International Journal of Speech Technology (2024) 27:483–502
https://doi.org/10.1007/s10772-024-10120-w

Pathological voice classification system based on CNN-BiLSTM network using speech enhancement and multi-stream approach

Soumeya Belabbas[1] · Djamel Addou[1] · Sid Ahmed Selouani[2]

Received: 17 November 2023 / Accepted: 21 June 2024 / Published online: 5 July 2024

Abstract
The paper developing a resilient speech classification system for individuals with voice disorders poses a formidable challenge due to the significant variability and distortions inherent in vocal signals. This article outlines the steps to create an effective classification system for pathological speech. The first step involved applying speech enhancement processing using the minimum mean square error (MMSE) enhancer to improve voice input data quality and intelligibility. Secondly, a multi-stream approach combined various acoustic vectors based on human auditory perception, including mel-spectrogram images, mel frequency cepstral coefficients (MFCC), power normalized cepstral coefficients (PNCC), and prosodic parameters like F0, Jitter, and Shimmer. Finally, a deep machine learning incorporating both a convolutional neural network (CNN) and a bidirectional long short-term memory (BiLSTM) network was employed to process these enhanced characteristics in a multi-stream framework, resulting in a powerful classification system architecture. In our experiments, we utilized two subsets from the Massachusetts Eye and Ear Infirmary (MEEI) database, each involving distinct causes of voice disorders. The first subset consisted of voice recordings from patients with vocal nodules, paralysis, and polyps, while the second subset included recordings from patients with mild ventricular compression, A–P squeezing, and gastric reflux. The results we obtained reveal that the CNN-BiLSTM system, coupled with a robust speech analysis interface based on the multi-stream approach and enhanced by the minimum mean square error (MMSE) processing, achieved the highest accuracy rates.

Keywords Convolutional neural network · Bidirectional long short-term memory · Mel-frequency cepstral coefficient · Power normalized cepstral coefficients · Multi-stream approach · Minimum mean square error

1 Introduction

Speaking is a complex process that requires precise coordination and control of various elements such as articulation, breathing, voicing, and prosody (Zhaoyan, 2016). However, any damage to the previous elements or any part of them results in articulation, fluency, or voice disorders (American Speech-Language-Hearing Association, 1993). Speech and voice disorders, although distinct, can sometimes overlap. Speech disorders affect the production of sounds and the sequence of words, while voice disorders concern abnormalities in vocal production. Several categories exist such as neurological, muscular, structural, functional, and psychogenic disorders. Certain disorders, such as dysarthria, can affect both speech and voice. Dysarthria is a motor speech disorder that can affect the clarity, fluency, intelligibility, pitch, volume, quality, or resonance of an individual's voice. It can be classified into different types based on its characteristics and underlying causes. Functional disorders are caused by an inability to use the vocal cord muscles (Chung et al., 2018). Organic disorders may result from structural issues, such as abnormal growths on the larynx, or neurological problems that affect the nerves controlling the larynx (Carding et al., 2016). On the other hand, psychogenic disorders are caused by emotional stress or trauma and may result

✉ Soumeya Belabbas
belabbas_soumeya@hotmail.com

Djamel Addou
daddou@usthb.dz

Sid Ahmed Selouani
sid-ahmed.selouani@umoncton.ca

[1] Laboratory of Speech Communication and Signal Processing (LSCSP), University of Sciences and Technology Houari Boumediene, Algiers, Algeria

[2] Laboratory of Research in Human-System Interaction (LARHSI), University of Moncton, Shippagan Campus, Moncton, Canada

from anxiety, depression, or conversion disorder (Zabret et al., 2018).

Although significant progress has been made in automatic speech classification system technologies over the past few decades, labeling disordered speech remains a challenging task (Duffy, 2019). Disordered speech poses a wide range of difficulties for current data-intensive deep neural networks (DNN) based speech classification system technologies, which mostly target normal speech. These challenges arise due to the variability in speech characteristics, the limited number of available recordings, and the use of traditional acoustic features that may not effectively capture the unique characteristics of pathological speech. These differences make the acoustic classifier components ineffective in correctly mapping pathological speech signals to labels. Developing pathological speech classification systems requires advanced machine-learning techniques, expertise in speech pathology, collaboration with healthcare professionals, and access to diverse and well-annotated datasets. Despite the challenges, progress is being made, and these systems hold immense potential to help individuals with voice disorders by improving their communication and quality of life (Jayaraman & Das, 2023).

1.1 Related work

Pathological speech classification refers to an automatic speech processing system to categorize and label speech from individuals with voice disorders. This area of research and its applications in speech technology holds significant clinical importance, with extensive prior research focused on voice disorder detection addressing the processing of pathological speech in various aspects.

To cope with the high variability of voice disorders, a study in Souissi and Cherif (2015) proposed an effective algorithm for voice pathologies detection, using short-term cepstral parameters, Linear Discriminant Analysis (LDA) for dimensionality reduction, and the Support Vector Machine (SVM) classifier. This study demonstrated that the detection of voice disorders can be efficient using only the original Mel Frequency Cepstral Coefficients (MFCC) ignoring their first and second derivative. Another aspect of improving the voice pathology classification systems consists of combining various machine learning techniques, (Amara et al., 2016) investigated the combined classifier GMM-SVM to distinguish normal voice from pathological speech arising from vocal tract pathologies, utilizing MFCC coefficients and modified Kullback–Leibler and Bhattacharyya distance approaches to enhance results. (Hossain & Muhammad, 2016) performed a fusion of the decisions of an extreme learning machine, Gaussian Mixture Model (GMM) and SVM, which had as input the fusion of MPEG-7 audio parameters and interlaced derivative pattern parameters. In Kadi et al. (2016), a model that simulates various aspects of the ear was introduced to enhance speech identification. This model's characteristics were integrated with MFCC to represent data from the Nemours and Torgo databases. The data was subsequently analyzed using GMM, SVM, and the combined GMM-SVM machine learning architectures.

