Effective Detection of Spasmodic Dysphonia Voice Disorder using Bidirectional LSTM Model

Nardjes MERZOUGUI[1], Mohamed Cherif AMARA KORBA[2], Fethi AMARA[1]

[1]LASA Laboratory, Electronic Department, Faculty of Technology,
Badji Mokhtar University Annaba, Algeria
[2]LEER Laboratory, Faculty of Science and Technology,
University of Souk Ahras, 41000
Souk Ahras, Algeria

nardjes.merzougui@univ-annaba.dz, amara_korba@univ-soukahras.dz, amafethi@gmail.com

Abstract **-** Spasmodic dysphonia presents as a neurological voice disorder characterized by spasms of the larynx muscles that control the vocal cords. This disorder can affect individuals of all ages. In this study, we will introduce an Automatic Voice Pathology Detection System (AVPDS), which uses an artificial intelligence technique known as bidirectional LSTM (Long Short-Term Memory) based on a recurrent neural network (RNN) architecture and trained to detect spasmodic dysphonia.
The system primarily utilizes Mel Frequency Cepstral coefficients (MFCCs) features, which are relevant acoustic features and reflect the dynamic characteristic of speech and vocal track. To evaluate the efficiency of our system we used the German Saarbrucken Voice Database (SVD). The best performances obtained for an accuracy equals to 96.20%, while sensitivity and specificity are 90.9% and 100%, respectively.

Keywords **-** Spasmodic dysphonia, Saarbrucken Voice Database, Bidirectional LSTM, Mel Frequency Cepstral Coefficients.

I. INTRODUCTION

Dysphonia is a type of phonation disorder with an impairment in the ability to produce normal voice sounds [1]. Patients of all ages and genders may be affected by it, and its lifetime prevalence is 30%. It is more common in patients who frequently use their voice, such as singers, teachers, coaches, and phone operators [2] [3]. Spasmodic dysphonia used in this study, is a neurologic disorder affect the voice and speech. It is difficult to diagnose this pathology; the invasive methods are dangerous to human health. To detect and classify automatically healthy and pathology of human voice, creating a non-invasive system. The automatic classification of voice pathology is tool to help medical person for correct diagnosis. Generally, automatic detection of voice disorder devises in three basic stages: pre-processing, features extraction and classification. Pre-processing is a process of preparing data for the analysis. Features extractions are extracted from audio record, there are four categories: spectral and cepstral measures MFCCs, linear predictive cepstral coefficients (LPCCs), cepstral peak prominence (CPP) and the perceptual linear prediction cepstral coefficients (PLPCCs); jitter and shimmer as perturbation measures; the Hurst exponent, approximate entropy, and sample entropy as complexity measures; and glottal source parameters in the time and frequency domains [4]. The most widely used features are cepstral features, especially MFCCs, which have been demonstrated to perform on par with or even better than many other feature types [5].

After extracting the vocal features, various classification methods are employed to detect voice pathology. Gaussian mixture models (GMM), support vector machines (SVM), artificial neural networks (ANN), deep learning (DL), and other classifiers are some of the various algorithms used for detection. In literature, GMM and SVM were applied to

detect voice pathology by extracting MFCCs, this purpose achieved an accuracy of 96.5% [6]. The identification of spasmodic dysphonia and the extraction of acoustic features for a comparative study of sustained phonation (/a/) and continuous speech signals.

Three classifiers, including the Levenberg Marquardt Back propagation algorithm, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM), were used to classify the normal and spasmodic dysphonic patients. The results of this study showed that the Levenberg BPN network achieved an accuracy of 96.7%, while SVM and KNN performed at 100% accuracy [7]. The sustained phonation of the vowel /a/ low pitch was classified and compared by the others using Decision Tree (DT), SVM, and KNN by extracting 11 features. The DT algorithm produced a better classification of 86.66% [8]. In [9], spasmodic dysphonia and other diseases were identified using a Convolutional Neural Network (CNN) approach. The model's sensitivity was 66%, its specificity was 91%, and its accuracy was 66.9%. MFCCs are extracted as input features to describe the voice in the form of data and have achieved highest accuracy of 92% using CNN algorithm [10].

