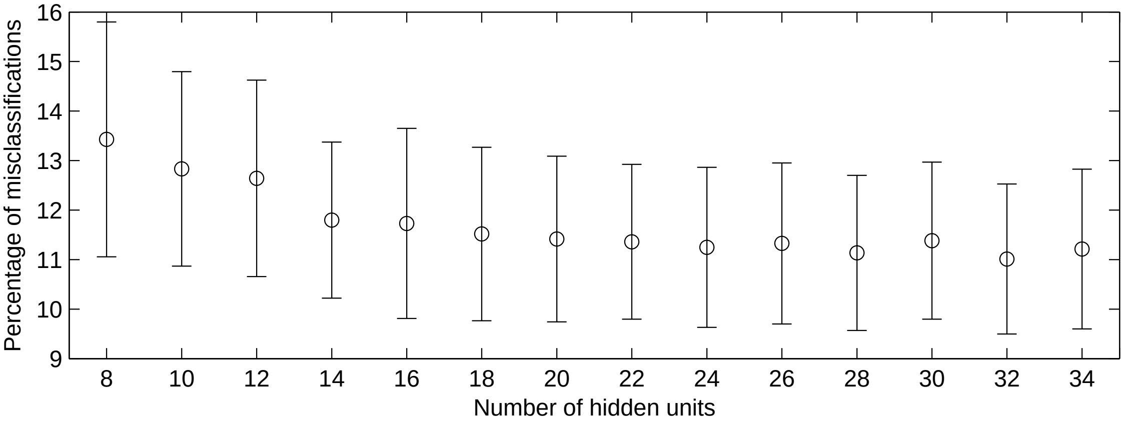

Figure 1: Percentage of misclassifications (three classes). Errors and standard deviation plotted as function of the number of hidden units of the NN.

Table 1: Results of the Tukey multiple comparison test. Groups are labeled using the number of hidden units of the neural network. Second column: Mean error of the percentage of misclassifications. Third column: groups that are not significantly different from the "group" in the first column, with a significance level of $\alpha = 0.05$.

Group	Mean (%)	Groups with means
(hidden units)		not significantly different
32	11.01124	32 28 34 24 26 22 30 20 18 16 14
28	11.13483	32 28 34 24 26 22 30 20 18 16 14
34	11.21348	32 28 34 24 26 22 30 20 18 16 14
24	11.24719	32 28 34 24 26 22 30 20 18 16 14
26	11.32584	32 28 34 24 26 22 30 20 18 16 14
22	11.35955	32 28 34 24 26 22 30 20 18 16 14
30	11.38202	32 28 34 24 26 22 30 20 18 16 14
20	11.41573	32 28 34 24 26 22 30 20 18 16 14
18	11.51685	32 28 34 24 26 22 30 20 18 16 14
16	11.73034	32 28 34 24 26 22 30 20 18 16 14
14	11.79775	32 28 34 24 26 22 30 20 18 16 14
12	12.64045	12 10 8
10	12.83146	12 10 8
8	13.42697	12 10 8

Table 2: Best confusion matrix for 14 hidden units.

	Classifications			Correct
Class	SD	MTD	Normal	Classifications
SD	20	1	0	95.24 %
MTD	5	10	0	66.67 %
Normal	0	0	53	100.00 %
TOTAL				93.26 %

Table 3: Best confusion matrix for 16 hidden units.

	Classifications			Correct
Class	SD	MTD	Normal	Classifications
SD	17	4	0	80.95 %
MTD	2	13	0	86.67 %
Normal	0	0	53	100.00 %
TOTAL				93.26 %

in the three tables are equal, we can see that those corresponding to each class (pathology) are different. The highest percentage of correct classifications of SD voices was obtained with 14 hidden units (see Table 2.) However the best result for MTD recognition was achieved using 16 hidden neurons (Table 3.) It can be observed that an intermediate result was obtained with 22 hidden units (Table 4.)

The minimum mean error (11.01 %, see Table 1) along the 100 experiments with the different numbers of hidden neurons was obtained with 32 units. The averaged confusion matrix for this case is shown in Table 5. With this network configuration, the mean of correct classifications was 88.99 %. It is important to add that, in this case, the mean of normal voices correctly classified was 99.96 % and joining SD and

Table 4: Best confusion matrix for 22 hidden units.

