Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 164 (2019) 678–685

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2019

Outliers Treatment to Improve the Recognition of Voice Pathologies

Letícia Silva[a,b], Juliana Hermsdorf [a,c] ,Victor Guedes[a,d], Felipe Teixeira[a], Joana Fernandes[a], Bruno Bispo[e], João Paulo Teixeira[a,f] *

[a]*Instituto Politécnico de Bragança, Bragança (IPB), Portugal*
[b]*Federal Technological University of Paraná, Cornélio Procópio Campus, Brazil*
[c]*Federal Technological University of Paraná, Campo Mourão Campus, Brazil*
[d]*Federal Technological University of Paraná, Medianeira Campus, Brazil*
[e]*Department of Electrical and Electronic Engineering, Federal University of Santa Catarina, Florianópolis, 88040-370, Brazil.*
[f]*Research Centre in Digitalization and Intelligent Robotics (CEDRI), Applied Management Research Unit (UNIAG). IPB, Bragança, Portugal*

Abstract

In some of the processes used in data analysis, such as the recognition of pathologies and pathological subjects, the presence of anomalous instances in the dataset is an unfavorable situation that can lead to misleading results. This article presents a function that implements the identification of anomalies in dataset using the boxplot and standard deviation methods. Also was used the filling technique to treat these anomalies, in which the anomalous point value were substituted by a limit value determined by the boxplot or standard deviation methods. To improve the outliers methods some normalization processes based on the z-score, logarithmic and squared root methodologies were experimented. These outliers treatment were applied to the dataset used in the recognition of vocal pathologies (dysphonia, chronic laryngitis and vocal cords paralysis vs control), performed by a MLP and LSTM neural networks. After the experiments, both the standard deviation and the boxplot methods with z-score normalization showed very useful for pre-processing the dataset for voice pathologies recognition. The accuracy was improved between 3 and 13 points in percentage.

* Corresponding author. Tel.: +351 273 30 3129; fax: +351 273 30 3051.
E-mail address: joaopt@ipb.pt

1877-0509
10.1016/j.procs.2019.12.235

*Keywords:*Standard Deviation; Box plot; z-score Normalization; Logaritmic Normalization; Squared Root Normalization; LSTM, MLP.

1. Introduction

Data mining extracts knowledge by applying statistical and computational techniques. It is an application that has come up to solve some of the difficulties faced by traditional analytical techniques. The scalability, the high dimensionality, the complexity and the distribution of the dataset are indicated as the source of these difficulties [1] [2].

Within data mining, two large areas are considered remarkable in many applications. The classification of data, which allows assigning a class to the data, based on the characteristics of its attributes, and the detection of outliers, that play a fundamental role in the discovery of patterns in data, searching for those whose characteristics differ from one another [1] [2].

Anomalies in datasets, also known as outliers, are instances that do not follow an expected standard behavior. Many definitions have been proposed for outliers. According to [3], an outlier has behavior highly inconsistent with the other instances, that it is questioned if it has been generated by a different mechanism [2] [3]. For [4], it is an element that seems to deviate sharply from the other members of the sample in which it occurs [1] [4].

The appearance of database outliers is caused mainly by human errors, instrument errors, population drift, fraudulent behavior and changes or failures in system behavior. However, there are outlier points that are natural datasets [1].

There are algorithms for identification and treatment of anomalous occurrences found in the literature [1], [2], [3], [4] and [5], and due to the relevance of the theme, new algorithms continue to be proposed. The choice of anomaly detection technique to be used depends primarily on the problem domain and the input data.

The anomaly detection algorithms are similar in relation to the application objective, but each of them uses its own approach of whether it should be considered an outlier or not, i.e. which method should be used to detect a point with outlier behavior and the best way to correct it [2].

The identification of anomalies is implemented in systems with different purposes, such as detection of computer network intrusion, bank fraud verification, disease detection, sports, statistics and detecting measurement errors [2]. Methods from different areas are used to identify outliers, such as within the machine learning area, the classifiers perform signal classification. The classifiers perform both the recognition of pathological voices and the recognition of speech pathologies, in the case of speech signals [6].