Investigating the Acoustic Voice Analysis methods (AVA) based on adaptive features is the major goal of the work presented in Emary et al. (2014), the Mel-Frequency Cepstral Coefficients (MFCCs with different variations in frequencies and amplitudes: Jitter and Shimmer), and the flux model mixture (GMM) was used in the AVA. A multivariate analysis of parameters that measure the various problems in the process of phonation is applied to analyze the importance of finding and sorting features that provide more information. In this work, the accuracy of the voice disorder classification system increased with the number of parameters (best accuracy with coefficients including 39 MFCCs, Jitter, and Shimmer), which means that the difference between normal and abnormal voices becomes noticeable using multiple parameters, also, the effect of the number of Gaussian which makes up the model is important, where a sufficient number of mixtures allows to represent data (features) optimally.

Several deep learning models have been explored for pathological speech domains. Deep leaning exploits data driven approaches to learn abstract pathological cues and improves the state-of-the-art performance in different classification tasks remarkably. These models require more complex architectural components [e.g., convolutional neural networks (Narendra et al., 2021; Shakeel et al., 2021, 2023; Vaiciukynas et al., 2018), long short-term memory networks (Mayle et al., 2019), autoencoders (Janbakhshi & Kodrasi, 2022a; Vásquez-Correa et al., 2017), etc.] and more data to be trained. As a result, they often achieve significantly higher performance. Mainstream deep learning-based dysarthric speech detection approaches typically rely on processing the magnitude spectrum of the short-time Fourier transform of input signals, while ignoring the phase spectrum that also contains inherent structures that are not immediately apparent due to phase discontinuity, (Janbakhshi & Kodrasi., 2022b) investigated the applicability of the unprocessed phase and the alternative phase representations such as the modified group delay (MGD) and instantaneous frequency (IF) spectra, it was shown that using phase representations as complementary features to the magnitude spectrum is beneficial for deep learning-based dysarthric speech detection and yielding a high performance. In Joshy and Rajan (2021), authors explored dysarthria severity classification using various deep learning architectural such as DNN, CNN, and LSTM, with MFCCs and their derivatives as features**.** Performances of these models are compared with a baseline support vector machine (SVM) classifier using the UA-Speech corpus and the TORGO database. The analysis of the results showed that a proper choice of a deep learning architecture can ensure better performance than the conventionally used SVM classifier.

To enhance pathological speech classification accuracy, researchers experimented with various methods. One approach involved using deep belief networks (DBN) to extract acoustic parameters, which was shown to be superior to Mel Frequency Cepstral Coefficients in a referenced study (Farhadipour et al., 2018). In another study (Souli et al., 2021), using scatter wavelet features in conjunction with a deep convolutional neural network (DCNN) improved performance for pathological classification.

In Chaiani et al. (2022), authors developed a voice disorder classification system employing a two-stage framework. The first stage incorporated a speech enhancement technique based on the minimum mean square error (MMSE). In the second stage, a CNN-LSTM network with the SinRU activation function was implemented. This system significantly improved the automatic classification of investigated voice disorders and the assessment of dysarthria severity levels, surpassing the performance of SVM, GMM, and GMM-SVM models presented in Kadi et al. (2016). In recent work (Mohammed et al., 2023), a novel method named the deep Multi-Modal and Multi-Layer Hybrid Fusion Network (MMHFNet) was introduced for extracting deep features. The authors conducted experiments using a deep learning algorithm based on the LSTM model, employing the Saarbruecken Voice Database SVD with both complete and balanced samples.

The classification of voice disorders based on a machine learning algorithm requires a large number of samples data for the training step. However, due to the sensitivity and particularity of medical data, it is difficult to obtain sufficient samples for model learning. To address this challenge, authors in Peng et al. (2023) proposed a pretrained OpenL3-SVM transfer learning framework for the automatic recognition of multi-class voice disorders. The framework combines a pre-trained convolutional neural network, OpenL3, and a support vector machine (SVM) classifier. The first step was to extract the Mel spectrum of the given voice signal and then input it into the OpenL3 network to obtain high-level feature embedding. Considering the effects of redundant and negative high-dimensional features, model overfitting easily occurred. Therefore, linear local tangent space alignment (LLTSA) was used for feature dimension reduction. The obtained dimensionality reduction features were used to train the SVM for voice disorder classification. Fivefold cross-validation was used to verify the classification performance of the OpenL3-SVM. The experimental results showed that OpenL3-SVM can effectively classify voice disorders automatically, and its performance exceeds that of the existing methods.

In Suresh and Thomas (2023), it was discovered that diverse feature selection strategies, machine learning classification algorithms and auditory feature combinations provided variations in the accuracy values of dysarthric speech severity level classification. Combining the Random Forest classifier with the Relief feature selection method and PCM-Other Spectral properties led to the highest classification accuracy. This work discussed a comparison study on the severity of dysarthric level classification utilizing various deep learning methods. For the UA-Speech and TORGO datasets, MFCCs were employed as features and analysis of the SVM-based classifier has been done. The outcomes showed that CNN and DNN both performed better than LSTM-based systems and are considerably superior to the often used SVM-based classifier.