Anilkumar V and R Venkata Siva Reddy in [11] employed the utilization of Bidirectional Long Short-Term Memory (BiLSTM) in the classification of various pathological voice conditions, including spasmodic dysphonia; with three features extraction: MFCC, constant-Q cepstral coefficients (CQCC) and All-Pole Group Delay Function (APGDF), their proposed system achieved an accuracy of 92.7% using the CQCC feature extraction and the BiLSTM classification model. A smart healthcare system was developed by [12] to detect dysphonia using a combination of CNN and BiLSTM, the extracted features were fused and processed, achieving an accuracy of 95.65%.

In this paper, we suggested an approach based on BiLSTM to investigate the effect of dynamic coefficients on the detection of spasmodic dysphonia. In the experiments, the Saarbrucken voice dataset (SVD) is used. The remaining sections of the paper are structured following: the proposed methodology is described is section II. Moving on to section III, we present in a detail the experimental results and discussions. Finally, we conclude our work.

II. PROPOSED METHODOLOGY

In this paper, we propose a non-invasive system to detect spasmodic dysphonia. Figure 1 presents this proposed, which is consist of three parts. First, the collected signal is preprocessed in wav format; then the features of the preprocessed audio signal are extracted; finally, an algorithm model has been developed for training and testing.

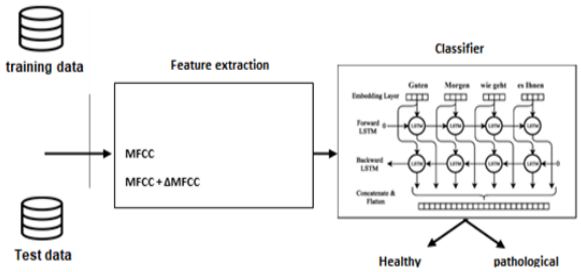

Fig. 1. Block diagram of the AVP

A) Dataset used

In our works, we used Saarbrucken voice disorder (SVD) database, which is freely downloadable [12] maintained by the Institute of phonetics at Saarland University. This database contains sustained vowels /a/, /i/ and /u/ with different intonations (normal, low, high, low-high-low), along with a spoken sentence in German “Guten Morgen, wie geht es Ihnen?” (“Good morning, how are you?”). All recorded SVD voices were sampled with a resolution of 16-bit at 50 kHz.

The voices analyzed were collected from 128 speakers divided into 64 healthy speakers (22 men and 42 women) and 64 patients with spasmodic dysphonia (22 men and 42 women). We download files from the web site [13] and we have used only the sentence “Guten Morgen, wie geht es Ihnen?”, the voices records were saved in “.wav” format.

B) Features extraction

1. Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are a suitable choice for acoustic feature

extraction in voice pathology detection due to their ability to effectively extract meaningful features from vocal signals. Several studies have shown that MFCC coefficients form the basis for numerous methods of voice disorder detection and classification, as evidenced by references [14-18]. They capture important characteristics of speech and vocal tract dynamics.

The extraction process of MFCCs is shown in Fig.2. The process can be described as follows: the voice signal in wav format is input, we apply pre-emphasis, framing, windowing, and FFT to the voice waveform. These operations transform the continuous voice signal into a one-dimensional array with larger values. Following that, a series of Mel-scale triangular band pass filters are used. Subsequently, a logarithmic operation is performed, and the vertical scale is adjusted. Finally, we compute the n-order MFCCs using discrete cosine transform (DCT).

