2021 18th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE). Mexico City, Mexico. November 10-12, 2021

Phoniatric System Based on Acoustical Analysis for Early Detection of Anomalies in Voice Production

Ana Laura Cazarin
Mecatrónica
Universidad Politécnica de Chiapas
Mexico
acazarin@upchiapas.edu.mx

Eladio Cardiel
Electrical Engineering Department
Centro de Investigación y de Estudios Avanzados IPN
Mexico City, Mexico
ecardiel@cinvestav.mx

Laura I. Garay-Jimenez
UPIITA
Instituto Politécnico Nacional
Mexico City, Mexico
lgaray@ipn.mx

Pablo Rogelio Hernández
Electrical Engineering Department
Centro de Investigación y de Estudios Avanzados IPN
Mexico City, Mexico
pablo.rogeli@cinvestav.mx

Victor Manuel Valadez Jiménez
Phoniatry service
Instituto Nacional de Rehabilitación
Mexico City, Mexico
vvaladez@inr.gob.mx

***Abstract*— An objective evaluation of voice quality using phoniatric parameters to analyze acoustic signals generated during phonations has been developed. The system used the Voice Handicap Index questionnaire to quantify the perception of the vocal capacity of people, complemented with the measurement of voice acoustic signal parameters obtained from sustained vocalizations. The parameters were statistically evaluated, and then a systematic analysis of seven classifiers with the preselected parameter was done with and without PCA, KNN, BPNN, and SVM obtained the best performance with a mean accuracy of 97.4%, 96%, and 94.7%. The proposal can be used to determine alterations in the production of the voice for medical diagnostics.**

Keywords—Phoniatric, Acoustical Analysis, Voice Quality

I. INTRODUCTION

The voice is the fundamental tool of human communication. As a consequence of technological communications, the use of voice and vocal disorders have increased, which occur in 3 to 9% of the population. [1]. Among the most affected working-age people, who use the voice as a work tool, are the teaching professionals, vendors, speakers, actors, singers, and speakers. Current commercial instruments in phoniatrics do not have a complementary objective evaluation of voice quality as a reference in the tests that interpret the state of phonation structures[2]. Due to this, the diagnosis of phonological medicine is based mainly on a set of clinical studies derived from evaluation protocols that use invasive techniques, which cause discomfort, fear, and sometimes injury to patients. In addition, these evaluations are carried out in specialized laboratories with limited access, which restricts the early detection of voice anomalies through fieldwork[3].

The analysis of voice for diagnosis was studied for several conditions such as continuous speech[4] or vocalizing a word[1][5], considering gender and age[2], self-perception[3], and even differences due to languages[5] and emotional conditions[6]. Pishgar et al. present a well-done summarize of research in voice diagnosis [7]. Voice signal is a complex signal that has to be analyzed with a multiparameter approach to identify pathological cases among healthy cases[8]. Several approaches for automatic searching, such as convolutional neural networks, SVM, KNN, CART, logistic regression, and even linear discriminant, have been used for classifying health and pathology differences[4][7][9][10]. The results showed that according to the sample size, type of diseases, number of parameters, and specific computational requirements, the performance could range from 60 to 92%. So a systematic review of the parameters and classifiers under identical conditions for this phoniatric system has to be part of the designing model.

This paper presents the development of a system based on sound analysis for the early detection of anomalies in the voice. The Voice Handicap Index questionnaire is applied as a first tool to quantify the perception of the vocal capacity. Then, this information is complemented with the measurement of voice acoustic signal parameters obtained from sustained vocalizations. This data set was analyzed and used to define a methodology based on classifiers to detect alterations in the voice production for diagnostics.

II. METHODOLOGY

The acoustic signal is a parameter that can be used to measure voice disturbances objectively.

978-1-6654-0029-9/21/$31.00 ©2021 IEEE

Its recording is non-invasive and allows establishing a diagnosis, helps in therapy monitoring, and generates lines of research in the multidimensional modeling of the phonatory system.

Questionnaire

The Voice Handicap Index-10 (VHI-10) questionnaire was adopted to complement the information provided by measuring anatomical and physiological variables. The tool is based on a self-assessment to quantify the perception of vocal disabilities[11].

