Objective Evaluation of the Pathological Voice Based on Deep Learning Neural Networks in an Algerian hospital environment

Mahraz Kabache and Mhania Guerti

Abstract− In this study, we propose a method based on Recurrent Neural Networks, to objectively evaluate the process of rehabilitation of the pathological voice, in an Algerian clinical environment. We choose Unilateral Laryngeal Paralysis as the pathology of the voice. In this paper, we used a Deep Learning system of pathological voice detection by Long Short Term Memory neural model (LSTM). As the dysphony studied in our work concerns essentially the laryngeal vibration, we choose the acoustic parameters based on the instability of the frequency and the amplitude of the laryngeal vibration: Jitter and Shimmer, Noise parameters and Cepstraux MFCC coefficients (Mel Frequency Cepstral Coefficients). A pathological voice detection rate of 88.65% shows important results brought by the rehabilitation technique adopted in Algerian clinical setting. The exclusive and abusive use of hearing to evaluate the effect of speech rehabilitation in the Algerian hospital environment remains insufficient. It is important to correlate perceptual data with objective methods based on detection and classification methods by introducing relevant acoustic parameters, for an effective and objective management of vocal pathology assessment.

***Keywords*− Voice Pathology, *Unilateral Laryngeal Paralysis,* Deep Learning, LSTM Recurring Neural Networks.**

NOMENCLATURE

LSTM	Long Short Term Memory.
MFCC	Mel Frequency Cepstral Coefficients
GMM	Gaussian Mixture Models.
KNN	K-Nearest Neighbors.
SVM	Support Vector Machines.
RNN	Recurring Neural Networks.
CPP	Cepstral Peak Prominence.
HPR	High-frequency Power Ratio.
HNR	Harmonic to Noise Ratio.
ANN	Artificial Neural Network.

I. INTRODUCTION

Assessing the quality of the voice is an important issue for laryngo-phoniatria in order to validate the relevance and effectiveness of the treatments proposed, whether they are rehabilitation or phono-surgery. In this sense, the ear-based judgment, also known as subjective or perceptive evaluation terminology, is the only method of analysis and evaluation of pathological voice used in Algerian clinical environment [1, 2, 3]. In this method, the rehabilitative speech therapist is the only one in charge of listening to the quality of the voice, which results in an unreliable perceptive evaluation, given that reliable perceptive analysis involving several expert auditors and several listening sessions is ultimately time consuming and human resources consuming, and does not allow regular use in clinical routine [3,4,5].

Manuscript received July 11, 2022; revised December 16, 2022.

M. Kabache and M. Guerti are with Ecole Nationale Polytechnique, Algiers, ALGERIA. (e-mails: mahraz.kabache@g.enp.edu.dz, mhania.guerti@g.enp.edu.dz).

Another major drawback of subjective evaluation is inter- and intra-listener variability in voice perception by a jury of experts. This variability can be influenced by the context, emotional state or attention of the listener [6].

In this work, we will develop a system of automatic detection and evaluation of pathological voice using LSTM-type Recurring Neural Networks. We will use in this system a discriminant acoustic analysis based on pathological acoustic parameters. The objective is to show that the use of RN in the re-education evaluation process with the introduction of pathological indices reflecting the malfunction of vocal strings, in the extraction phase of acoustic vectors, can significantly improve and facilitate voice assessment during rehabilitation.

II. LARAYNGEAL PARALYSIS

Dysphonia is an alteration of the voice resulting in the isolated or combined achievement of the three acoustic parameters of the voice which are the pitch, intensity and timbre. The main causes of dysphonia are functional disorders, organic alterations or neurological affected. Laryngeal immobility is defined as a complete decrease or stop of the abduction and/or abduction movement of the larynx (Figure 1). Depending on their laryngeal topography (uni or bilateral, position rather abduction or adduction), they will expose to a vital risk due to respiratory or swallowing problems and to a functional risk related to the various functions of the larynx: phonation, swallowing and breathing.

Unilateral paralysis accounts for 90% of laryngeal paralysis. They are more common on the left, probably for anatomical reasons (longer path on this side) [2]. The voice of a laryngeal paralysis is blown and hoarsely with a significant air leak causing shortness of breath at the end of sentences and a continuous projected voice impossible.