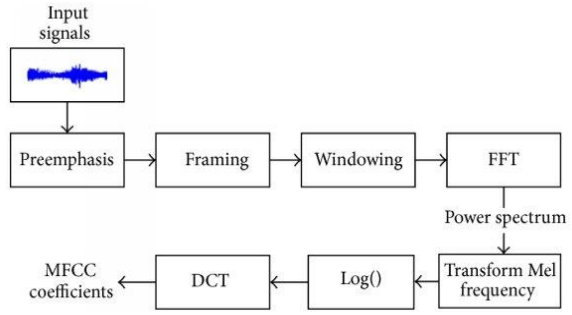

Fig. 2. The extraction process of MFCCs

- Pre-emphasis is operation performed by applying a high-pass filter to the audio waveform. This filter accentuates the change in amplitude between consecutive samples, effectively amplifying the high-frequency content. The most common form of pre-emphasis is achieved using a first-order finite impulse response (FIR) filter with a simple equation:

$$y(t) = x(t) - \alpha \times x(t-1) \quad (1)$$

Where: y(t): the pre-emphasized signal at time t; x(t): the original audio signal at time t; x(t-1): the original audio signal at the previous time step; α: the pre-emphasis coefficient, typically set to a small value like 0.95.

- Framing: the speech signal is divided into small duration blocks of 20-30 ms known as frames. Each frame contains N samples, and adjacent frames are separated by M. Framing is necessary because speech signals change over time, but when we analyze them within these short time intervals, they appear relatively stable. This allows us to perform short-time spectral analysis.

$$frame\ num = \frac{(N - win + inc)}{inc} \quad (2)$$

- Windowing: each frame is multiplied by a window function (e.g., Hamming) to reduce spectral leakage and minimize the effects of discontinuities at the frame boundaries.

$$x_1[i] = x_0 \times w[i] \quad (3)$$

- Fast Fourier Transform (FFT): the windowed frames are passed through an FFT to convert them from the time domain. This results in a series of spectra, one of each frame.

$$X[k] = \Sigma_{n=0}^{N-1} x[n] e^{-j\left(\frac{2\pi}{N}\right)kn} \quad (4)$$

- Mel Filter-bank: the power spectrum obtained from the FFT is then passed through a bank of Mel filters, which are triangular filters spaced on the Mel scale of frequency. These filters emphasize the perceptually relevant frequency bands of the signal.

$$y(m) = \Sigma_{k=1}^{N} H_m(k) |X[k]|^2 \quad (5)$$

- Logarithmic Compression: the logarithm of the filter-bank energies is taken to mimic the logarithmic response of the human auditory system. This step also helps to linearize the energy distribution.
- Discrete Cosine Transform (DCT): the log-filter-bank energies are transformed using the DCT. This decorrelates the features and reduces the dimensionality of the feature vector.

$$s(m) = \log\left(\Sigma_{n=0}^{N-1} |X(k)|^2 H_m(k)\right) \quad (6)$$

$0 \leq m \leq M$

M : the number of filters

- Cepstral Coefficients: the resulting DCT coefficients are known as MFCCs. Typically, only a subset of these coefficients is retrained for further analysis, as higher order coefficients often

contain less relevant information.

$$c(n) = \Sigma_{m=0}^{N-1} s(m) \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right] \quad (7)$$

Where: $n = 1,2,3,4, \ldots L$; L: the order of MFCC.

2. Root Mean Square Energy (rms)

The RMS Energy is a feature that measures the squared values of the mean of the squared values of samples within a segment or frame of a signal. It provides insight into the magnitude or-energy of the signal during that specific time interval. Mathematically is defined:

$$RMS = \sqrt{\left[\frac{(x_1^2 + x_2^2 + \cdots + x_n^2)}{n}\right]} \quad (8)$$

Where : - x_1, $x_2 \ldots x_n$ are the individual values in the samples, n is the total number of samples.

3. Zero Crossing Rate (zcr) -

The zero-crossing rate (ZCR) is used to measure the rate which a signal changes its sign within a given time frame. Within each frame, the number of times the signal crosses the zero-amplitude level (I.e., changes from positive to negative) is counted. The count of zero-crossings is divided by the duration of the frame to calculate the ZCR. It's often expressed as the number of zero-crossing per unit of time, such as zero-crossing per second.

$$ZCR = \frac{number\ of\ zcr}{total\ duration} \quad (9)$$

Where:

$$total\ duration = \frac{N}{Fs} \quad (10)$$

N: is the number of samples, and Fs: is the sample rate.