A. Analysis Parameters

The fundamental frequency (F0), Shimmer, and Jitter values were considered the most used parameters for diagnosing a pathological voice. In addition, the harmonic-noise ratio HNR and the smoothed Cepstral peak were determined [1][5][7][8].
The F_0 is the parameter related to the natural physical characteristics of the phonation structures. The vibration of the vocal cords is evaluated with the Jitter index, measuring the regularity of the fundamental frequency and the duration variability of the wave in consecutive cycles, and with the Shimmer index, evaluating the stability of the sound system by measuring the variability of the amplitude of cyclic sounds.
The Harmonic-Noise Ratio (HNR) allows detecting larynx problems related to vibrating the vocal folds due to abnormal growth or palsy. These problems can generate air leaks that, in turn, cause noise due to turbulence.
With the use of Smoothed Cepstral Peak Prominence (CPPS), a higher discrimination power has been obtained, using both sustained vowel / a / and sentences. Therefore, it proves more diagnostic accuracy in detecting dysphonia in the Spanish language [5][7].

B. Study cases

In a sample of 19 male patients with pathologies of different origins and 19 healthy male subjects, sustained phonations of the vowel / a / were processed. In pathological cases, the physician selected the subjects sample according to the diagnosis, in a range of 21-40 years old, with a mean and standard deviation of 31.46 ±5.9. The recordings were made by a phoniatric specialist in a clinical appointment with the consent of the patient and using the traditional methods. The considered diagnosed cases were functional pathologies, that is, pathologies caused by injuries to speech organs or by congenital causes. The pathologies were caused by poor vocal management, possible nodule formation, or transient alterations. Then a set of eight healthy male subjects were used for the final integration test. All healthy subjects were 21-40 years old and signed a written consent after being informed according to the Declaration of Helsinki.

C. Record of phonations

An electret microphone was used, with a quasi-flat frequency response in the 20 to 20 kHz range and a typical sensitivity of -44 dB. The device was placed in a mask with a low-pressure seal (Hudson RCI, China), in front and 4 cm from the mouth of the study subject, and connected to the amplifier circuit suggested by the manufacturer for MAXIM MAX6666 device. The ESP32 microcontroller (Espressif Systems, Shanghai) was used, including an A / D converter with a sampling frequency of 8 kHz for data acquisition. It is based on a 32-bit Tensilica Xtensa LX6 microprocessor, low power consumption, and high-frequency clock (typically 160 MHz). Then a MicroSD Card Adapter was used for data storage, with a Serial Peripheral Interface (SPI) communication interface and a microSD memory with 2GB storage capacity. Finally, a Graphic User Interface (GUI) was implemented to interact with the user. This interface includes questionnaires and clinical data collected from the subjects.

D. Acoustical parameters

From the widely used clinical tests and the Praat software, 15 parameters were obtained: the fundamental frequency, Jimmer and Shitter to evaluate the disturbances in amplitude and frequency, the HNR factor to assess the harmonic-noise relationship, and spectral measurements [1][7][12].

E. Statistics

Using the Graph Pad PRISM® program, the values of the parameters obtained were subjected to a normality analysis. Subsequently, a Mann Whitney test was performed individually to know the dispersion of the data and then to determine if there were significant differences between the data of healthy people and those with some pathology. Finally, a systematic multidimensional two-way variation (Two-way ANOVA) test was performed to check the effectiveness of the parameters according to whether they presented statistically significant differences between the group of healthy and pathological subjects.

F. Selection of the classifier

After the statistical analysis, a neural network was implemented in MATLAB® using the groups identified. In this case, deep learning networks were non considered because of the size of the dataset. So a neural network consisting of the input layer with data of the thirteen parameters obtained from the analysis in the software Praat, a hidden layer of 13 neurons, and a layer with binary output were proposed. The database was divided into 70% for training, 15% for validation, and 15% for tests, randomly selected. The evaluation was carried out using confusion matrices to evaluate the mean accuracy of test, validation, and training stages repeated ten times.
Due to the number of attributes to be considered for the classification and the reduced number of samples available, a pertinence and effect test was carried out. Principal Component Analysis (PCA) technique was used to define the number of variables that provide more information under an automatic classification scheme and its impact on training time and evaluation time to be used online. Performance among the seven most common classifiers in the automated classification

domain was compared. Table 1 shows the twenty-one variants of the classifiers tested to identify the best behavior according to the most significant hyperparameters.

Table 1. Test of seven classifiers with main differences.