The study in Ankışhan and İnam (2021) aims to introduce the new feature vector in the hybrid axis and multi-model in order to diagnose these disorders with more conventional methods. Two types of fusion models (feature and decision level fusion) are used to increase the classification accuracy of the multi-model. It is seen from the experimental results that the proposed feature vector helps to classify pathological data successfully, depending on their pathological conditions. Together with the proposed multi-model, both LSTM and CNN are found to be similarly successful in the classification of data in multi-model architecture. Also in Ksibi et al. (2023), a multi-model architecture which is a coupled CNN–RNN machine learning algorithms for the classification of healthy and pathological audio samples, and a two-level cascaded architecture that enables the accurate identification of pathological voices from the input dataset by incorporating gender information and manually extracted features are proposed.

A study introduced a novel approach for categorizing four common voice disorders (functional dysphonia, neoplasm, phonotrauma, and vocal palsy) (Wang et al., 2022). Instead of a single vowel, this approach utilizes continuous Mandarin speech. The researchers first transformed acoustic data into Mel-frequency cepstral coefficients and then employed a bi-directional long short-term memory network (BiLSTM) to capture the sequential traits of the signal. The results of the experiments demonstrate that this proposed framework yields notable improvements in accuracy and unweighted average recall, compared to systems that only utilize a single vowel.

Our study builds upon these previous works by proposing a system classifying general voice disorders and based on a three-stage framework integrating speech enhancement techniques, a multi-stream approach, and a combined CNN-LSTM architecture to enhance pathological voice classification.

1.2 Objective and contribution

In our recent research, we introduced an efficient approach to develop a precise pathological voice classification system, we used speech enhancement techniques to improve the quality and intelligibility of pathological voice and then, we optimized the acoustic analysis by exploiting robust algorithms to identify a set of combined acoustic parameters which are used as input to a classifier combining two machine learning architectures, namely CNN-BiLSTM, to characterize voice disorders. Our ultimate goal is to create a robust and efficient system for voice pathology classification that can be of assistance to individuals with voice disorders. Here's how we addressed this objective:

(i) Preprocessing for Enhanced Feature Extraction: We implemented a preprocessing step to improve the quality of the audio signal, focusing on preparing it for effective feature extraction. This step is crucial for improving the model's ability to learn from the data.

(ii) Multi-Stream Feature Integration: We employed a multi-stream approach that incorporates various relevant acoustic features, such as Mel-frequency cepstral coefficients (MFCCs), Power normalized cepstral coefficients (PNCCs) and prosodic features. This allows the model to capture a wider range of information within the pathological speech signal, leading to more comprehensive classification of disorders.

(iii) CNN-BiLSTM Classification Model: We designed a classification system using a combination of a CNN and a BiLSTM network. The CNN effectively captures local patterns within the acoustic features, while the BiLSTM leverages the sequential nature of speech data. This combination enables the model to learn complex relationships between features and disorders.

We evaluated the performance of our system on the MEEI pathological database. This database is a well-established resource for voice disorders research and allows us to assess the effectiveness of our system in classifying this specific voice disorder.

The paper's structure is as follows: In Sect. 2, we outline the three-stage voice disorders classification process, encompassing a voice enhancement algorithm, a multi-stream approach employing various parameters, and a presentation of the combined deep learning architectures. Section 3 delves into the database utilized, the algorithms employed for improving the pathological speech classification, the different activation functions proposed, and the evaluation outcomes concerning two sets of pathological voice severity levels. The experimental findings are deliberated in Sect. 4. Finally, we present our conclusions in Sect. 5.

2 Proposed approach

Our proposed system for classifying pathological speech is built upon a robust architecture consisting of three primary components, as depicted in Fig. 1. The initial component primarily focuses on speech quality, to improve the signal's relevance through the application of an enhancement technique. Our research postulates that the speech signal in pathological conditions, shares similarities with signals produced in noisy environments or with added noise. Consequently, we employ the enhancement methods to improve the efficiency and intelligibility of the pathological speech signal, there by mitigating the impact of voice disorders. The second component of our system amalgamates the set of enhanced streams into a multidimensional acoustic vector, contributing to an increased accuracy rate for the trained system. Finally, the set of enhanced and amalgamated features is passed through the deep neural network architecture, which combines both the Convolutional Neural Network and the BiLSTM models. The combined model CNN-BiLSTM yields efficient classification parameters.

2.1 Enhancement of pathological voice

Speech enhancement is a specialized approach for mitigating noise in audio recordings. While The term 'noise reduction' can have a broader scope, speech enhancement specifically targets the detection and removal of unwanted noise in audio to enhance clarity, intelligibility, or overall listening experience, Fig. 2 shows the Basic steps of a speech enhancement system (Kulkarni et al., 2016).

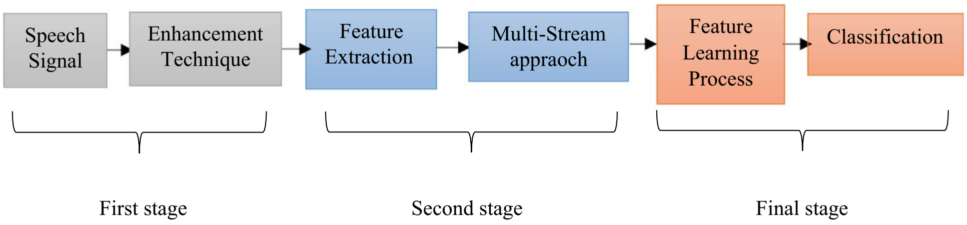


Fig. 1 Synoptic system for classifying pathological speech

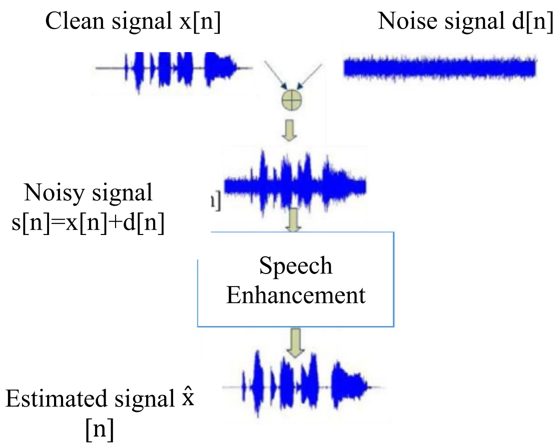


Fig. 2 Basic steps of speech enhancement system

Numerous algorithms are employed for speech enhancement, with signal processing historically being the most prevalent denoising method.

