Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 138 (2018) 250–257

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 2018

Long Short Term Memory on Chronic Laryngitis Classification

Vitor Guedes[a,b], Arnaldo Junior[b], Joana Fernandes[a], Felipe Teixeira[a], João Paulo Teixeira[a,c] *

[a]Instituto Politécnico de Bragança, Bragança 5300, Portugal
[b]Universidade Tecnológica Federal do Paraná, Câmpus Medianeira, Brasil
[c]Research Centre in Digitalization and Intelligent Robotics (CEDRI), Applied Management Research Unit (UNIAG), Instituto Politécnico de Bragança (IPB), Bragança 5300, Portugal

Abstract

The classification study with the use of machine learning concepts has been applied for years, and one of the aspects in which this can be applied is for the analysis of speech acoustics applied to the analysis of pathologies. Among the pathologies present, one of them is chronic laryngitis. Thus, this article aims to present the results for a classification of chronic laryngitis with the use of Long Short Term Memory as a classifier. The parameters of relative jitter, relative shimmer and autocorrelation was used as input of the LSTM. A dataset of about 1500 instances were used to train, validate and test along 4 experiments with LSTM and one feedforward Artificial Neural Network (ANN). The results of the LSTM overcome the ones of the feedforward ANN, and was about 100% accuracy, sensitivity and specificity in test set, denoting a promising future for this classification tool in the voice pathologies diagnose.

Keywords: Machine Learning, Vocal acoustic analisys, voice pathologies diagnose, LSTM, ANN.

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

10.1016/j.procs.2018.10.036

1. Introduction

Communication is one of the most important actions created and carried out by individuals that governs the whole process of growth and interaction in the social environment, and as we well know, it can be realized by numerous channels, among them we can cite as an example the visual communication, gestural, through words and voice.

The human voice is created in the larynx where there is the presence of the vocal cords that requires an air flow of the lungs to enter into vibration, which later is filtered and amplified in the vocal tract to generate the speech [1]. Sometimes these processes can be changed due to numerous diseases (pathologies) that the larynx can get, one of them is laryngitis.

Laryngitis is inflammation of the larynx that causes the person to have a hoarse voice and can last normally around two weeks (acute) or even in extreme cases with more weeks or months, which is already characterized as chronic laryngitis [2].

This article aims to classify the pathology of chronic laryngitis through machine learning studies, but precisely with the algorithms of Classical Neural Networks and Recurrent Neural Networks with Long Short Term Memory, using as parameters some acoustic signal properties.

Nomenclature

LSTM	Long Short Term Memory
RNN	Recurrent Neural Network
ANN	Artificial Neural Network

2. Long Short Term Memory

The Long Short Term Memory (LSTM), presented by Hochreiter and Schmidhuber in 1997 [3], aims to solve the vanishing gradient problem of the Backpropagation through time algorithm for Recurrent Neural Networks. Basically, these network topologies were not enough to work with a large variation over time for the inputs, that is, the power to remember inputs previously presented to the network and to use them as memory for future training interactions. Figure 1 illustrates the problem.

Some information relevant for the future is extracted by the RNN at Event 1 (dennoted by the black circle). As time passes by, and new events are processed, the information vanishes. This is problematic when the information should be used in the distant future.

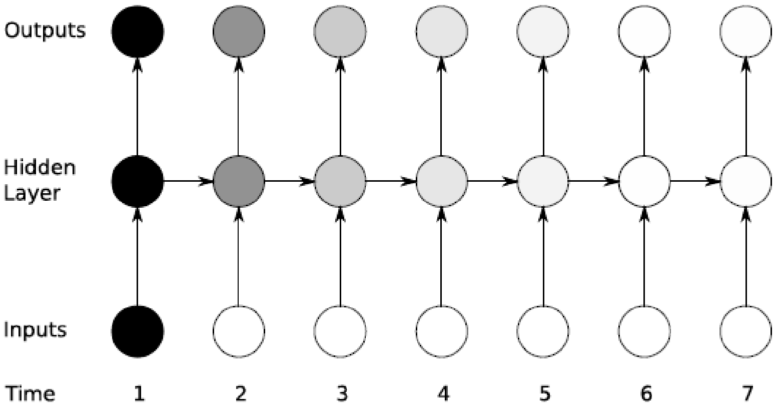


Fig. 1. Vanishing gradient problem (retrieved from [4]).