Digital Object Identifier (DOI): 10.53907/enpesj.v2i2.170

Fig. 1: Pictures representing two larynx, right healthy vocal cords, left unilateral paralysis of the left vocal cord.

III. Pathological Voice Classification Techniques

An automatic evaluation system can discriminate between normal and pathological samples and classify voice pathologies. The process of differentiating between normal and pathological subjects is a two-class problem called pathology detection. On the other hand, identify different types of pathology is a multi-class problem called pathology classification.

Among the methods used in the detection and automatic classification of pathological voice, we cite Artificial Neural Network ANN, models based on Gaussian mixtures adapted from a generic GMM speech model, the K Nearest Neighbor classifier KNN, SVM support vector machines, etc. Table. I summarises research work on pathological speech and voice detection and classification systems, and their methodologies on the acoustic analysis used, the type of classifier and the corpus chosen.

Table. I
MODELS OF CLASSIFIERS

Reference	Selected Features	Classifier Model	Corpus used
[7]	MFCC	GMM	Vowels [a], [i] and [u]
[8]	MFCC	SVM	Speech segment of 1.5 sec
[9]	MFCC, Jitter, Shimmer et HNR	SVM	Vowels [a], [i] and [u], continuous speech
[10]	MFCC	Comparison SVM et GMM	Vowels [a], [i] [u], continuous speech and EGG signals
[11]	MFCC	hybrid model SVM /GMM	Vowels
[12]	HNR, critical band of the energy spectrum	KNN	Vowels
[13]	Voice signal preprocessing	DNN	Vowels [a], [i] [u] and EGG signals

IV. Long Short Term Memory Recurring Networks

Long Short Term Memory Network (LSTM) are the special type of Recurrent Neural Networks capable of learning long term dependencies (figure II).

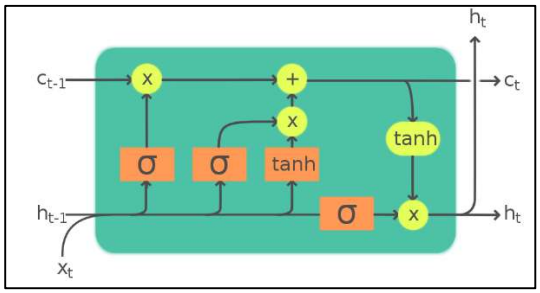

Fig. 2: *Diagram of an LSTM cell [14]*

LSTM cell contains three gates: *forget gate, Input gate* and *Output gate*. Forget gate layer decides what information has to be kept or thrown away from the cell state. It takes input as h_{t-1} and x_t and outputs a number between 0 and 1 using the sigmoid function σ, the output f_t as in the Equation (1). Value of 0 indicates completely remove and 1 to completely keep this.

$$f_t = \sigma\,(W_{hf}\, h_{t-1} + W_{xf} x_t) \quad (1)$$

Now we need to decide what information has to be stored in the cell state. It has two parts, firstly input gate layer using to decide what values has to be updated and then tanh layer generates a vector of new candidate values that has to be added. i_t is the function used by input gate layer and C is the vector of new candidate values by *tanh* layer as shown in the equation (2) and (3).

$$i_t = \sigma\,(W_{hi}\, h_{t-1} + W_{xi} x_t) \quad (2)$$

$$C = tanh\,(W_{hh}\, h_{t-1} + W_{xh} x_t) \quad (3)$$

The updated cell status is indicated by the following equation:

$$C_t = f_t * C_{t-1} + i_t * C \quad (4)$$

Finally, we need to decide what will be the output using output gate. First we run the sigmoid layer using o_t as shown in the Equation (5) and then its output is multiplied by *tanh* to get the output which is shown in the equation (6):

$$o_t = \sigma\,(W_{ho}\, h_{t-1} + W_{xo} x_t) \quad (5)$$

$$h_t = o_t * tanh(C_t) \quad (6)$$

W_{hi}, W_{hf}, W_{hh}, W_{ho} *are the* recurrent connections between the previous hidden layer and current hidden layer. W_{xi}, W_{xh}, W_{xf}, W_{xo} *are* the weights matrix that connects the inputs to the hidden layer. *σ and tanh* are the activation functions.