The Minimum Mean Square Error MMSE algorithm is a widely adopted technique in speech enhancement (Gupta et al., 2011), primarily focused on noise reduction in speech signals while preserving speech quality. The primary goal is to minimize the Mean Square Error (MSE) through the utilization of an estimator. The MMSE estimator of the short-term power spectrum is given by Ephraim and Malah (1985), and the calculation is detailed below:

The MMSE algorithm functions in the time–frequency domain and is typically applied to short-time segments of the signal, utilizing techniques such as Short-Time Fourier Transform (STFT). It is given by:

$$S_{(n,k)} = X_{(n,k)} + D_{(n,k)} \tag{1}$$

here n corresponds to the time domain, while k pertains to the spectral domain. $S_{(n,k)}$ represents the noisy signal, which is the sum of the clean signal $X_{(n,k)}$ and the additive noise signal $D_{(n,k)}$.

To estimate the clean voice signal $\hat{X}$, a gain function is utilized to attenuate the noisy signal and it is expressed as follows:

$$\hat{X}_{(n,k)} = S_{(n,k)} \times G \tag{2}$$

The signals are enhanced by the MMSE estimator, which minimizes the mean-square error between the magnitude spectra of the clean and estimated signals, leading to noise attenuation without distorting the signal too much (Loizou, 2007). Its spectral gain function denoted as $G_{(\varepsilon,v)}$ and $v_{(n,k)}$ are calculated as:

$$G_{(\varepsilon,v)} = \frac{\varepsilon_{(n,k)}}{1+\varepsilon_{(n,k)}} \times e^{\left\{\frac{1}{2}\int_{v_{(n,k)}}^{\infty}\frac{e^{-u}}{u}du\right\}} \tag{3}$$

$$v_{(n,k)} = \gamma_{(n,k)} \times \frac{\varepsilon_{(n,k)}}{1+\varepsilon_{(n,k)}} \tag{4}$$

Among the methods used to estimate the noise power spectrum $E\left[\left|D_{(n,k)}\right|^2\right]$ are:

- External noise estimation

This method involves obtaining an estimate of the noise signal, $D_{(n,k)}$, from an external source or a non-speech segment of the recording. This noise estimate can then be used to calculate the power spectral density PSD of the noise, which is the average power of the noise signal at each frequency bin.

The $E\left[\left|D_{(n,k)}\right|^2\right]$ can be approximated by the average value of the noise PSD across the desired frequency range.

- Voice activity detection (VAD) and spectral subtraction

This method leverages the fact that speech and noise often occupy different regions in the time–frequency domain. A Voice Activity Detection (VAD) algorithm can be used to identify speech segments in the signal. VAD plays a crucial role in noise estimation for speech processing applications, ensuring accurate estimation of noise characteristics by identifying and utilizing noise-only segments of the audio signal.

During non-speech segments (assumed to be dominated by noise), the estimated noise PSD can be calculated.

During speech segments, the estimated noise PSD can be subtracted from the noisy signal's spectrum to obtain an estimate of the clean speech spectrum. This estimated clean speech spectrum can then be used to calculate $E\left[\left|X_{(n,k)}\right|^2\right]$, the power spectral density of the clean signal.

The a priori Signal-to-Noise Ratio SNR $\varepsilon_{(n,k)}$ represents the specific spectral bin k at a given time n and is determined by calculating the ratio of the power of the clean signal to the power of the noise signal, defined by:

$$\varepsilon_{(n.k)} = \frac{E\left[\left|X_{(n,k)}\right|^2\right]}{E\left[\left|D_{(n,k)}\right|^2\right]} \tag{5}$$

The a posteriori SNR $\gamma_{(n,k)}$ represents the measured SNR of the given specific spectral bin 'k' at a given time 'n' and is determined by calculating the ratio of the squared magnitude of the noisy signal to the power of and the noise signal, defined by:

$$\gamma_{(n,k)} = \frac{\left|S_{(n,k)}\right|^2}{E\left[\left|D_{(n,k)}\right|^2\right]} \quad (6)$$

The a priori SNR $\varepsilon_{(n,k)}$ is calculated using the decision-directed approach, expressed as:

$$\varepsilon_{(n,k)} = a\frac{X^2_{(n,k)}(l-1)}{E\left[D_{(n,k)}(l-1)^2\right]} + (1-a)max\left[\gamma_{(n,k)}(l) - 1{,}0\right] \quad (7)$$

where a represents the weighting factor.

For the first frame l =0, the expression is as follows:

$$\varepsilon_{(n,k)}(0) = a + (1-a)max\left[\gamma_{(n,k)}(0) - 1{,}0\right] \quad (8)$$

The MMSE algorithm employed in our system utilizes a priori SNR $\varepsilon_{(n.k)}$, to estimate the noise level and attenuate it in the signal. This a priori SNR is crucial for calculating the gain function that minimizes the mean squared error between the estimated clean signal and the original clean signal. Common methods for estimating a priori SNR include external noise estimation and Voice Activity Detection (VAD) with spectral subtraction. In our system, we implemented the Voice Activity Detection (VAD) approach because this algorithm accurately detects and differentiates speech from non-speech segments, contributes to enhanced speech quality, reduces bandwidth usage, improves energy efficiency and increases automation in speech-related tasks. These advantages make VAD a valuable tool in audio processing, speech recognition and classification.