The present work aims to treat and correct the anomalies identified in the datasets available in applications related to the diagnosis of voice pathologies, using as input the acoustic parameters jitter, shimmer, autocorrelation and HNR [7] [8]. Resulting in the maximization of the precision in the identification of pathological subjects and pathologies. The classification of pathologies is performed by a Multi-Layer-Perceptron (MLP) Artificial Neural Network (ANN) [9] and by a Long Short-Term Memory (LSTM) ANN [10].

2. Theoretical Framework

2.1. Outliers Identification Tests

The main methods of outliers identification are distinguished according to the criteria used, such as classification, distance, density, clusters and statistics [2]. The tests used in this work are the standard deviation and boxplot.

Outliers influence the calculations of the mean, standard deviation and histograms. As a result, it causes the distortion of conclusions and generalizations about the analyzed data set. Therefore, the presence of outliers in the dataset can lead to erroneous interpretations [11].

The Standard Deviation (SD) method is the most commonly used tactic to treat outliers because of its simplicity. Its utility is limited to bunch-shaped and reasonably symmetric data. Observations outside the range of two or three standard deviations above and below the observations average can be considered as outliers in the data [5]

A Box Plot (BP) is a simple and widely used tool that allows an easy visualization of the similarity between the parameters and groups of voices [5]. This method is less sensitive to outlier's values than the standard deviation. It does not make distributive assumptions, as it does not depend on the average or standard deviation [12]. Therefore, it is applicable to the normal or Gaussian distributions, i.e. symmetric distributions or with slight asymmetry [11].

The limits of the boxplot are calculated from the quartiles, lower quartile (Q1) corresponds to 25% of the data and the upper one (Q3) corresponds to 75%. Then the interquartile range (IQR) is calculated which is equivalent to the distance between the quartiles (IQR = Q3 - Q1) [12].Values that exceeds the upper and lower external limits, of boxplot, are considered outliers, i.e. atypical values. The internal limits are located at a distance of 1.5 IQR below Q1 and above Q3 [Q1-1.5IQR, Q3 + 1.5IQR]. The outer limits are located at a distance of 3 IQR below Q1 and above Q3 [Q1-3IQR, Q3 + 3 IQR] [12].

For SD and BP methods, if the data has a normal distribution, it is easier to estimate the likelihood of having outliers. Respectively, about 68%, 95%, and 99.7% of the data from a normal distribution are within one, two, and three standard deviations of the average. Therefore, an instance located beyond two or three standard deviations has a high probability of being an outlier [5]. A typical rule of thumb is to point out as an anomaly the observations that deviate more than three standard deviations from the average in a normal data distribution, considering only one dimension [13].

In addition, when data is highly distorted, transformations to normality are a common step in identifying outliers, when using a method that is quite effective in a normal distribution. Normalization operations, are useful when identifying outliers and it is important to note that the expected outliers before and after the transformation are different [5].

2.2. Normalization

Some modeling tools have their benefits by normalization, e.g., neural network, k-nearest neighbors algorithm (KNN), clustering, because such normalization operations are intended to minimize problems such as data redundancy and distorted results in the presence of anomalies [2]. Some transformations were made in the dataset. Thus, what is sought is to make a scale to stabilize a variance, and decrease the asymmetry and approximate the normal distribution of the variable [14]. Initially, the first type of transformation was applied, a statistical method from the standard Z-score, normalizing the data to have mean equal to 0 and standard deviation equal to 1, [15]. The second type was the logarithmic transformation of the elements of the set. The third type is the power transformation, with exponent ½, also called normalization through the square root [14].

3. Methodology

3.1. Acoustics Parameters

The speech parameters were extracted from a German voice bank called Saarbrucken Voice Database (SVD) [16], using the algorithms developed in [17-18]. The database with the extracted parameters is composed of pathological subjects with 19 different diseases and healthy subjects, in a total of 901 subjects. Among the pathologies, only 194 Healthy subjects, 69 subjects with Dysphonia, 41 subjects with Chronic Laryngitis and 169 subjects with Vocal Cords Paralysis were used for classification in this work. The parameters used in this work are relative jitter, relative shimmer, HNR, and Autocorrelation [17].The groups of pathologies used are based on the database developed in [17-18], namely: *Control* or healthy subjects, chronic *Laryngitis*, *Dysphonia* and vocal cord *Paralysis*. Each subject has nine observations of each parameter extracted in the 9 speech files: the sustained vowels /a/, /i/ and /u/ in the low, normal and high tones [8] [18].