V. Materials and Methods Applied

A. Selected population

The subjects selected for this works consists of nine Algerian female patients aged 42 to 56 with Unilateral Laryngeal Paralysis (ULP), six are left and three are right. Patients over the age of 56 were eliminated in this study to give reliability to our results. A recording is made after a 9-month speech rehabilitation. In this study, we only selected patients who followed a regular rehabilitation protocol. The same corpus was pronounced by 3 normal female speakers between the ages of 40 and 50 years, not presenting voice disorders (reference standard).

B. Equipment and protocol recording

The voice corpus was recorded with an external M-audio pro sound card, with a Signal/Noise ratio of 100 dB and 16 bits of resolution. We selected a sampling frequency of 44.1 kHz. A dynamic microphone of the Sennheiser e815S type is used for recording with sound software Sound Forge version 10. The voice recordings were made in an acoustically quiet room to eliminate parasitic sound sources. When recording the vocal corpus, a distance of 5 cm is respected between the microphone and the patient's mouth. The microphone is placed at 45° laterally to the mouth, its gain has been adjusted to have an optimal quality of the recording and to avoid the saturation of the sound.

C. Corpus used

The corpus of sound recordings includes the vowel [a] for the various pathological parameters. The chosen corpus consists of 450 normal and pathological voice samples, used for learning, validation and testing. The detection corpus consists of 194 samples between normal and pathological.

D. Multi variable extraction of acoustic parameters

After the input signal preprocessing step a multi-variable acoustic analysis is applied to each frame. In order to have an optimal discrimination of our detection system, we took the Jitter and the Shimmer to evaluate the stability of the frequency and amplitude of the laryngeal vibration F0 of the voice, HNR, HPR, H_1-H_2 and CPP for noise analysis and we used the Mel Frequency Cepstral Coefficients (MFCC) (table. II).

Table. II
ACOUSTIC PARAMETERS USED

Acoustics Parameters		Dimension of the acoustic vector
Cepstral Parameter	MFCC	12
Fundamental frequency stability parameters	Jitter	1
	Shimmer	1
Noise parameters	HNR	1
	HPR	1
	$H_1 - H_2$	1
	CPP	1

E. System Architecture and Learning

The system has 6 layers of neurons: an input layer, an output layer and 4 hidden layers. The number of neurons in the input layer is set to 18. These inputs correspond to the input acoustic vector coefficients for each frame (window). The first hidden layer is limited to 100 neurons, it corresponds to the LSTM cell layer. This is followed by a dropout layer to avoid over-learning or learning by heart (Overfitting). The hidden third layer has two completely connected neurons (Fully Connected), represents the number of classes to ensure non-linearity of network activities. The hidden fourth layer is a softmax activation function with two neurons. The output layer contains a single neuron for decision (Pathological P or Normal N) (Figure III).

The learning rate is set at 0.001, the Dopout (neuron abandonment) is 0.5, the Epoch number is 50 and the batch size is set at 25. As a reminder, in Deep Learning, the Batch size is the number of learning examples in a forward and backward pass through the network. An epoch represents a pass back and forth only once of all learning data.

A. Confusion Matrix and Performance Evaluation

Pathological voice classification performance is represented by a two-dimensional array called the Confusion Matrix. Real voices are arranged in rows and predicted voices in columns (Table. III) [16, 17].

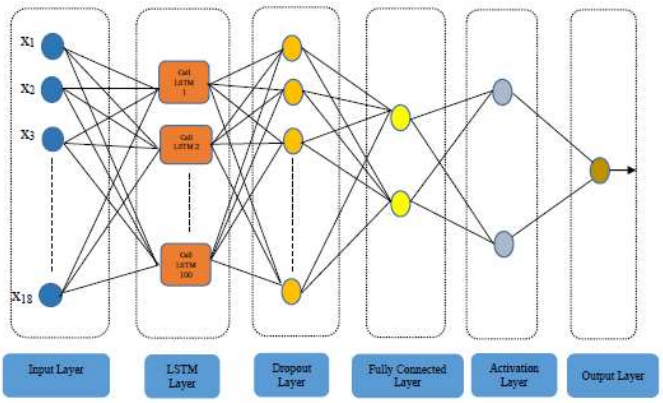

Fig. 3: Pathological Voice Detection System Architecture

B. Confusion Matrix and Performance Evaluation

Pathological voice classification performance is represented by a two-dimensional array called the Confusion Matrix. Real voices are arranged in rows and predicted voices in columns (Table. III) [16, 17].