	Classifications			Correct
Class	SD	MTD	Normal	Classifications
SD	18	3	0	85.71 %
MTD	3	12	0	80.00 %
Normal	0	0	53	100.00 %
TOTAL				93.26 %

Table 5: Average of the 100 confusion matrices for 32 hidden units. Pathological voices are classified as pathological (SD or MTD) in the 97.63 % of cases, and normal voices in 99.96 % are correctly classified.

	Classifications			Correct
Class	SD	MTD	Normal	Classifications
SD	15.78	4.48	0.74	75.14 %
MTD	4.45	10.44	0.11	69.60 %
Normal	0.01	0.01	52.98	99.96 %
TOTAL				88.99 %

MTD in a single class (pathological voices), a 97.63 % of correct classifications was achieved.

In order to compare our results with previous works, we tested the ability of our classifier for discrimination between pathological and normal voices. For this purpose we changed the output layer, leaving now two output neurons. Increasing the number of hidden units from one to 14, and running 100 realizations in each case, we obtained the minimum mean error while working with eight hidden neurons. Furthermore, in many realizations (even in those ran with only one hidden unit) the classification was perfect. In Table 6 we present an averaged confusion matrix for the 100 realizations. The percentage of correct classifications reached 98.94 % (97.58 % of pathological voices and 99.87 % of normal voices correctly classified). This result is better than the highest percentage of correct classifications found in the literature (96.5 % of correct detections, presented in [5]).

In Fig. 2 the mean percentage of misclassifications is depicted. The minimum error was obtained with eight neurons. A Tukey test with a significance level of $\alpha = 0.05$ was applied, indicating that the MLP with 2 or more neurons in the hidden layer are not significantly different.

Table 6: Classification in two categories. Average of the 100 confusion matrices for 8 hidden units.

	Classifications		Correct
Class	Pathological	Normal	Classifications
Pathological	35.13	0.87	97.58 %
Normal	0.07	52.93	99.87 %
TOTAL			98.94 %

4 Discussions and Conclusions

In the previous section we have presented a new approach for the automatic classification of pathological voices. In particular we have focused our attention on two different aspects: i) to discriminate between normal and pathological voices and ii) to discriminate between normal, spasmodic dysphonia and muscle tension dysphonia, two pathologies for which it did not exist till now a proper automatic differential diagnosis and misdiagnoses is quite often.

These preliminary results obtained in this work suggest that a voice pathology classification is possible using only acoustical measures that are well–known by both the speech physicians and the therapists. This is a very important characteristic of our approach due to the knowledge that the specialists have on these parameters. However, further studies are in progress in order to incorporate the information that can be extracted of read text in order to improve the automatic classification.

The SD pathology was recognized in a 95.24 % (see Table 2.) In the case of MTD, it arrived to a 86.67 % of correct classifications (see Table 3.) For normal voices, a perfect recognition was achieved, for example see Tables 2, 3, and 4. Considering the percentage of correct classifications in the three classes, the best result was obtained reaching a 93.26 % of successful classifications (for example, in Tables 2, 3, and 4.) In case of separating pathological and normal voices, it is also possible to reach a perfect discrimination and, as it can be seen in Table 6, these results overcome those published to date. The results with a MLP with eight hidden neurons averaged 98.94 % of correct classifications in 100 realizations, overcoming the best reported percentage of correct classifications (96.5 %, [5].) classifiers for pathological and normal voices, the obtained results allow to conclude that in the present application the Multilayer Perceptron neural networks, with eight hidden unites, are more successful than SVM. The contribution of this work lies in that it is possible to distinguish between two pathologies that often cause erroneous diagnosis by not highly specialized professionals. As long as we know, there are not previous works for automatic discrimination between the analyzed pathologies. In order to introduce the presented automatic classification tools in the clinical usage it would be necessary to refine and test it on a higher number of normal and pathological voices. In future works, these results