An independence of female and male gender was considered, as noted in [19], for the parameters studied, i.e. there is the concatenation between data of female and male individuals.

Relative or local jitter is the mean absolute of the difference between consecutive glottal periods, divided by the mean period and expressed as a percentage [7], [20-21]. Pathologic voices are correlated with higher *jitter*.

Relative shimmer is defined as the absolute mean difference between amplitudes of consecutive glottal periods, divided by the mean amplitude, expressed in percentage [20] [21]. Pathologic voices are correlated with higher shimmer.

HNR is the relation between signal and noise, i.e. it consists of the relationship between the periodic component and the non-periodic component, glottal noise, in a speech segment [8] [22]. Pathologic voices are correlated with lower HNR.

Autocorrelation is the cross-correlation of a signal with itself. It is a mathematical tool for determining repeating patterns, such as the presence of a periodic signal camouflaged by noise. The greater the result of an autocorrelation, the greater is the repetition of similar events along the signal [8]. Pathologic voices are correlated with lower Autocorrelation. Relative jitter and shimmer and HNR are expressed by eq. 1, 2 and 3, where T_i and A_i are the length and magnitude of glottal period i, and N the number of glottal periods. H is the periodic component and $1\text{-}H$ the noise component.

$$\text{jitter (relative)} = \frac{1}{N-1}\sum_{i=2}^{N}|T_i - T_{i-1}| \tag{1}$$

$$\text{Shim (relative)} = \frac{\frac{1}{N-1}\sum_{i=2}^{N}|A_i - A_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \times 100 \tag{2}$$

$$HNR(dB) = 10 \times \log_{10}\frac{H}{1-H} \tag{3}$$

3.2. Voice Pathologies Recognition ANN's Architecture

The pathology recognition was made alone for each pathology. Therefore, three Recognition Models were implemented to classify each subject in control (healthy) or pathologic (with the tested pathology).

Two different architecture ANNs were used for the identification and treatment of outliers with the objective of analyze the eventual improvement of performance in the task of voice pathologies recognition with different models.

The first ANN is a MLP with Feed-Forward architecture [9], with 8 nodes in the hidden layer, whereas in the output layer there is only one node. The activation function for the hidden layer is Logistic-Sigmoidal transfer function and the output layer is the Linear transfer function.

The input parameters are relative jitter (*jitter*), relative shimmer (*Shim*) and *HNR*, for the 9 speech files. The input matrix consists of 27 columns x N rows, where, 27 correspond to 9 observations by the 3 parameters and N is the number of subjects. The output layer is composed of a single node that contains the classification as pathological (output=1) or non-pathological (output=0).

Because different initial values of the ANN, depending on the seed, the final score can be different. Therefore, 20 training session of the ANN were performed and the best result was retained.

The subjects were divided for each group of the MLP in training, validation and test sets. These relationships are described in Table 1, the dataset was divided into 70% for training, 15% validation and 15% for test. Validation set is used to early stop training to avoid overfitting, and test set to evaluate the performance.

The second ANN is a recurrent neural network (RNN) architecture, called Long Short Term Memory (LSTM) [10].

The input parameters for the LSMT are relative jitter (*jitter*), relative shimmer (*Shim*) and *Autocorrelation*, organized in 3 columns of parameters and 9xN lines for each individual. In this case, each speech files is considered one instance, therefore there are 9 instance for each subject, and the input has only 3 nodes. The output layer has a softmax function with 2 neurons, with binary classification as 0 or 1. Again, the ANN had 20 sessions of train and the best result were retained.

Unbalanced dataset for this LSTM resulted in an overfitting, therefore the dataset had to be rearranged to be approximately balanced (similar number of subjects in each class), and the over plus subjects of the other class were keep out. The data are divided into approximately 70% for training, 15% validation and 15% test, as in Table 1.