Test	Classifier	Kernel
1	Support Vector Machine (SVM)	Linear
2		Cuadratic
3		Cubic
4		Medium Gaussian
5		Coarse Gaussian
6	Logistic regression	
7	Trees	Fine (Levels=100)
8		Medium (Levels=20)
9		Coarse (Levels=4)
10	Discriminant	Linear
11		Cuadratic
12	Naive Bayes	Gaussian
13	KNN	Fine (K=1, euclidean distance)
14		Medium (K=10, euclidean distance)
15		Medium (K=10, cosine distance)
16		Medium (K=10, cubic distance)
17	Ensemble	Boost trees
18		Bagged trees
19		Subspace discriminant
20		RUS Boost
21		Subspace KNN

The robustness of the method was determined using a cross-validation test with k = 10 folds, and the evaluation was done with accuracy and misclassification cost. In all cases, the average value, the standard deviation, and the maximum and minimum were obtained to establish the best option for method and conditions.

Finally, the method conformed with the selected classifier, the questionnaire, the vocal recording option, and the clinical parameters extraction was tested with eight healthy subjects. As was expected, the result was that any of them required clinical attention.

III. RESULTS

A. Acoustic analysis

The acoustic analysis of vocalizations obtained from 19 subjects with pathologies from different origin sites and 19 healthy subjects was performed. The values of the selected parameters are shown in Table 2 with the results of the T-Student post-test. Firstly, the sample was sectioned between healthy subjects and subjects with some pathology to obtain the ranges in which each group was. A difference was observed between arithmetic means of the parameters, in addition, that the standard deviation of each parameter was higher in the case of subjects with pathology in contrast to healthy subjects.

B. Statistical analysis

A systematic multidimensional statistical analysis was carried out with the GraphPad PRISM® program. All parameters were separated into groups considering the ranges and units of measurement. The results of the two-ways ANOVA comparing by groups are: 1) In group one, the fundamental frequency (F_0) value was raised above 350 Hz in patients with pathologies. 2) In group two, the harmonic-noise ratio values obtained from the acoustic signals of healthy people are concentrated below 10 dB and above 15 dB in subjects with pathology. 3) Regarding the CPP, the value is maintained above 13 dB in healthy people cases, unlike concentrated values between 5 dB and 10 dB in pathological cases. 4) The standard deviation of group three harmonic-noise ratio (SD HNR) showed no statistically significant differences. This parameter measures dispersion over the central tendency, and the previous ones were values of central tendency. In contrast, in healthy cases, the standard deviation of the fundamental frequency did not reach a higher value than 10 Hz, unlike above 100 Hz values in patients with some pathology. 5) In group five, the local Shimmer measures, apq3, apq5, and dda, which are the measures of disturbance of the amplitude of phonation, showed different distributions, which were helpful for classification since a difference in the concentration of values of the healthy and pathological groups was observed. In general, pathological subjects tend to show more variation in the magnitude of phonation, and these parameters allow an evaluation of their low change for both states.

6) In the Local Jitter, rap, ppq5, and ddp measurements of group six, the fundamental frequency variability between the two health states might be helpful in automatic classification since there was a difference in the concentration of the values of healthy groups and pathological ones. 7) In group seven, the distribution of the absolute local Jitter was observed. Here, the highest concentration of healthy subjects was maintained below the upper limit reached by patients with pathology. However, the data dispersion did not show a statistically significant difference with the used sample in this ANOVA study.

8) Finally, in group eight, the local Shimmer dB was determined. It was observed that the maximum value obtained from the analysis of the signals of healthy subjects was positioned a little more than 1 dB, unlike patients with a pathology that can reach values greater than 1.5 dB. However, the distributions overlapped, which indicates that the groups

cannot be linearly separable. According to these results, thirteen parameter values were selected and proposed as input elements for the classifier's test, looking for a robust parameter combination for functional pathologies. The standard deviation of Fo and absolute local Jitter were discharged.

Table 2. Selected parameters values obtained from acoustic analysis in Praat with t-student p <0.05 comparison.

Parameter	**Healthy**		**Pathologic**		**t-student**
	Mean	Σ	Mean	Σ	p-value
Fundamental frequency F_0 (Hz)	127.2	14.19	183.7	77.41	**0.0075**
F_0 Standard deviation (Hz)	1.526	0.827	26.39	40.55	**0.0062**
***Jitter* local (%)**	0.395	1.713	1.7132	1.507	**0.0001**
***Jiter* local absoluto (µs)**	31.47	113.8	113.82	128.1	**0.0118**
***Jitter* rap (%)**	0.22	1.01	1.0101	0.94	**0.0002**
***Jitter* ppq5 (%)**	0.236	1.023	1.0228	0.889	**0.0003**
***Jitter* ddp (%)**	0.66	3.03	3.0304	2.819	**0.0002**
***Shimmer* local (dB)**	0.397	1.567	1.5672	0.297	**< 0.0001**
***Shimmer* local (%)**	4.382	18.01	18.01	3.833	**< 0.0001**
***Shimmer* apq3 (%)**	2.337	9.021	9.0209	2.081	**< 0.0001**
***Shimmer* apq5 (%)**	2.812	11.51	11.511	2.97	**< 0.0001**
***Shimmer* dda (%)**	7.012	27.06	27.062	6.242	**< 0.0001**
CPPS (dB)	15.98	8.529	8.529	2.721	**< 0.0001**
HNR (dB)	18.89	6.415	6.415	3.655	**< 0.0001**
HNR (dB) Standard Deviation	3.257	2.776	2.776	0.608	**0.0428**