C) Classifier description

Bidirectional LSTM is a type of recurrent neural network (RNN) architecture used in deep learning and signal processing. It is an extension of the traditional LSTM network, which is designed to address the vanishing long-range dependencies in sequential data. The key characteristic of a BiLSTM network is that it simultaneously processes input sequences in both the forward and backward directions. This bidirectional processing allows the network to consider not only the past context (previous elements in the sequence) but also the future context (subsequent elements in the sequence) when making predictions or capturing dependencies. This is particularly useful when information from both directions is essential for understanding the context. Fig. 3. presents the architecture of BiLSTM used in our work.

Here's how a BiLSTM network works:

- **Forward LSTM**: this component captures dependencies and features in a forward direction by processing the input sequence from begin to end.

- **Backward LSTM**: simultaneously, reversing the order of processing the input sequence, a second LSTM records features and dependencies in the opposite direction.

- **Combination:** the forward and backward LSTMs' outputs are typically combined in some way (e.g., concatenation or addition) to create a unified representation of the input sequence that encodes both forward and backward information.

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

During training and testing, ensure that we are evaluating the model's performance using appropriate metrics such as accuracy, sensitivity, specificity, precision and F1_score to measure the effect of features dynamics on vocal pathology. We used four combinations when the number of coefficients is changes.

- Accuracy is a typical measure used to assess how well a classification model is performing. It indicates the percentage of accurately classified samples among all instances in the dataset, it is calculated using the formula:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (11)$$

- Sensitivity quantifies the percentage of true positives that the model correctly detects, it is calculated using the formula (12). This metric is crucial because it indicates how well the model identifies voices as ill.

$$sensitivity = \frac{TP}{TP + FN} \times 100\% \quad (12)$$

- Specificity indicates the voice as healthy. It is calculated using the formula:

$$specificity = \frac{TN}{TN+FP} \times 100\% \quad (13)$$

- Precision measures the proportion of correctly predicted positive out of all instances predicted as positive, mathematically, it is calculated using the formula:

$$precision = \frac{TP}{TP+FP} \times 100\% \quad (14)$$

- F1_score is a measurement that yields a single value by combining sensitivity and precision, is calculated using the formula:

$$f1\ score = 2 \times \frac{precision \times sensitivity}{precision + sensitivity} \quad (15)$$

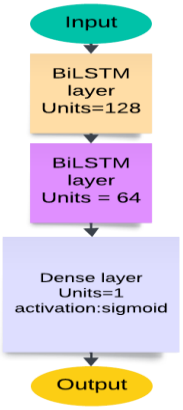

Fig. 3. Bidirectional LSTM model architecture

The influence of MFCC with various coefficients on performance is shown in Table 1, shows that achieved highest accuracy of 92.3%, sensitivity of 90.9%, 100% of precision and specificity, and f1_score of 90.9%.

The result of performance with combining MFCC, RMS, delta-RMS and ZCR is shown in table 2, the best performance is where number of coefficients of MFCC equal 12, 96.2% accuracy, 90.9% sensitivity, 100% precision and specificity, and 95.2% of f1_score.

The effect of the combination of MFCC, delta-MFCC, RMS, delta-RMS and ZCR on performance is presented in table 3. It outperformed in terms of 92.3% of accuracy, 90.9% precision and sensitivity, 93.3% of specificity and 90.9% of f1_score. When we combined MFCC, delta-MFCC, delta-delta-MFCC, RMS, delta-RMS and ZCR, the results of performance are described in table 4.