3.3. Dataset Pre-processing - Normalization and Outliers Treatment

The pre-processing consist in identify the outliers and change its value by a limit value determined according to the method used for the outliers identification. The BP and SD methods were used for the dataset of each Classification Model. For each method No normalization, z-score normalization, logarithmic normalization and squared root normalization were experimented.

The method applied for the treatment of outliers is that of filling, after an outlier is identified. Its value is replaced by the limit value, established according to the chosen method. The limit is calculated by the BP methods, where the limit is calculated from the interquartile range or SD, where the limit is calculated by the distance in 3 standard deviations from the mean. The replacement of the outlier by the Lower Limit Value (LLV) occurs for values smaller than the LLV. In addition, the substitution by the Upper Limit Value (ULV) occurs for values greater than the ULV [23].

For new subjects added to the data set, the verification must be done by the threshold value determined with the previous dataset.

Table 1. Dataset for each recognition problem using the MLP [9] and LSTM [10].

Classifier	Recognition Model	Total subjects (100%)	Training	Validation	Test
MLP	Control x Dysphonia	263	185	39	39
	Control x Laryngitis	235	165	35	35
	Control x Paralysis	363	255	54	54
LSTM	Control x Dysphonia	1143	802	156	185
	Control x Laryngitis	727	524	93	110
	Control x Paralysis	3267	2360	416	491

4. Results and Discussion

First, the normalization procedure will be discussed and then the outlier's treatment method will be analyzed.

4.1. Normalization Procedure Comparison

Regardless of the type of used network, whether MLP network or LSTM network, the result obtained is the logarithmic method that has the data set with a distribution closer to normal or Gaussian, considering a generality of data sets.

Thus, Fig.1 is presented as an example containing the dataset of MLP, in which the control and paralysis groups are compared, based on a critical analysis of the relative jitter parameter. The goal is to identify the most appropriate standardization method.

Fig. 1 contains a normal probability chart with a red distribution trend line indicating a normal distribution and the blue dots represent the instances. Logarithmic normalization has an almost linear behavior that resembles the red line. Thus, it is emphasized that the normalization obtained from the logarithm is more adequate because it approximates the data of a normal distribution.

Even if the data are more complete from a normal distribution, after normalization with a logarithmic technique, this fact can not lead to a greater classification capacity of the neural networks. Thus, for comparison purposes, both neural networks were tested for all forms of normalization, and outlier's treatment methods were used, even in the absence of normalization.

