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 dysphonia 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

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>Long Short Term Memory.</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel Frequency Cepstral Coefficients</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Models.</td>
</tr>
<tr>
<td>KNN</td>
<td>K-Nearest Neighbors.</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines.</td>
</tr>
<tr>
<td>RNN</td>
<td>Recurring Neural Networks.</td>
</tr>
<tr>
<td>CPP</td>
<td>Cepstral Peak Prominence.</td>
</tr>
<tr>
<td>HPR</td>
<td>High-frequency Power Ratio.</td>
</tr>
<tr>
<td>HNR</td>
<td>Harmonic to Noise Ratio.</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network.</td>
</tr>
</tbody>
</table>

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].

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.

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].

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 adduction 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.

II. LARYNGEAL PARALYSIS

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@enp.edu.dz; mhania.guerti@enp.edu.dz).

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

2716-912X © 2022 Ecole Nationale Polytechnique
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, identifying 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 classification and their methodologies on the acoustic analysis used, the type of classifier and the corpus chosen.

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).

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 $\sigma$, 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. $i$ 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).

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

$$C = \tanh (W_{ch} h_{t-1} + W_{xc} x_t) \quad (3)$$

The updated cell state 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):

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

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

$W_{hi}, W_{xf}, W_{xi}, W_{ho}, W_{xc}, W_{xo}$ are the recurrent connections between the previous hidden layer and current hidden layer. $W_{hi}, W_{xi}, W_{ho}, W_{xo}$ are the weights matrix that connects the inputs to the hidden layer. $\sigma$ and $\tanh$ are the activation functions.

V. MATERIALS AND METHODS APPLIED

A. Selected population

The subjects selected for this work 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, H₁-H₂ and CPP for noise analysis and we used the Mel Frequency Cepstral Coefficients (MFCC) (table. II).

<table>
<thead>
<tr>
<th>Acoustics Parameters</th>
<th>Dimension of the acoustic vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cepstral Parameter</td>
<td>MFCC 12</td>
</tr>
<tr>
<td>Fundamental frequency stability parameters</td>
<td>Jitter 1</td>
</tr>
<tr>
<td>Shimmer</td>
<td></td>
</tr>
<tr>
<td>Noise parameters</td>
<td>HNR 1</td>
</tr>
<tr>
<td>HPR</td>
<td></td>
</tr>
<tr>
<td>H₁-H₂</td>
<td></td>
</tr>
<tr>
<td>CPP</td>
<td></td>
</tr>
</tbody>
</table>

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 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 Dropout (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].

<table>
<thead>
<tr>
<th>Voice Detected</th>
<th>Normal Voice</th>
<th>Pathological Voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Voice</td>
<td>Normal Voice</td>
<td>Pathological Voice</td>
</tr>
<tr>
<td>Voice</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Pathological Voice</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>True Positive</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>False Negative</td>
<td>FN</td>
<td>TP</td>
</tr>
<tr>
<td>True Negative</td>
<td>TN</td>
<td>FP</td>
</tr>
<tr>
<td>False Positive</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

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.

\[
AC = \frac{TP+TN}{TP+FP+FN} \times 100
\]
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:

\[ TPR = \frac{TP}{TP + FN} \times 100 \] (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.

\[ TNR = \frac{TN}{TN + FP} \times 100 \] (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.

![Figure 4: Evolution of the Accuracy Rate during learning depending on the number of Epoch](image)

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 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) and pathological voices after rehabilitation of the different acoustic parameters.

<table>
<thead>
<tr>
<th>Voice Detected</th>
<th>Normal Voice</th>
<th>Pathological Voice</th>
<th>Total</th>
<th>Accuracy AC (%)</th>
<th>Sensitivity TPR (%)</th>
<th>Specificity TNR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real voice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal Voice P</td>
<td>102</td>
<td>8</td>
<td>110</td>
<td>95.87</td>
<td>92.72</td>
<td>100</td>
</tr>
<tr>
<td>Pathological Voice</td>
<td>00</td>
<td>84</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
subjective analysis through listening. Methods such as objective evaluation through acoustic and in the detection and evaluation of pathological voice with other therapy process ineffective, but it can help the speech pathologist networks alone, as an objective evaluation method in the speech pathological voices in the detection phase makes the use of neural dependence. The problem of confusion between normal and LSTM cell in order to avoid problems related to a long-term through several LSTM layers, but also propagate over time in an is that the input values transmitted to the network not only pass through several LSTM layers, and also propagate over time 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.

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


Mhazraz 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.

Mhnia Guerti is currently a Full Professor and Research Director in the Department of Electronics at Ecole Nationale Polytechnique of Algiers (Algeria). She received her MSE in 1984, from the ILP Algiers in collaboration with the CNET - Lannion (France). She got her PhD from ICP -INPG (Grenoble Polytechnique of Algiers) in 1999. 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.