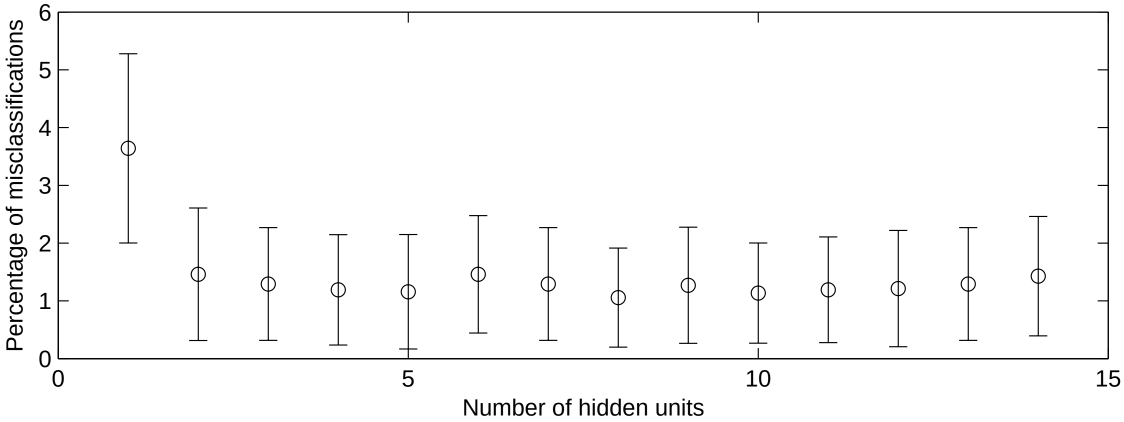

Figure 2: Percentage of misclassifications (two classes). Errors and standard deviations plotted as function of the number of hidden units of the NN.

could be improved using read text. obtained results are very promising and suggest that to achieve this goal is close. Nevertheless, the presented preliminary results are very promising and suggest that it is close to reach our goal.

5 Acknowledgments

This work is supported by A.N.P.C.T. under Project PICT N^{o} 11-12700, Universidad Nacional de Entre Ríos and CONICET.

References:

[1] C. Jackson-Menaldi, *La voz patológica*, Editorial Médica Panamericana, 2002.

[2] R. T. Sataloff, *Clinical assessment of voice*, Plural Publishing, Inc., 2005.

[3] B. Boyanov and S. Hadjitodorov, Acoustic analysis of pathological voices, *IEEE Engineering in Medicine and Biology Magazine* 16 (4), 1997,pp. 74–82.

[4] M. de Oliveira Rosa, J. C. Pereira, and M. Grellet, Adaptive estimation of residue signal for voice pathology diagnosis, *IEEE Transactions on Biomedical Engineering* 47, 2000, pp. 96–104.

[5] V. Parsa and D. G. Jamieson, Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech, *Journal of Speech, Language, and Hearing Research* 44, 2001, pp. 327–339.

[6] J. I. Godino–Llorente and P. Gómez–Vilda, Automatic detection of voice impairments by means of short–term cepstral parameters and neural networks based detectors, *IEEE Trans. Biomed. Eng.*, 51, 2004, pp. 380–384.

[7] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, Discrimination of pathological voices using a time–frequency approach, *IEEE Trans. Biomed. Eng.* 52, 2005, pp. 421–430.

[8] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, 2nd Edition, Wiley, 2001.

[9] P. Lachenbruch, R. Mickey, Estimation of error rates in discriminant analysis, *Technometrics* 10 (1968) pp. 1–11.

[10] V. Parsa, D. G. Jamieson, A comparison of high precision F0 extraction algorithms for sustained vowels, Journal of Speech, Language, and Hearing Research 42 (1999) pp. 112–126.

[11] R. J. Baken, R. F. Orlikoff, Clinical Measurement of Speech and Voice, 2nd Edition, Singular Thomson Learning, 2000.

[12] G. de Krom, A cepstrum-based technique for determining a harmonic–to–noise ratio in speech signals, Journal of Speech, Language, and Hearing Research 36 (1993) pp. 254–266.

[13] C. M. Bishop, Neural Networks for Pattern Recognition, Oxford, U. K., Clarendon, 1995.

[14] M. Riedmiller, H. Braun, A direct adaptive method for faster backpropagation learning: The RPROP algorithm, in: Proceedings of the IEEE International Conference on Neural Networks, IEEE, USA, 1993, pp. 586–591.

[15] J. A. Rafter, M. L. Abell, J. P. Braselton, Multiple comparison methods for means, SIAM Review 44 (2002) pp. 259–278.