2.2 Multi-stream approach

In our research, we introduced an effective acoustic parameterization method known as the multi-stream approach. This method involves the consolidation of various relevant acoustic parameters into a combined parametric vector. This vector is subsequently employed as input for our pathological voice classifier. The primary advantage of this approach lies in its straightforward integration within the classification system's architecture. Within our application, we have integrated multiple parameters to enhance the relevance of our voice disorders classification system. These parameters encompass MFCC (Davis & Mermelstein, 1980), PNCC (Kim & Stern, 2016), Mel-Spectrogram (Kishore, 2011), parameters reflecting variation in frequency and amplitude (Jitter and Shimmer), and the prosodic parameter F0.

The acoustic analysis plays a pivotal and important role as it yields parameters that capture essential aspects of speech signals. The MFCC coefficients are highly proficient in representing the distinctive characteristics of speech sounds and phonemes since they are derived from models of human auditory perception. Their performances are further enhanced when combined with the robust coefficients PNCC.

The PNCC features are obtained through a gammatone filter bank emphasizing lower frequencies akin to the mel frequency filter bank utilized in MFCC coefficients. The processing features include additional steps compared to MFCC. These steps include the replacement of the logarithmic nonlinearity in MFCC processing by a power-law nonlinearity to remove small signals and variability, the use of medium-time processing with a duration of 50–120 ms to analyze the parameters that characterize environmental degradation, which makes it possible to estimate the degradation of the environment more accurately, the use of a form of asymmetric nonlinear filtering to estimate the acoustic noise level for each time slot and frequency bin, this approach removes slowly changing variables, the development of computationally-efficient realizations of the algorithms above that support 'online' real-time processing, and finally, a signal processing block that performs temporal masking is implemented. The diagram in Fig. 3 illustrates the sequential stages in calculating MFCC, PNCC, and Mel Spectrogram coefficients.

Furthermore, Mel-Spectrogram coefficients, known as a time–frequency representation, seamlessly align with machine learning architectures and excel at capturing diverse spectral characteristics within speech signals. MFCC coefficients, which stem from Mel Spectrogram coefficients, involve an additional step that entails the computation of the Discrete Cosine Transform (DCT) (Sumin et al., 2021).

Jitter and Shimmer coefficients serve as two essential metrics for vocal signal analysis. The Jitter coefficient represents the variation of the fundamental frequency (F0) throughout the temporal evolution of the utterance. It indicates the variability or disturbance of the time period (T0) across several oscillation cycles (Westzner et al., 2005).

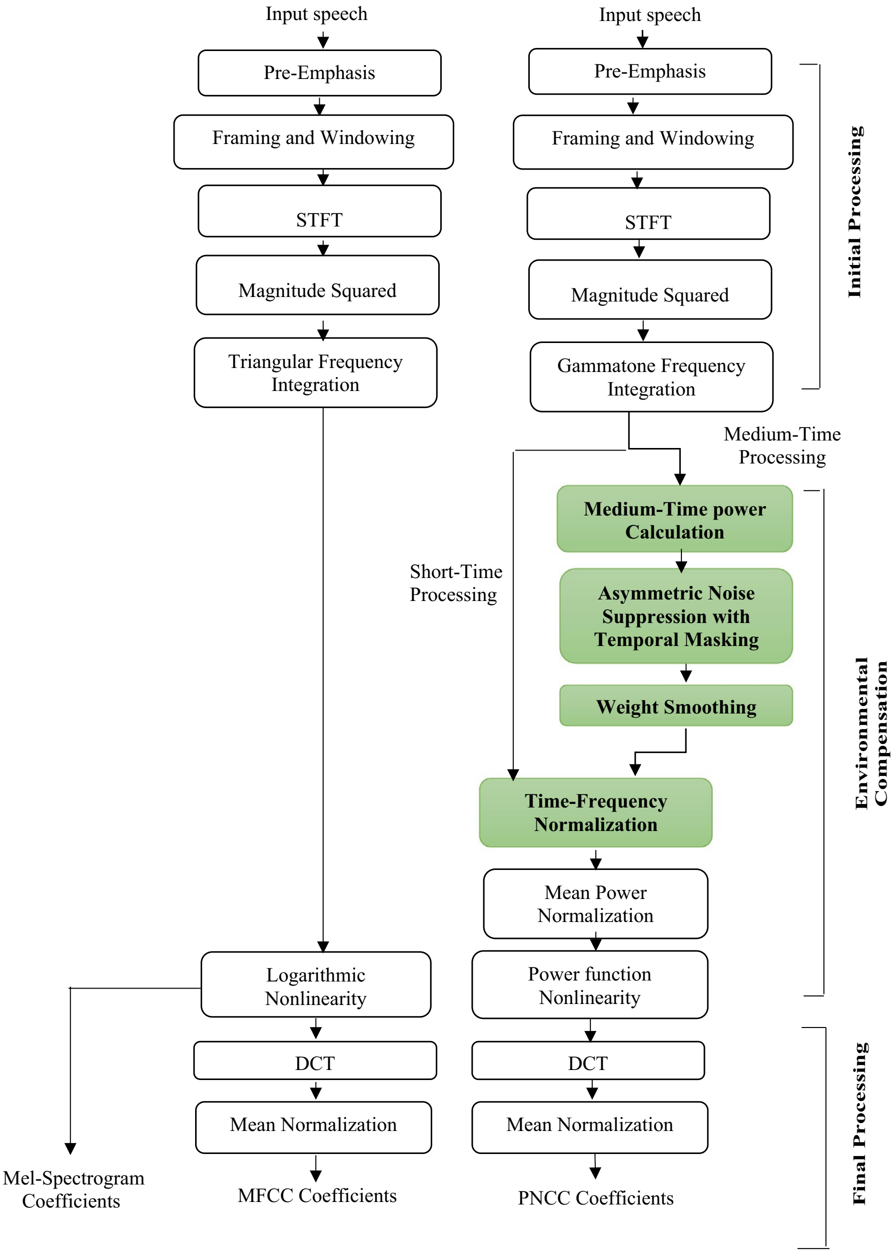
The values for Jitter can be measured in different parameters, such as absolute, relative, relative average perturbation (rap), and the period perturbation quotient (ppq5) (Brockmann et al., 2011; Teixeira et al., 2013). The Jitter absolute is the cycle-to-cycle variation of fundamental frequency, it is the average absolute difference between consecutive periods, expressed as:

$$Jitter(Absolu) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - T_{i+1}\right| \quad (9)$$

where Ti is the extracted glottal period lengths and N: is the number of extracted glottal periods.

While the relative Jitter or local Jitter is the average absolute difference between consecutive periods, divided by the average period, it is expressed as a percentage:

Fig. 3 Comparative analysis of three feature extraction techniques: PNCC, MFCC, and Mel Spectrogram



$$Jitter(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - T_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N}T_i} \times 100 \tag{10}$$

The Shimmer (dB) is expressed as the variability of the peak-to-peak amplitude in decibels (Brockmann-Bauser, 2012), it is the average absolute base-10 logarithm of the difference between the amplitude of consecutive periods, multiplied by 20:

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20 \times log\left(\frac{A_{i+1}}{A_i}\right)\right| \tag{11}$$

where Ai is the extracted peak-to-peak amplitude data and N is the number of extracted fundamental frequency periods.

The Shimmer relative is defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude, expressed as a percentage:

$$Shimmer(Relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - A_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \times 100 \tag{12}$$

2.3 CNN-BiLSTM stacking architecture

Our application introduces a system that leverages a combined deep learning architecture CNN-BiLSTM to improve the accuracy of pathological voice classification. The overall system architecture is depicted in Fig. 4. The system takes as input a matrix of combined streams encompassing different parameters like MFCC, PNCC, and parameters reflecting variation in frequency and amplitude (Jitter and Shimmer). These parameters are processed through a convolutional neural network (Albawi et al., 2017), which comprises a convolutional and max-pooling layer. Each convolution layer employs filters to extract relevant features from the input sequence, producing a feature map as defined in Eq. (13). The advantage of convolutional operations is that nodes in each layer connect to specific node regions in the adjacent layer.

$$h_{i,j} = \sum_{k=1}^{n} \sum_{l=1}^{n} c_{k,l} x_{i+k-1,j+1} \tag{13}$$

where h corresponds to the feature maps, c represents the convolution kernel, x is the input image and n denotes the width and height of the kernel.

Max pooling layers offer the advantage of dimensionality reduction by selecting the maximum value within each window, thus retaining essential information (Gholamalinezhad & Khosravi, 2020). Following the convolutional operation, the CNN-based system applies a Rectified Linear Unit ReLU transformation to the resulting feature map. This introduces non-linearity to the model, facilitating rapid learning. Despite its simplicity, the ReLU activation function proves highly effective in various models, making it the preferred choice for hidden layers. The ReLU activation function is represented by Eq. (14):

$$ReLU(x) = \begin{cases} 0 & if\ x < 0 \\ x & if\ x \geq 0 \end{cases} \tag{14}$$

The data from the max-pooling layer of the CNN network is then processed by a Bi-directional Long Short- Term Memory system (Staudemeyer & Morris, 2019). The Bidirectional Long short-term memory BiLSTM is a type of recurrent model that processes the sequence input in both forward and backward track. The BiLSTM layer is capable of capturing short and long-term contextual dependencies of the sequence input. According to the Eqs. (15–20), The forward track unfolds the network from the first time instance to the last instance, whereas the backward track does the reverse by changing all $t - 1$ to $t + 1$. The two tracks work in parallel, each keeps separate weights and biases. Their hidden states h are simply stacked together at each time t and are transmitted as input to the two tracks of the higher layer. BiLSTM is designed to mitigate the vanishing and exploding gradient problems often encountered when training on long sequences in Recurrent Neural Networks (RNNs). This feature makes BiLSTM an ideal choice for speech modeling. Furthermore, stacking BiLSTM layers on top of one other results in two-dimensional depth for learned patterns, considering both the time dimension and the feature hierarchy dimension. Within each layer, BiLSTM progresses over time t as follows (Alhussein & Muhammad, 2019; Xiaoyu, 2018):