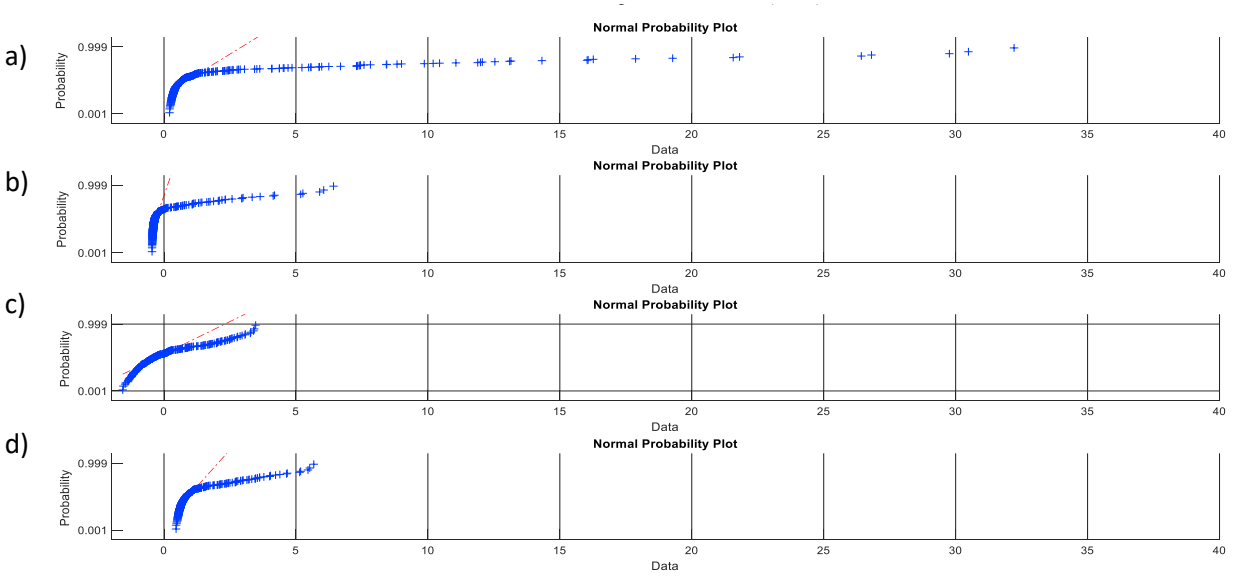

Fig. 1 – Normalization of input data (*jitter*): (a) original; (b) z-score normalization; (c) logarithmic normalization; (d) squared root normalization.

4.2. Outlier's Treatment Comparison

In order to demonstrate the results of the outliers processing techniques and their performance, Fig. 2 shows one example with the *jitter* parameter for the Control/Laryngitis with MLP dataset, with the original values and after the application of the BP and SD methods.

On the original data most of the values seems similar face the very larger values corresponding to outliers. The ANN should face more difficulties in classifying pathologic values, because after the outliers every other values seem at the same level. On the other hand, the outliers treatment with the boxplot method seem that the upper limit is very low leaving low difference between higher and lower values. Finally, the treatment with the SD method seem the most convenient because it is possible to view the difference between high and low values. However, nothing has been concluded about the accuracy of the method yet, it is necessary to take into account the results of the classification.

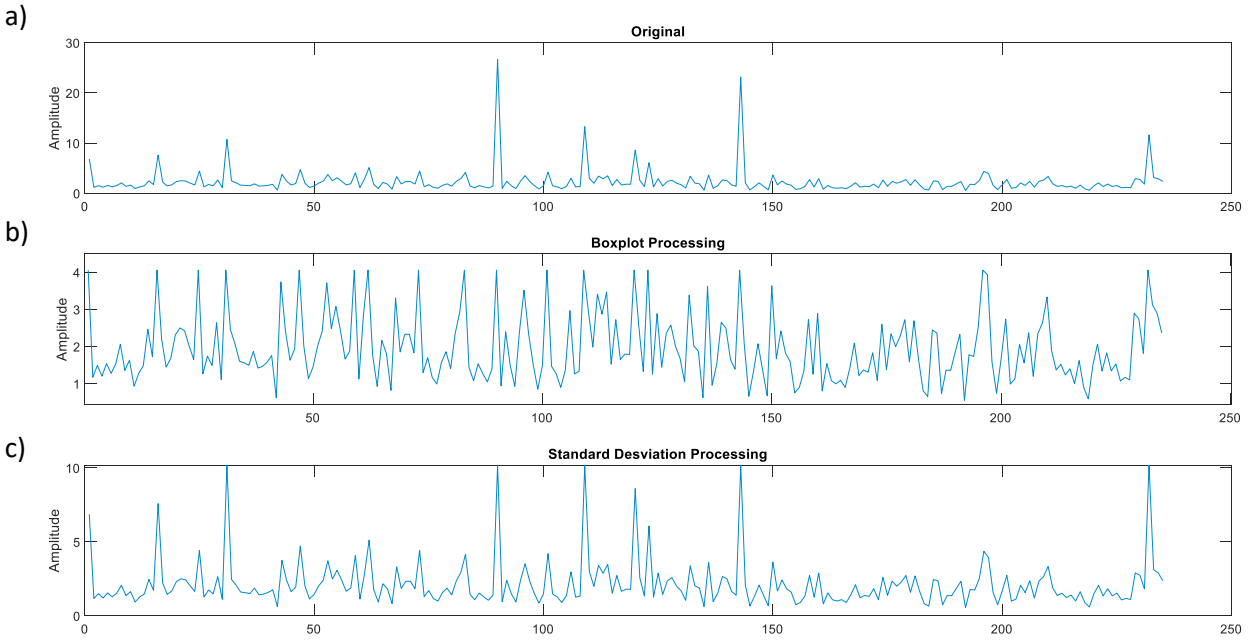

Fig. 2 - *Jitter* feature of the MLP dataset: (a) original; (b) boxplot processing; c) standard deviation processing.