TABLE 1. PERFORMANCE OF THE (AVPDS) OBTAINED USING MFCC FEATURES

Nbr MFCC coefficients	Accuracy (%)	Sensitivity (%)	Precision (%)	Specificity (%)	F1_score (%)
12	84.60	81.80	81.80	86.70	81.80
14	**92.30**	81.80	**100.0**	**100.0**	90.00
16	**92.30**	**90.90**	90.90	93.30	**90.90**
18	88.50	**90.90**	83.30	86.70	87.00
20	88.50	**90.90**	83.30	86.70	87.00
22	**92.30**	**90.90**	90.90	93.30	**90.90**

TABLE 2. PERFORMANCE OF THE (AVPDS) OBTAINED USING MFCC FEATURES, RMS, Δ RMS AND ZCR

Nbr MFCC coefficients	Accuracy (%)	Sensitivity (%)	Precision (%)	Specificity (%)	F1_score (%)
12	96.20	90.90	100.0	100.0	95.20
14	92.30	90.90	90.90	93.30	90.90
16	92.30	90.90	90.90	93.30	90.90
18	88.50	90.90	88.30	86.70	87.00
20	92.30	90.90	90.90	93.30	90.90
22	84.60	90.90	76.90	80.0	83.30

TABLE 3. PERFORMANCE OF THE (AVPDS) OBTAINED USING MFCC, Δ MFCC, RMS, Δ RMS AND ZCR

Nbr MFCC coefficients	Accuracy (%)	Sensitivity (%)	Precision (%)	Specificity (%)	F1_score (%)
12	88.50	81.80	90.00	93.30	85.70
14	80.80	90.90	71.40	73.30	80.00
16	92.30	90.90	90.90	93.30	90.90
18	88.50	81.80	90.00	93.30	85.70
20	84.60	81.80	81.80	86.70	81.80
22	92.30	90.90	90.90	93.30	90.90

TABLE 4. PERFORMANCE OF THE (AVPDS) OBTAINED USING MFCC, Δ MFCC, ΔΔ MFCC, RMS, Δ RMS AND ZCR

Nbr MFCC coefficients	Accuracy (%)	Sensitivity (%)	Precision (%)	Specificity (%)	F1_score (%)
12	92.30	90.90	90.90	93.30	90.90
14	92.30	90.90	90.90	93.30	90.90
16	92.30	90.90	90.90	93.30	90.90
18	88.50	90.90	83.30	86.70	87.00
20	84.60	90.90	76.90	80.00	83.30
22	92.30	90.90	90.90	93.30	90.90

In summary, Table 2 shows the optimal outcomes for each combination. High accuracy achieved of 96.20%. Balanced sensitivity and specificity are respectively 90.90 % and 100 %.

The results of comparative analysis with related work in Table 5 were truly encouraging as our approach suggest that the model performs well in correctly identifying both positive and negative case.

TABLE 5. ACCURACY RESULTS COMPARING WITH RELATED WORK.

Xiaoping and all [10]	A.V and all [11]	Ghulam and all. [12]	Proposed system
92.00%	92.70%	95.65%	96.20 %

IV. CONCLUSION

This work proposed a non-invasive system to detection voice disorder based on BiLSTM. The proposed AVPDS evaluated using SVD database. We observe that the second-degree dynamic coefficients, namely delta-delta MFCC, RMS, and ZCR, make a substantial contribution to the system's performance. The F1-score consistently surpasses 90% across various numbers of MFCC coefficients. It's worth highlighting that the system's performance begins to decline when the number of MFCC coefficients exceeds 12. This observation underscores the delicate balance between feature dimensionality and performance in our system.

In future work, we will investigate to detection voice pathology using CNN model with others acoustic features to enhance the system's performance.

V. REFERENCES

[1] Benninger MS, Ahuja AS, Gardner G, Grywalski C (1998) Assessing outcomes for dysphonic patients. Journal of Voice 12: 540–550.

[2] Stachler RJ, Francis DO, Schwartz SR, Damask CC, Digoy GP, Krouse HJ, McCoy SJ, Ouellette DR, Patel RR, Reavis CCW, Smith LJ, Smith M, Strode SW, Woo P, Nnacheta LC. Clinical Practice Guideline: Hoarseness (Dysphonia) (Update). Otolaryngol Head Neck Surg. 2018 Mar;158(1_suppl): S1-S42.

[3] Van Houtte E, Van Lierde K, Claeys S. Pathophysiology and treatment of muscle tension dysphonia: a review of the current knowledge. J Voice. 2011 Mar;25(2):202-7.