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right) \tag{15}$$

$$g_t = tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{16}$$

$$c_t = f_{t\,o}\, c_{t_1} + i_{t\,o}\, g_t \tag{17}$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right) \tag{18}$$

$$h_t = o_{t\,o}\, \tanh\left(c_t\right) \tag{19}$$

$$z_t = softmax\left(W_{hz}h_t + b_z\right) \tag{20}$$

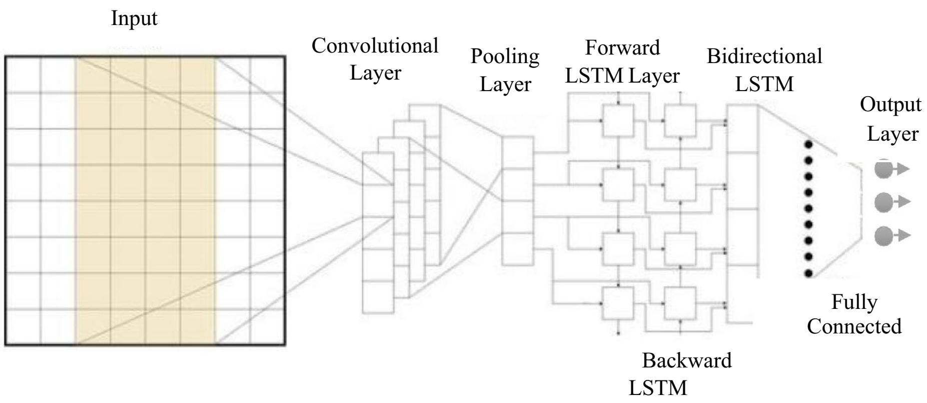


Fig. 4 Architecture of the CNN-BiLSTM proposed

where *ft* represents the forget gate. The parameters *it* and *ot* denote the input and output of the gate respectively, while *gt* represents the modulation gate. The forget gate addresses the vanishing gradient problem by ensuring that the problematic product component has elements close to one (Xing Luo, 2019).

Lastly, an output layer using Softmax function with multiple classes is essential and it is expressed by Eq. (21) as follows (Xing Luo, 2019):

$$softmax(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}} \tag{21}$$

Subsequently, the cross-entropy loss is computed by measuring the discrepancy between the actual labels and the predicted ones.

Figure 5 below summarizes the different methods proposed in our research to build the pathological speech classification system.

3 Experimental configuration and outcomes

In this section, we will offer an insight into the datasets employed to assess the proficiency of our pathological speech classification system. We will delve into the different speech enhancement techniques employed and elucidate how amalgamating acoustic parameters augmented our system's performance. Our system is grounded in the CNN-BiLSTM network, and we have adopted distinct activation functions to enhance accuracy. Furthermore, we will disclose the findings of our pathological classification experiments, based on two sets of pathological data.

3.1 Data source

The datasets employed in our study originated from the MEEI Voice Disorders database, which was developed by the MEEI Voice and Speech lab. This database comprises more than 1400 voice samples and has become a valuable resource in the field of voice pathology detection and classification, despite some inherent limitations (Disordered Voice Database, 1994). Notably, it is commercially accessible through Kay Elemetrics. A notable limitation of this database is that normal and pathological voices were recorded in dissimilar environments and at varying sample frequencies. The recordings within this database encompass sustained phonation of the vowel /ah/ including 53 samples from normal voices and 657 from pathological voices, as well as the first sentence of the rainbow passage, comprising 53 normal voice samples and 662 pathological voice samples. Of these samples, 77 pathological vowels and all normal vowels were recorded at a 50 kHz sample rate, while the remaining 580 pathological vowels were sampled at 25 kHz. Additionally, 36 of the normal rainbow sentences were recorded at 25 kHz and 17 at 10 kHz. Among the pathological sentences, 648 were sampled at 25 kHz, 13 at 10 kHz, and one at 50 kHz. The database encompasses various tools for voice condition assessment, including stroboscopy, acoustic aerodynamic measures, and a physical examination of the neck and mouth, with the information provided by Kay Elemetrics. For our study, we specifically selected only sustained vowel /a/ samples.

3.2 Overview of selected pathologies

Within the medical domain, voice disorder databases serve as a crucial resource for the development of pathological speech classification systems designed to distinguish between various pathologies afflicting individuals with vocal disorders. In this classification system, pathological voice is considered as a noisy signal. Table 1 provides the initial set of data, highlighting the voice pathologies that can adversely affect the vocal cords, leading to voice-related challenges. Vocal cord nodules, for instance, represent non-cancerous growths that manifest as a result of vocal strain, misuse, or abuse. These growths are commonly observed in individuals who extensively use their

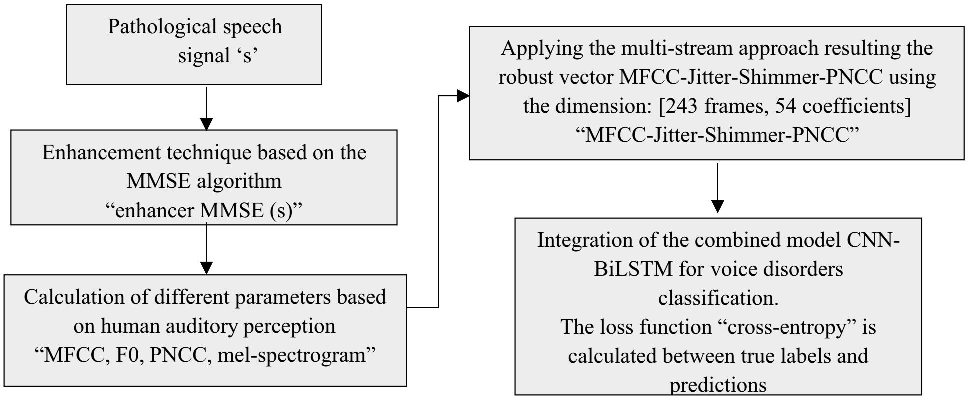

Pathological speech signal 's'
Enhancement technique based on the MMSE algorithm "enhancer MMSE (s)"
Calculation of different parameters based on human auditory perception "MFCC, F0, PNCC, mel-spectrogram"
Applying the multi-stream approach resulting the robust vector MFCC-Jitter-Shimmer-PNCC using the dimension: [243 frames, 54 coefficients] "MFCC-Jitter-Shimmer-PNCC"
Integration of the combined model CNN-BiLSTM for voice disorders classification. The loss function "cross-entropy" is calculated between true labels and predictions