Thus, Long Short Term Memory is defined with a Recurrent Neural Network topology that presents memory block to work with data that varies over time, each block contains one or more memory cells protected by three multiplier units, input gate, forget gate and output gate, each with its specificity [4]. Figure 2 illustrates the operation of an LSTM block.

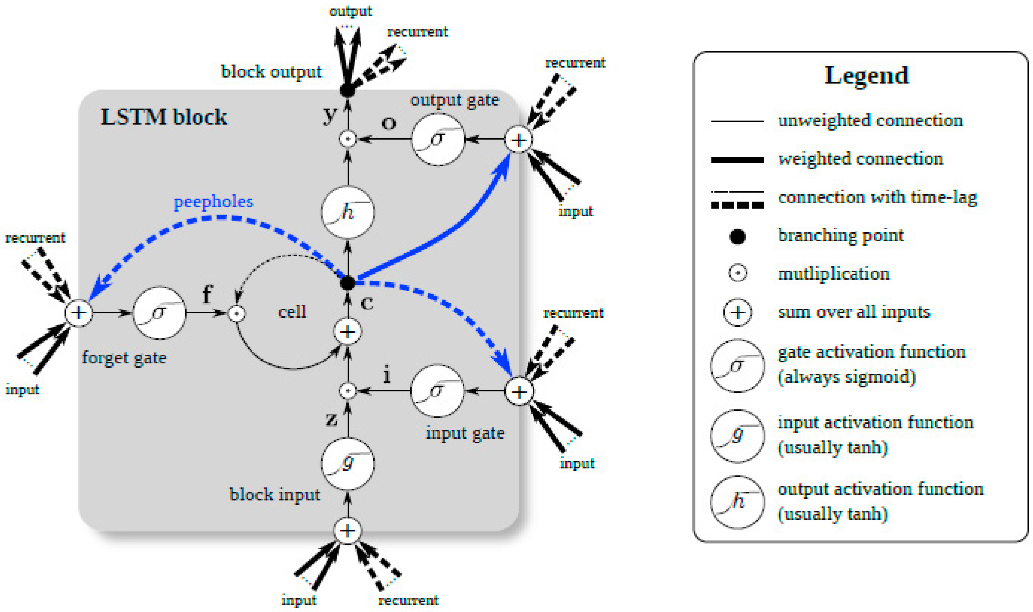


Fig. 2. LSTM Block (retrieved from [5]).

Before explaining the operation of each of the gates it is necessary to present the notations that will be used in the following equations, as can be seen in the Equation 1 [5].

$$\begin{aligned}
&InputWeights : W_z, W_i, W_f, W_ô \in ¡^{NxM} \\
&\mathrm{Re}\,currentWeights : R_z, R_i, R_f, R_o \in ¡^{NxM} \\
&PeepholeWeights : p_i, p_f, p_o \in ¡^{N} \\
&BiasWeights : b_z, b_i, b_f, b_o \in ¡^{N}
\end{aligned} \tag{1}$$

Where W_z, W_i, W_f and W_o are the weights matrix of the blocks input, input gate, forget gate and output gate respectively, the R_z, R_i, R_f and R_o are the recurrent weights matrix of the blocks input, input gate, forget gate and output gate respectively, the p_x are the Peepholeweights and b_x the bias weights.

As seen in figure 2 the gateway first protects the memory cell with the sum of the input at the current time and the previous time inputs, then a sigmoid function is assigned to establish how much information will have input, the nearest of zero is that if the information were null, and closer to one the total information is passed. Subsequently a multiplication is performed with the generated values of the hyperbolic tangent [5]. It is also assigned a Peephole to look at the current state of the cell, this serves for the network to improve learning in problem that require a precise time [4]. Equation 2 represents the forward flow of the input gate.

$$\begin{aligned}
\bar{Z}^t &= W_Z x^t + R_2 y^{t-1} + b_z \\
Z^t &= g(\bar{Z}^t) \\
\bar{i}^t &= W_i x^t + R_i y^{t-1} + p_i \odot c^{t-1} + b_i \\
i^t &= \sigma(\bar{i}^t)
\end{aligned} \tag{2}$$

Where Z^t is the output of the block input at instant t, x^t is the input at instant t, y^{t-1}, is the output of the recurrent layer at instant t-1, g is the activation function of the block input (hyperbolic tangent), i^t is the output of the input gate at instant t, c^{t-1} is the value of cell at the instant t-1 ,σ is the activations function of the input gate (sigmoid).