C. Classification test

The best average performance with n = 10 training of the BPNN is shown in Table 3. In addition, Table 4 shows the statistics of the four best-optimized proposals under a cross-validation test with k = 10 folds. When a PCA method was previously applied to classification, the number of parameters was reduced from 13 to 2, covering 99.3% of the total variance. However, a decrease in accuracy of 6.25 ± 2.47 was presented. The average cost in training time was increased by 4%, in a range of 14 to 53 seconds, but the prediction time varied considerably according to the method used.

In general, without PCA, it has a range of 29-950 observations per second (obs/s), while with PCA, it was reduced to 29 to 130 obs/s. This parameter was not calculated for the neural network in parts, but the training, validation, and testing process times were not greater than 28 seconds, using the 13 variables.

Table 3. The best performing neural network and its configuration.

Backpropagation Neural Network		CE	%E	Accuracy	#epoch	Time (s)
Net:13-13-2; data set selection: aleatory with 75%,15%,15% subsets; training method: Scales Descendent Gradient	Max	10.72	27.78	100.00	54.00	28.00
	mean	4.72	4.69	96.01	28.70	10.31
	min	1.17	0.00	77.77	9.00	0.00
	sdt	3.42	8.95	7.27	17.94	10.19

Table 4. Performance and configuration of the four classifiers with the best average accuracy and the configuration

Model	Method	Accuracy (%)	Missclassification cost	PCA	Training time (s)	Prediction speed (obs/s)
K=1, distance metric=Euclidean, Distance weight=equal, Standardize data	KNN	97.40	1	13/13, 100%	26.784	250
K=10, distance metric=Euclidean, Distance weight=square inverse, Standardize data	KNN	97.40	1	13/13, 100%	26.394	120
Kernel=fine gaussian SVM, Kernel scale =0.9, box constraint level=1;2 classes one vs one, standardize data	SVM	94.7	2	13/13, 100%	20.9	450
Cosine KNN, K=10, distance metric= Minkowski (cubic), Distance weight=equal, Standardize data	KNN	92.40	3	13/13, 100%	28.99	220

On average, the classifiers that performed the best were KNN, BPNN, and SVM, with 97.4%, 96%, and 94.7%, respectively, without PCA.

For KNN classifier, a greater dependence was observed in the adjustment of the hyperparameters ($\sigma^2 = 18.41$) than with SVM (($\sigma^2 = 1.69$) or BPNN ($\sigma^2 = 7.27$). Although assemblies were tested, an average accuracy of 92.1% was obtained at best combination, but the high cost in response time and training time rises must be considered.

Once BPNN was selected, a final test was done with eight healthy subjects. The maximum score of HIV-10 was 32% of the total score that could be achieved in this survey. As expected, results showed that no one required clinical attention. The general results are presented in Table 5.

IV. Discussion

One of the aims of studying these anomalies in voice production was to use as many parameters as reported in the clinic and then analyzed them systematically to select the best option to determine alterations in the production of the voice associated with medical diagnostics of functional pathologies. The first approach was with statistical analysis and then with automatic classification. Results are congruent with previously reported, and the statistic results could be used as a guide in the clinical examination in a non-laboratory study.

If the priority is to perform automatic online classification of the parameters, the best option is SVM, without PCA, since up to 450 observations per second can be made, and the training time is also the shortest even though the classification percentage was 94% on average for this data set. If accuracy is a priority, the proposed KNN method reaches 97.4% average accuracy and is appropriate for online use. BPNN has training cost among the best; besides, once the best weights were found, the test time and computational requirements were minimal, so it was chosen as part of the phoniatric system and implemented in a visual interface for this proposal. The final integration test of the system with the group of eight healthy subjects shows consistency with the healthy training set shown in Table 5.

V. Conclusions

A non-invasive voice measurement system during phonation to detect early voice production anomalies was developed. The system is based on the analysis of sustained vocalizations, it can be used as a medical first-attention tool for fieldwork. In future work, a broad sample could be used or contrasted with the Spanish dataset if available.