Table. III
CONFUSION MATRIX

		Voice Detected		Total
		Normal Voice	Pathological Voice	
Real Voice	Normal Voice P	True Positive TP	False Negative FN	P= TP+FN
	Pathological Voice N	False Positive FP	True Negative TN	N= FP+TN

If a voice is positive (P) and is detected as positive, that is a positive voice correctly detected, it is counted as a True Positive (TP). If it is detected as negative, then it is considered a False Negative (FN). If a voice is negative and is detected as negative, it is considered as True Negative (TN), if it is detected as positive, so it is considered as False Positive (FP).

In order to measure the performance of a voice pathology detector or of the classification of the type of pathology, three main indices are taken into consideration: Accuracy, Sensitivity and Specificity [16, 17].

F.1. Accuracy

Is one of the measures commonly used for detection and classification performance. It is defined as a ratio between correctly detected voices and the total number of voices.

$$AC = \frac{TP+VN}{TP+FP+VN+FN} * 100 \quad (8)$$

F.2. Sensitivity

Represents the True Positive Rate (TPR), it is the ability of a classifier to detect positive samples correctly classified in relation to the total number of positive samples. It is estimated by the following equation:

$$TPR = \frac{TP}{TP+FN} * 100 \quad (9)$$

F.3. Specificity

It concerns negative or pathological samples, it represents the True Negative Rate (TNR), it is the ability of a classifier to detect negative samples correctly classified in relation to the total number of negative samples.

$$TNR = \frac{TN}{TN+} * 100 \quad (10)$$

VI. OBTAINED RESULTS AND DISCUSSION

Figure 4 shows the training of the network, it shows the evolution of the Detection Rate (Accuracy) and the Cost (Loss) for the validation data according to the number of Epoch (or iteration number). We tested several values for this parameter, 50 Epoch is enough for a good convergence of the network and that no drop in performance appears beyond.

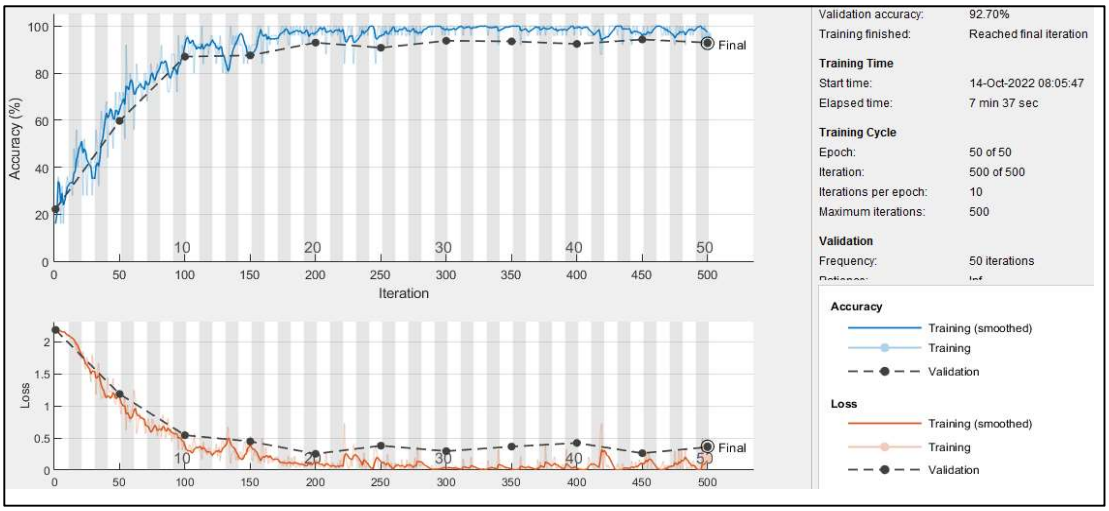

Fig. 4: Evolution of the Accuracy Rate during learning depending on the number of Epoch

Tables IV and V illustrate the confusion matrices obtained by our detection system, for RUP before and after rehabilitation, indicating the overall detection rate represented by the accuracy as well as the sensitivity and specificity of the system.