In relation to the forget gate (represented by Equation 3) it is attributed how much the information present in the memory cell will be forgotten, given the current inputs and the recurrences.

$$\begin{aligned}
\bar{f}^t &= W_f x^t + R_f y^{t-1} + p_f \odot c^{t-1} + b_f \\
f^t &= \sigma(\bar{f}^t)
\end{aligned} \tag{3}$$

Where f^t is the output of the forget gate at instant t.

Already the calculation for cell memory is given by the interactions with the gateways of entry and forgetting plus the state value of the cell in the previous time. This brings you to the exit gate, which takes into account the current and recurring entries, the peephole, and the value of the current cell. This result is passed by a sigmoid function resulting next to the current value of the cell at the output of the LSTM block [5]. Equation 4 represents the case.

$$\begin{aligned}
c^t &= Z^t \odot i^t + c^{t-1} \odot f^t \\
\bar{o}^t &= W_o x^t + R_o y^{t-1} + p_o \odot c^t + b_o \\
o^t &= \sigma(\bar{o}^t) \\
y^t &= h(c^t) \odot o^t
\end{aligned} \tag{4}$$

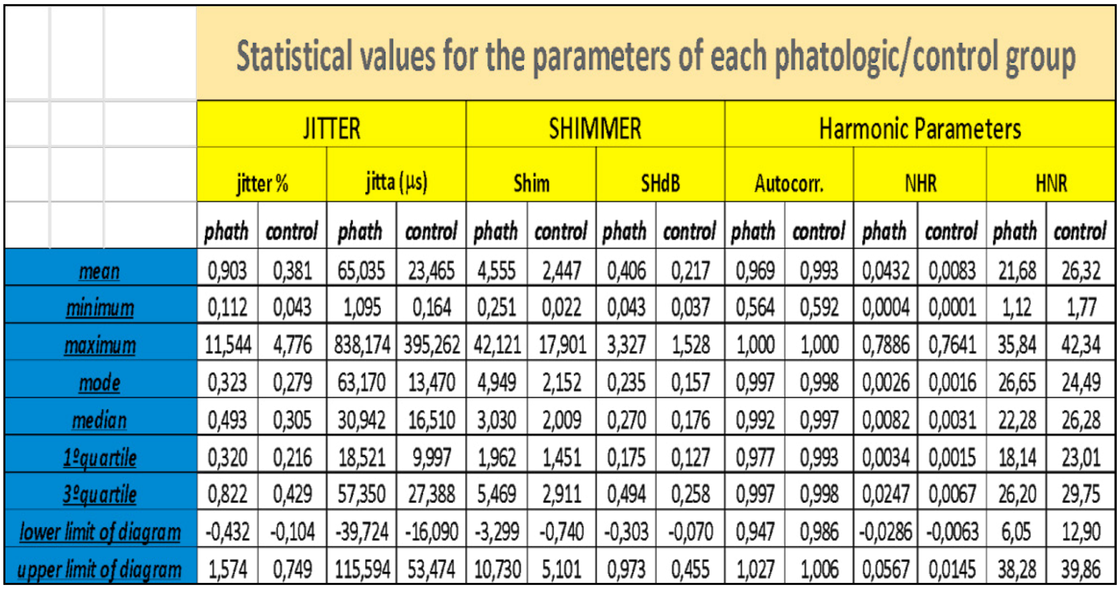
Where o^t is the output of the output gate at instant t, h is the activation function of the cell output (hyperbolic tangent) and y^t the output of the cell.

All these calculations are part of the Forward process of the LSTM network, it is also possible to obtain the Backpropagation Through Time equations and the adjustment of the weights and other elements described in [3] [4] [5].