[4] Madhu Keerthana Yagnavajjula, Paavo Alku, Krothapalli Sreenivasa Rao, Pabitra Mitra. Detection of Neurogenic Voice Disorders Using the Fisher Vector Representation of Cepstral Features. J Voice. 2022 Nov.

[5] J.A. Gómez-García, L. Moro-Velázquez, J.I. Godino-Llorente. On the design of automatic voice condition analysis systems. Part II: review of speaker recognition techniques and study on the effects of different variability factors Biomed Signal Process Control, 48 (2019), pp. 128-143

[6] Fethi Amara, Mohamed Fezari and hocine bourouba. An Improved GMM-SVM System based on Distance Metric for Voice Pathology Detection. Appl. Math. Inf. Sci. 10, No. 3, 1061-1070 (2016).

[7] Snekhalatha Umapathy, Shamila Rachel, Rajalakshmi Thulas. Automated speech signal analysis based on feature extraction and classification of spasmodic dysphonia: a performance comparison of different classifiers.Received: 13 July 2017 / Accepted: 27 September 2017. © Springer Science+Business Media, LLC 2017.

[8] Elmounder Hadjaidji, Mohamed Cherif Amara Korba, Khaled Khelil. Spasmodic dysphonia detection using machine learning classifiers. IEEE Xplore 2012. 978-1-6654-417-1.

[9] Hao-Chun Hu, Shyue-Yih Chang, Chuen-Heng Wang, Kai-Jun Li, Hsiao-Yun Cho, Yi-Ting Chen, Chang-Jung Lu, Tzu-Pei Tsai, Oscar Kuang-Sheng Lee. Deep Learning Application for Vocal Fold Disease Prediction Through Voice Recognition: Preliminary Development Study. J Med Internet Res. 2021 Jun; 23(6): e25247. Published online 2021 Jun 8. doi: 10.2196/25247.

[10] Xiaoping Xie, Hao Cai, Can Li and Fei Ding. A Voice Disease Detection System Based on MFCCs and Single-Layer CNN. IEEE Xplore. Apr 2023.

[11] A. V and R. V. S. Reddy, "Classification of voice pathology using different features and Bi- LSTM," 2023 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES), Tumakuru, India, 2023, pp. 1-4.

[12] Ghulam Muhammad, Musaed Alhussein. Convergence of artificial intelligence and internet of things in smart healthcare: a case study of voice pathology detection. Ieee Access 9, 89198-89209, 2021.

[13] W. J. Barry and M. Pützer. Saarbrucken voice database. Institute of Phonet- ics, University of Saarland. Accessed: Mar. 10, 2017. [Online]. http://www.stimmdatenbank.coli.uni-saarland.de/.

[14] Fang, S. H., Tsao, Y., Hsiao, M. J., Chen, J. Y., Lai, Y. H., Lin, F. C., & Wang, C. T. (2019). Detection of pathological voice using cepstrum vectors: A deep learning approach. Journal of Voice, 33(5), 634-641.

[15] Ali, Z., Alsulaiman, M., Muhammad, G., Elamvazuthi, I., & Mesallam, T. A. (2013, November). Vocal fold disorder detection based on continuous speech by using MFCC and GMM. In 2013 7th IEEE GCC Conference and Exhibition (GCC) (pp. 292-297). IEEE.

[16] Grzywalski, T., Maciaszek, A., Biniakowski, A., Orwat, J., Drgas, S., Piecuch, M., ... & Szarzynski, K. (2018, December). Parameterization of Sequence of MFCCs for DNN-based voice disorder detection. In 2018 IEEE International conference on big data (big data) (pp. 5247-5251). IEEE.

[17] Benba, A., Jilbab, A., Hammouch, A., & Sandabad, S. (2015, March). Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease. In 2015 International conference on electrical and information technologies (ICEIT) (pp. 300-304). IEEE.

[18] Shetty, S., Hegde, S., & Dodderi, T. (2018, February). Classification of healthy and pathological voices using mfcc and ann. In 2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAECC) (pp. 1-5). IEEE.