Table 5. Comparison of the mean and standard deviation of the test sample against previous results with the thirteen parameters selected.

Parameters	Pathologial		Healthy		Test	
	mean	σ	mean	σ	mean	σ
F_0	183.74	77.414	127.15	14.192	124.53	20.338
F_0 DS	26.39	40.55	1.526	0.827	3.6389	1.570
CPPS	8.529	2.721	15.98	2.001	16.255	2.646
HNR	6.415	3.654	18.895	3.397	25.813	2.765
Shimmer local dB	1.5672	0.296	0.397	0.260	0.123	0.023
Shimmer local	18.010	3.832	4.381	2.755	1.330	0.279
Shimmer apq3	9.020	2.080	2.337	1.460	0.733	0.1742
Shimmer apq5	11.511	2.970	2.811	1.889	0.814	0.181
Shimmer dda	27.062	6.241	7.012	4.380	2.200	0.522
Jitter ddp	3.030	2.819	0.66	0.354	1.637	0.560
Jitter local	1.713	1.506	0.395	0.183	0.955	0.286
Jitter rap	1.01	0.94	0.22	0.118	0.545	0.186
Jitter ppq5	1.023	0.889	0.236	0.1	0.551	0.162

ACKNOWLEDGMENT

The authors thanks Eng. Nora Daniela Sánchez Rodríguez for her valuable collaboration in the original recordings dataset.

REFERENCES

[1] N. G. Elisei, "Acoustic Analysis of Normal and Pathological Voices Using Two Different Systems: Anagraf and Praat," *Rev. Psicol. y Ciencias Afines*, vol. 29, pp. 339–357, 2012, [Online]. Available: http://www.redalyc.org/articulo.oa?id=18026361002.

[2] M. Brockmann, M. J. Drinnan, C. Storck, and P. N. Carding, "Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task," *J. Voice*, vol. 25, no. 1, pp. 44–53, 2011, doi: 10.1016/j.jvoice.2009.07.002.

[3] Z. Wang, P. Yu, N. Yan, L. Wang, and M. L. Ng, "Automatic Assessment of Pathological Voice Quality Using Multidimensional Acoustic Analysis Based on the GRBAS Scale," *J. Signal Process. Syst.*, vol. 82, no. 2, pp. 241–251, 2016, doi: 10.1007/s11265-015-1016-2.

[4] F. Kazinczi, K. Mészáros, and K. Vicsi, "Automatic detection of voice disorders," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9449, pp. 143–152, 2015, doi: 10.1007/978-3-319-25789-1_14.

[5] J. Delgado-hernández and N. L. A. Jiménez-álvarez, "Precisión diagnóstica del pico cepstral de mayor prominencia en el cepstrum suavizado (CPPS) en la detección de la disfonía en español," vol. 6, no. January, pp. 1–6, 2019.

[6] C. Müller, F. Caffier, T. Nawka, M. Müller, and P. P. Caffier, "Pathology-Related Influences on the VEM: Three Years' Experience since Implementation of a New Parameter in Phoniatric Voice Diagnostics," *Biomed Res. Int.*, vol. 2020, 2020, doi: 10.1155/2020/5309508.

[7] M. Pishgar, F. Karim, S. Majumdar, and H. Darabi, "Pathological Voice Classification Using Mel-Cepstrum Vectors and Support Vector Machine," *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018*, pp. 5267–5271, 2019, doi: 10.1109/BigData.2018.8622208.

[8] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters," *Procedia Technol.*, vol. 9, pp. 1112–1122, 2013, doi: 10.1016/j.protcy.2013.12.124.

[9] M. A. Mohammed *et al.*, "Voice pathology detection and classification using convolutional neural network model," *Appl. Sci.*, vol. 10, no. 11, pp. 1–13, 2020, doi: 10.3390/app10113723.

[10] J. Lee and M. Hahn, "Automatic assessment of pathological voice quality using higher-order statistics in the LPC residual domain," *EURASIP J. Adv. Signal Process.*, vol. 2009, 2009, doi: 10.1155/2009/748207.

[11] C. A. Rosen, A. S. Lee, J. Osborne, T. Zullo, and T. Murry, "Development and validation of the voice handicap index-10," *Laryngoscope*, vol. 114, no. 9 I, pp. 1549–1556, 2004, doi: 10.1097/00005537-200409000-00009.

[12] H. Traunmüller and A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," *Dep. Linguist. Univ. Stock.*, vol. 97, pp. 1905191–5, 1994.