We noticed a total detection (specificity) of pathological voice (100%) before rehabilitation with a high rate of system accuracy (95.87 %). After rehabilitation, we observed a confusion between normal and pathological voices resulting in a specificity rate of 83.33 % that decreased the accuracy of the system (88.65 %). This low rate of specificity is explained by a difference in the values of pathological parameters used in the values of pathological parameters used in acoustic analysis between normal and pathological voices after rehabilitation.

The sensitivity of our detection system is considered very high (92.72 %) given the inter-speaker and intra-speaker variability factor of the corpus that can cause performance difficulties for automatic speech recognition systems in general.
The significant difference between the specificity and sensitivity of the detection system is explained by the differences between the reference (normal) an pathological voices after rehabilitation of the different acoustic parameters.

Table. IV
CONFUSION MATRIX AND DETECTION RATE OBTAINED BEFORE REHABILITATION

		Voice Detected		Total	Accuracy AC (%)	Sensitivity TPR (%)	Specificity TNR (%)
		Normal Voice	Pathological Voice				
Real voice	Normal Voice P	**102**	8	110	**95.87**	**92.72**	**100**
	Pathological Voice N	00	**84**	84			

Table. V
CONFUSION MATRIX AND DETECTION RATE OBTAINED AFTER REHABILITATION

		Voice Detected		Total	Accuracy AC (%)	Sensitivity TPR (%)	Specificity TNR (%)
		Normal Voice	Pathological Voice				
Real voice	Normal Voice P	**102**	8	110	**88.65**	**92.72**	**83.33**
	Pathological Voice N	14	**70**	84			

VII. CONCLUSION

The LSTM ANN was implemented for pathological voice detection for the first time. Other machine learning tools such as vector support machines and artificial neural networks were already used in similar work [6] but for other pathologies. The application of LSTM recurrent neural networks, in an automatic classification (detection) system of pathological voice, allowed us to have appreciable results. The advantage of these Networks is that the input values transmitted to the network not only pass through several LSTM layers, but also propagate over time in an LSTM cell in order to avoid problems related to a long-term dependence. The problem of confusion between normal and pathological voices in the detection phase makes the use of neural networks alone, as an objective evaluation method in the speech therapy process ineffective, but it can help the speech pathologist in the detection and evaluation of pathological voice with other methods such as objective evaluation through acoustic and subjective analysis through listening.

REFERENCES

[1] M Kabache and M. Guerti, "Multi parametric method for the objective Acoustic Evaluation of the Voice Produced by laryngectomy patients", *Instrumentation, mésure et métrologie*, vol. *20, no. 3,* pp. 137-142, 2021, DOI: https://doi.org/10.18280/i2m.200303

[2] M Kabache and M. Guerti, "Acoustic Analysis of Voice Signal of Patients with Unilateral Laryngeal Paralysis a view to objective evaluation after rehabilitation", *Revue de Traitement de Signal*, vol.38, no. 5, pp.1339-1344, 2021, DOI:10.18280/ts.380508

[3] K. Ferrat and M. Guerti, "A study of sounds produced by Algerian esophageal speakers", *African Health Sciences*, vol. 12, no. 4, 2012, doi: 10.4314/ahs.v12i4.9.

[4] Yu. Ping M. Ouaknine, J. Revis, A. Giovanni, "Objective Voice Analysis for Dysphonic Patients: A Multiparametric Protocol Including Acoustic and Aerodynamic Measurements. *Journal of Voice*, vol. 15 no. 4 pp. 529–542. 2001, https://doi.org/10.1016/S0892-1997(01)00053-4.