3. Development

The parameters Jitter, Shimmer and Autocorrelation were used. They were extracted from the database developed by Joana Fernandes et.al [6]. These parameters belong to the control classes (healthy people) and the class of chronic laryngitis (sick people), and are divided by the extraction of the vowels / a /, / i / and / u / into three different tones, low, normal and high, that is, nine instances for each person. Figure 3 shows some statistic of the database.

The database developed by Joana Fernandes et al. is based on the Saarbrücken Voice Database (SVD) developed by Barry and Pützer [7]. They are sustained speech audios of the previously mentioned vowels with sampled signal at 50 kHz and 16 bits resolution. The audios are also divided into the male and female gender, however such separation can be ruled out [8] for the training of the networks.



Statistical values for the parameters of each phatologic/control group

	JITTER				SHIMMER				Harmonic Parameters					
	jitter %		jitta (µs)		Shim		SHdB		Autocorr.		NHR		HNR	
	phath	control	phath	control	phath	control	phath	control	phath	control	phath	control	phath	control
mean	0,903	0,381	65,035	23,465	4,555	2,447	0,406	0,217	0,969	0,993	0,0432	0,0083	21,68	26,32
minimum	0,112	0,043	1,095	0,164	0,251	0,022	0,043	0,037	0,564	0,592	0,0004	0,0001	1,12	1,77
maximum	11,544	4,776	838,174	395,262	42,121	17,901	3,327	1,528	1,000	1,000	0,7886	0,7641	35,84	42,34
mode	0,323	0,279	63,170	13,470	4,949	2,152	0,235	0,157	0,997	0,998	0,0026	0,0016	26,65	24,49
median	0,493	0,305	30,942	16,510	3,030	2,009	0,270	0,176	0,992	0,997	0,0082	0,0031	22,28	26,28
1ºquartile	0,320	0,216	18,521	9,997	1,962	1,451	0,175	0,127	0,977	0,993	0,0034	0,0015	18,14	23,01
3ºquartile	0,822	0,429	57,350	27,388	5,469	2,911	0,494	0,258	0,997	0,998	0,0247	0,0067	26,20	29,75
lower limit of diagram	-0,432	-0,104	-39,724	-16,090	-3,299	-0,740	-0,303	-0,070	0,947	0,986	-0,0286	-0,0063	6,05	12,90
upper limit of diagram	1,574	0,749	115,594	53,474	10,730	5,101	0,973	0,455	1,027	1,006	0,0567	0,0145	38,28	39,86

Fig. 3 Statistical values for the parameters of each group of pathological and control.

3.1. Parameters Jitter, Shimmer and Autocorrelation

"Jitter is the measure of variation of the glottic period between successive cycles of vocal fold vibration" [9]. People with difficulties in controlling vocal cord vibration tend to have high jitter values. Jitter can be measured in four types, relative Jitter, absolute jitter, relative average perturbation (rap), and five-point Period Perturbation Quotient (ppq5), however only the first two were annotated in the database [6]. Relative jitter was used as one of the LSTM network inputs because in gender independent [10].

The Shimmer is related to the "magnitude variation along the glottic periods" [9]. Reductions in good glottis resistance as lesions can cause variations in such magnitudes. Like Jitter, the Shimmer can also be measured in four types, but in the database only two of these, the relative shimmer (Shim) and the absolute shimmer (ShdB) were registered. It was also used the relative as input of the network.

For the harmonic parameter, the autocorrelation parameter is used to calculate the similarity of repeated periods along the signal. The higher the value of autocorrelation, the greater the repetition in the signal.