4.3. General Comparison

Table 2 presents the best accuracy in test set after the application of normalization methods combined with the outliers detection and treatment methods, for MLP and LSTM. First column shows the best accuracy without outliers treatment. The method that achieved the higher accuracy for each binary classification is highlighted.

In the case of Control/Dysphonia using the MLP, the accuracy improved from 69% to 80% with the SD1. In the Control/Laryngitis (MLP), the accuracy improved from 80% to 89% also with SD1 method. In the Recognition between Control/Paralysis, using MLP, the accuracy improved from 72% to 78%, again with SD1 method, but also with BP1 method.

For the LSTM ANN, under the Recognition between Control/Dysphonia, the accuracy improved from 59% to 68% with the BP and BP2 methods. The BP1 and SD1 methods achieved a 67% accuracy, very close to the maximum. In the case of Control/Laryngitis using the LSTM classification, the accuracy improved from 63% to 76%, with BP1 processing method. In the case of Control and Paralysis, the improvement was from 68% to 72% with the BP1 method. In this case, the SD1 method achieved 71 %, very close to the maximum accuracy.

Using the MLP ANN, the standard deviation and z-score normalization (SD1) method achieved the best accuracy for vocal pathology identification, but the BP1 achieved also very close accuracy. When the classification is done with the LSTM, the boxplot with z-score normalization (BP1) achieved the best performance, but SD1 also still close.

Concerning the normalization, the z-score performed better than logarithmic and squared root methods. It seem that although the logarithmic normalization method is better to normalize the dataset (see section 4.1), for the classification procedure, the z-score allows better classification accuracy.

Concerning the outliers detection and treatment, both SD and BP method perform with similar accuracy when combined with the z-score.

Table 2. Comparison of ANN accuracy in test set for each pathology identification.

Classifier	Binary Classification	NP (%)	BP (%)	BP1 (%)	BP2 (%)	BP3 (%)	SD (%)	SD1 (%)	SD2 (%)	SD3 (%)
ANN	Control Dysphonia	69	69	77	64	62	74	80	77	72
	Control Laryngitis	80	74	83	80	74	71	89	86	83
	Control Paralysis	72	72	78	74	70	76	78	70	74
LSTM	Control Dysphonia	59	68	67	68	64	66	67	66	61
	Control Laryngitis	63	68	76	62	67	65	68	70	65
	Control Paralysis	68	71	72	69	63	70	71	71	69

NP - No outliers Processing, BP – Boxplot method, BP1 – Boxplot method with z-score normalization, BP2 – Boxplot method with logarithmic normalization, BP3 – Boxplot method with squared root normalization, SD – Standard Deviation method, SD1 – SD with z-score normalization, SD2 – SD with logarithmic normalization, SD3 – SD with squared root normalization.

5. Conclusion

The preprocessing of the dataset for vocal pathologies recognition procedure was analyzed and the improvement in accuracy was experimentally compared. For this, three classification models for Dysphonia, Chronic Laryngitis and Vocal cords Paralysis using two different ANN architectures, namely the MLP and LSTM architectures was used.

The preprocessing consists in the outliers identification and treatment. The Boxplot (BP) and standard deviation (SD) methods were compared. Since these methods can deal better with Gausssian distribution, a normalization procedures using the z-score, the logarithmic and squared root methods were experimented.

Concerning the normalization, although the logarithmic method showed better fitting with a normal curve, the z-score achieved higher accuracy improvement.

Concerning the outliers method both methods, BP and SD showed similar accuracy with the z-score normalization. The SD method improve between 3 and 11 point in percentage the accuracy, meanwhile the BP method improve the accuracy between 3 and 13 points in percentage.

As a final conclusion, the BP or SD method with z-score normalization is very recommendable for pre-processing of the datasets for voice pathology recognition.

Acknowledgements

This work is supported by the Fundação para a Ciência e Tecnologia (FCT) under the project number UID/GES/4752/2019.

References

[1] L. Berton. (2011) "Caracterização de classes e deteccção de outliers em redes complexas". PhD dissertation. *Universidade de São Paulo.* São Carlos, São Paulo.