[5] E. Saltürk, T. Özdemir, Z. Lütfi Kumral, , E. Karabacakoğlu, H. Kumral, G. Yildiz, , Y. Mersinlioğlu, , G. Atar, , G. Berkiten, Y. Yildirim, "Subjective and objective voice evaluation in Sjögren's syndrome", *Logopedics Phoniatrics Vocology*, vol. 42 no. 1, pp. 9-11. 2017, https://doi.org/10.3109/14015439.2015.1116606

[6] J. Kreiman, B. R Gerratti, , G. B. Kempster., A. Erman, G. S. Berke, "Perceptual Evaluation of Voice Quality: Review, Tutorial, and a Framework for Future Research*", Journal of Speech and Hearing Research*, vol. 36, pp. 21-40, 1996, https://doi.org/10.1044/jshr.3601.21

[7] L. Salhi, T. Mourad, A.Cherif, "Voice Disorders Identification Using Multilayer Neural Network". *The International Arab Journal of Information Technology*, vol. 7, no. 2, pp. 177-185, 2010, https://www.researchgate.net/publication/220413929

[8] C.M. Vikram and K. Umarani, "Pathological Voice Analysis To Detect Neurological Disorders Using MFCC & SVM", *International Journal of Advanced Electrical and Electronics Engineering*, vol. 2, no. 4, 2013, DOI:10.1109/EITech.2015.7163000

[9] F. Teixeira, J. Fernandes, V. Guedes, A. Junior, and J. P. Teixeira, "Classification of control/pathologic subjects with support vector machines," *Procedia Computer Science*, vol. 138, pp. 272–279, 2018.

[10] F. Amara, " An Improved GMM-SVM System based on Distance Metric for Voice Pathology Detection", *Applied Mathematics & Information Sciences*, vol.10, no. 3, pp. 1061-1070, 2016, DOI:10.1016/j.procs.2018.10.039

[11] Xiang Wang et al., "Discrimination between pathological and normal voices using GMM-SVM approach," *Journal of Voice*, vol. 25, no. 1, 2011, https://doi.org/10.1016/j.jvoice.2009.08.002

[12] S. Kumara , K. Anantha, U. Niranjan, "Study of Harmonics-to-Noise Ratio and Critical-Band Energy Spectrum of Speech as Acoustic Indicators of Laryngeal and Voice Pathology*", EURASIP Journal on Advances in Signal,* 2007, DOI:10.1155/2007/85286

[13] H. Pavol, B. Jesus, H. Alonso, J. Mekyska_, Z. Galaz_, R. Burget, Z. Smekal, "Voice Pathology Detection Using Deep Learning: a Preliminary Study" arxiv, 1907.05905, pp. 1-4, 2019, https://doi.org/10.48550/arXiv.1907.05905

[14] C. P. Tang, K. L. Chui, Y. K. Yu, Z. Zeng, K. H. Wong, "Music Genre classification using a hierarchical Long Short Term Memory (LSTM) model", *International Workshop on Pattern Recognition IWPR*, 2018, DOI:10.1117/12.2501763

[15] V. Gupta, "Voice Disorder Detection Using Long Short Term Memory (LSTM) Model*", Arxiv,* 1812.01779, *2018,* https://doi.org/10.48550/arXiv.1812.0177

[16] B. Sabir, F. Rouda, Y. Khazri, B. Touri, and M. Moussetad, "*Improved algorithm for pathological and normal voices identification", International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 1, pp. 238-243, DOI:10.11591/ijece.v7i1.pp238-243

[17] A. Tharwat, "Classification assessment methods", *Applied Computing and Informatics*, vol. 17 no. 1, pp. 168-192, 2021, 10.1016/j.aci.2018.08.003

Mahraz Kabache received his Diploma of Engineer on electronics from university of Blida, Algeria in 1999 and the Diploma of Magister in automatic speech processing at the University of Bouzaréah, Algeria in 2006. He is currently a PhD candidate in electronics at the Ecole Nationale Polytechnique, Algiers. His research interests on speech and treatment of the pathological signal.

Mhania Guerti is currently a Full Professor and Research Director in the Department of Electronics at Ecole Nationale Polytechnique of Algiers (Algeria). She received her MSc in 1984, from the ILP Algiers in collaboration with the CNET - Lannion (France). She got her PHD from ICP -INPG (Grenoble - France), in 1993. She is specialised in Speech and Language Processing. Professor M. GUERTI has supervised many students in Master and PHD degree. Her current research interests include the areas of Speech Processing, Audiovisual Systems, Acoustics, Biomedical Engineering and Technology and Medical Image Processing.