3.2. Implementation

As mentioned previously, three input parameters were used in the LSTM network, so a unit sequence relation is assigned. For the lstm network modeling, a fully connected input layer with three neurons and an input shape (3.1) is assigned, then two intermediate LSTM layers with three neurons and a sigmoid activation function are added, followed by a Batch layer Normalization to adjust and scale the data, thus accelerating the training process [11]. At the end a layer fully connected with two neurons representing the control and chronic laryngitis classes and sigmoid activation function. For the calculation of the error, it is categorical and use cross entropy. The optimization was done with stochastic gradient descent. A classical neural network with the same hyper-parameters and size is also implemented, however the intermediate layers are changed to two fully connected layers. The code implementation was developed in Python 3.6, supported by the Keras, Scikit-learn and Tensorflow libraries as Backend

4. Results and discussion

Given the database, these were organized in the experiments according to Table 1 whose total of instances is of 1454 and each row of the matrix is considered a new instance. Five experiments were performed, four with recurrent LSTM with different division of the data between training, validation and test sets, and one with classical ANN. A classical neural network and a recurrent LSTM network were implemented because we considered a simple classification problem, assuming that the classical network would not have many problems.

Table 1. Data organization.

Experiments	**Train**		**Validation**		**Test**		**x**
x	**nº**	**%**	**nº**	**%**	**nº**	**%**	**Total**
LSTM 1	1084	74,5	270	18,6	100	6,9	100
LSTM 2	818	56,3	273	18,8	363	25,0	100
LSTM 3	1050	72,2	186	12,8	218	15,0	100
LSTM 4	1049	72,1	186	12,8	219	15,0	100
RNA	1084	74,56	270	18,6	100	6,9	100

However, the results in Table 2 show, on the contrary, the classical neural network did not perform as relevant as the other experiments, reaching 85% accuracy on the test set. For the other experiments an accuracy of 100% is a problem because there can always be the suspicion of overfitting. In order to overcome this problem, the LSTM 4 experiment was performed with 72% of data in the training, 12% for validation and 15% for the test, where in each training season a shuffling of the training data is performed, thus reaching a value of 99% accuracy on the test set, vanishing the idea of overfitting. Figure 4 represent the validation graph of the LSTM 4 experiment, where it is possible to visualize the accuracy performance along the training epochs.

Table 2. Result of the experiments.

Experiments	Validation		Test	
x	Accuracy	error	Accuracy	error
LSTM 1	0,988	0,132	1.0	0,131
LSTM 2	0,996	0,024	1.0	0,344
LSTM 3	1.0	0,061	1.0	0,575
LSTM 4	1.0	0,006	0,995	0,117
RNA	0,969	0,505	0,850	0,464

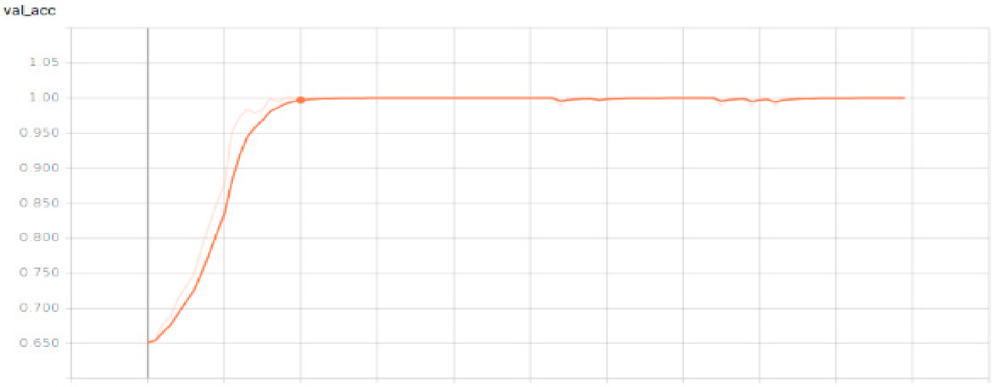


Fig. 4 Validation accuracy during the training process of the Experiment 4.

From the test set of Experiment 4 it was also possible to perform the confusion matrix and determine the sensitivity and specificity according to Equations 5 and 6. In the case of Figure 5, the sensitivity percentage is 99.0%, and the specificity is 100% meaning that all positives are really positives.

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

Where *TP* are the true positives and *FN* the false negatives, *TN* the true negatives and *FP* the false positives.