[2] G. O. Campos. (2015) "Estudo, avaliação e comparação das técnicas de detecção não supervisonadas de outliers". PhD dissertation. *Universidade de São Paulo*. São Carlos, São Paulo.

[3] D.M. Hawkins. (1980) "Identification of outliers". London, Chapman and Hall (Vol. 11).

[4] F. E. Grubbs. (1969) "Detecting outlying observations in samples". *Technometrics*, *11*(1): 1-21.

[5] S. Seo. (2006) "A review and comparision of methods for detecting outliers in universe data sets" PhD dissertation, *University of Pittsburgh*.

[6] H. Cordeiro. (2016), "Reconhecimento de patologias da voz usando técnicas de processamento da fala". PhD dissertation. *Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.* Lisboa.

[7] Teixeira, J. P., Gonçalves, A. (2016), "Algorithm for jitter and shimmer measurement in pathologic voices", *Procedia Computer Science - Elsevier.* 100, Pages 271–279.

[8] Fernandes, J., Teixeira, F., Guedes, V. Junior, A. & Teixeira, J. P. (2018), "Harmonic to Noise Ratio Measurement - Selection of Window and Length", Procedia Computer Science - Elsevier. Volume 138, Pages 280-285.

[9] F. Teixeira. (2019), "Deep Learning with jitter, shimmer and HRN for voice pathologies identification". HCIST.

[10] Guedes, V., Junior, A., Teixeira, F., Fernandes, J., & Teixeira, J. P. (2018), "Long Short Term Memory on Chronic Laryngitis Classification", Procedia Computer Science - Elsevier. Volume 138, Pages 250-257.

[11] Lima, L. F. M., Marson, A., da Silva, D. V. O., Hayashi, C. R. M., & Hayashi, M. C. P. I. (2017), "Métricas científicas em estudos bibliométricos: detecção de outliers para dados univariados". In Questão, 23(5), 254-273.

[12] J. W. Turkey. (1970), "Exploratory data analysis". *Addison-Wesley Publishing Company*.

[13] D. C. Howell. (1988), "Statistical methods for psychology". Belmont, *CA: Wadsworth*. (4th ed.)

[14] F. A. Pino. (2014), "A questão da não-normalidade: uma revisão" São Paulo: Revista de Economia Agrícola. 61(2), 17-33.

[15] MathWorks, "Normalize: Z-score," [Online]. Available: https://fr.mathworks.com/help/matlab/ref/double.normalize.html?searchHighlight=normalize&s_tid=doc_srchtitle.. [Acessed on 29 January 2019].

[16] W. J. Barry e M. Putzer, "Saarbrucken voice database". *Institute of phonetics. University of Saarland*. [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de/. [Accessed on 20 September 2018].

[17] Fernandes, J., Cena, L., Teixeira, F., Guedes, V., Santos, J., Teixeira, J. P, (2019), "Parameters for Vocal Acoustic Analysis - Cured Database", Procedia Computer Science – Elsevier.

[18] Fernandes, J., Teixeira, F., Fernandes, P. & Teixeira, J. P. (2018), "Cured Database of Sustained Speech Parameters for Chronic Laryngitis Pathology", In Proceedings of 31st International Business Information Management (IBIMA) Conference. Milan.

[19] Teixeira, J. P. & Fernandes, P. O. (2014), "Jitter, Shimmer and HNR classification within gender, tones and vowels in healthy voices". *Procedia Technology - Elsevier*, Vol. 16, Pages 1228-1237.

[20] Teixeira, J. P. & Fernandes, P. O. (2015), "Acoustic Analysis of Vocal Dysphonia". Procedia Computer Science - Elsevier 64, 466–473.

[21] Teixeira, J. P., Gonçalves, A. (2014), "Accuracy of Jitter and Shimmer Measurements". Procedia Technology - Elsevier, Vol. 16, Pages 1190-1199.

[22] P. Boersma. (1993), "Accurate short-term analysis of the fundamental frequency and the harmonic-to-noide ratio of a sample sound". *In Proceedings of the institute of phonetic sciences* (Vol. 17, No. 1193, pp. 97-110).

[23] Mathworks, "Filloutlier," [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/32849-htk-mfcc-matlab. [Acessed on 17 November 2019].