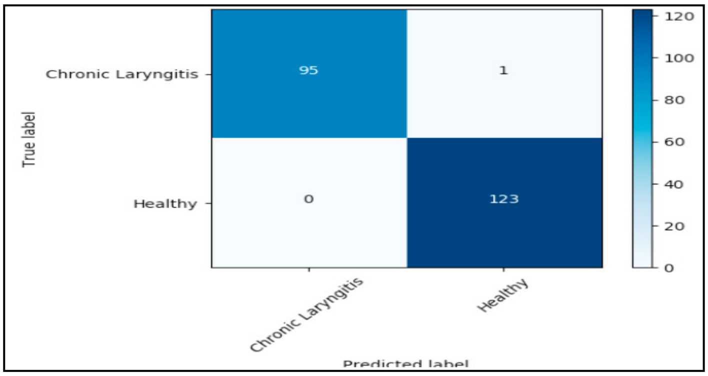


Fig. 5 Confusion matrix.

5. Conclusion and Future Work

The LSTM ANN were implemented for the classification of voice pathologies for the first time. Other machine learning tools like Support Vector Machines and feedforward Artificial Neural Networks were already used in similar works [12] [13] but for other pathologies and several pathologies in the pathologic group.

This experimented used only 3 parameters at the input of the LSTM for each subjects meanwhile other works had used much more parameters.

With the results obtained using only this 3 parameters the implementation of a Long Short Term Memory network can be recommended for this classification problem compared to the classical network. Although preliminary results may already be considered at a very good level, for future work it is interesting to use these LSTM to train and classify and identify more pathologies and make use of other parameters like Mel Frequency Cepstral Coefficient (MFCCs) extracted from continuous speech.

Acknowledgements

This work is supported by the Fundação para a Ciência e Tecnologia (FCT) under the project number UID/GES/4752/2016.

References

[1] Brockmann-Bauser, M.(2011) “Improving jitter and shimmer measurements in normal voices.” Institute of Cellular Medicine, Medical School, Newcastle University.

[2] Dworkin-Valenti JP, Sugihara E, Stern N, Naumann I, Bathula S, Amjad E. (2015) “Laryngeal inflammation.” Ann Otol Rhinol (2): 1058-66.

[3] Hochreiter, Sepp, and Jürgen Schmidhuber. (1997) “Long short-term memory.” *Neural computation* 9.8: 1735-1780.

[4] Alex Graves. (2012) “Supervised Sequence Labelling with Recurrent Neural Networks. Studies in Computational Intelligence.” Springer

[5] Greff, Klaus, Srivastava, Rupesh Kumar, Koutník, Jan, Steunebrink, BasR, and Schmidhuber, Jurgen. (2015) “Lstm: A search space odyssey”.

[6] Joana Fernandes, Filipe Teixeira, Paula Odete Fernandes, João Paulo Teixeira. (2018) “Cured Database of Sustained Speech Parameters for Chronic Laryngitis Pathology”. Proceedings of 31st IBIMA Conference, Milan.

[7] Barry, W.J., Pützer, M. Saarbrücken Voice Database, Institute of Phonetics, Univ. of Saarland, http://www.stimmdatenbank.coli.unisaarland.de/

[8] Teixeira, João Paulo, Joana Fernandes, Filipe Teixeira, e Paula Odete Fernandes. (2018) “Acoustic Analysis of Chronic Laryngitis - Statistical Analysis of Sustained Speech Parameters”, In 168-175, Funchal, Madeira, Portugal.

[9] Teixeira, J. P., Gonçalves, A. (2014) “Accuracy of Jitter and Shimmer Measurements. Procedia Technology.” Elsevier, Volume 16, 1190-1199.

[10] J. P. Teixeira and P. O. Fernandes. (2014) "Jitter, Shimmer and HNR classification within gender, tones and," in Procedia Technology, 16, 1228-1237.

[11] S. Ioffe and C. Szegedy. (2015) “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” In ICML arXiv:1502.03167

[12] Teixeira, J. P., Fernandes, P. O. & Alves, N. (2017) “Vocal Acoustic Analysis – Classification of Dysphonic Voices with Artificial Neural Networks”, Procedia Computer Science - Elsevier 121, 19–26.

[13] Cordeiro, Hugo T, (2016) Reconhecimento de Patologias da Voz usando Técnicas de Processamento da Fala. PhD thesis at Universidade Nova de Lisboa.