Fig. 5 Proposed approach steps

Table 1 Number of female and male records and average age for the first dataset

Diseases	Female records	Average female age ± std	Male records	Average male age ± std	Total
Nodules	19	28.47 ± 9.87	1	47 ± 0	20
Paralysis	27	52.56 ± 16.94	31	53.48 ± 19.12	58
Polypoid	21	46.43 ± 12.39	4	51.5 ± 16.84	25

voices, such as singers, actors, teachers, and public speakers. Vocal cord nodules can bring about symptoms such as hoarseness, breathiness, and rough or raspy voice quality (Karkos & McCormick, 2009). Vocal cord paralysis, on the other hand, occurs when nerve damage impairs the proper movement of one or both vocal cords to move properly. This condition can stem from various factors, including viral infections, neurological disorders, surgical trauma, or nerve injuries affecting the vocal cords. Vocal cord paralysis often leads to voice problems, difficulties in breathing, and swallowing issues. It may result in a weak, breathy, or strained voice (Toutounchi et al., 2014). Polypoid lesions, also known as vocal cord polyps, are benign growths that form on the vocal cords, typically as a consequence of vocal abuse or trauma, such as chronic coughing, screaming, or excessive yelling. These polyps come in various sizes and shapes and are known to induce voice changes, including hoarseness, roughness, and a reduction in vocal range (Zhuge et al., 2016).

The second subset of pathologies includes A–P squeezing, gastric reflux, and mild ventricular compression, with an equal distribution of male and female speaker files as detailed in Table 2. To ensure comparability, groups with similar average ages were carefully selected.

Gastric reflux can lead to voice disorders when stomach acid enters the esophagus and larynx, causing irritation and inflammation. This can result in symptoms such as hoarseness, throat clearing, and a sensation of a lump in the throat. Chronic reflux can lead to more severe vocal fold damage over time (Vakil et al., 2006).

A posterior-anterior squeezing typically occurs due to improper vocal fold closure. The vocal folds are not coming together evenly, resulting in a squeezing sensation during phonation. Etiologies may include vocal fold nodules, polyps, muscle tension dysphonia, or neurological conditions affecting vocal fold movement (Behrman et al., 2003).

This ventricular compression disorder involves compression of the ventricular folds (false vocal folds) during phonation, often resulting in a strained or breathy voice quality. People with ventricular compression disorder may have a voice with the following qualities: severe dysphonia (abnormal voice), low pitch, roughness, and strain. Ventricular compression disorder can have various etiologies, including excessive muscle tension, vocal fold misalignment, laryngeal pathology, neurological conditions, and psychological factors (Bailly et al., 2014).

These disorders affect vocal mechanisms differently and can result in various changes and difficulties in voice quality, potentially leading to dysarthria. The severity of dysarthria depends on the specific characteristics and severity of the underlying disorder, as well as the individual's overall health and any pre-existing neurological conditions. Severe dysarthria, such as that caused by significant A–P squeezing or ventricular compression, can markedly impact speech intelligibility due to pronounced changes in voice quality (Kent & Kim, 2008). Symptoms of gastric reflux can exacerbate throat irritation and voice quality issues in cases of severe dysarthria.

3.3 Experimental setup

To improve the performance of our pathological speech classification system, we employed a multi-stream approach. This approach entails the amalgamation of diverse acoustic parameters that aptly represent the characteristics of the speech signal. We selected parameters that demonstrated high efficiency and resilience, primarily based on models of human auditory perception. Subsequently, we adapted these parameters to our classification task by incorporating them into a combined CNN-BiLSTM system.

The calculation of these parameters was carried out utilizing a Hamming window with a duration of 25 ms and a shift

Table 2 Number of female and male records and average age for the second dataset

Diseases	Female records	Average female age ± std	Male records	Average male age ± std	Total
A–P squeezing	15	49.87 ± 18.62	15	46.8 ± 22.19	30
Gastric reflux	15	46.6 ± 17.37	15	44 ± 14.71	30
Ventricular compression (mild)	12	41.25 ± 12.83	13	46.46 ± 13.40	25

of 10 ms. For optimal training of our system, we established the learning parameters, which are detailed in Table 3 below. The number of the BiLSTM units used in our application differs depending on the evaluated systems.

3.4 Impact of multi-stream approach

In our study, we developed a CNN-BiLSTM network and integrated robust acoustic parameters as inputs into the system. To enhance the performance of our system, we employed a multi-stream approach, which involved amalgamating various acoustic parameters into a single feature vector. This vector was created by concatenating the MFCC coefficients including the energy, and their first and second derivatives, along with the 13 PNCC coefficients and parameters measuring vocal signal perturbations like Jitter, and Shimmer.

The selection of acoustic parameters significantly influences the performance of pathological speech classification systems. In our study, we incorporated a range of acoustic parameters, each chosen for its specific contributions:

- MFCC (Mel-frequency cepstral coefficients): These coefficients capture the spectral envelope of speech signals and are widely used in speech processing tasks due to their effectiveness in representing spectral characteristics and facilitating tasks such as speaker identification and speech recognition.
- PNCC (Power normalized cepstral coefficients): PNCC coefficients aim to capture perceptually relevant aspects of speech signals and have demonstrated superior robustness in noisy environments and voice disorder contexts compared to MFCCs.
- Jitter and Shimmer: These acoustic measures are fundamental in voice quality analysis, providing insights into vocal fold stability, regularity, and vibratory characteristics. They are particularly valuable for detecting pathological voice disorders and distinguishing between normal and disordered speech.

Table 3 Configuration of various parameters comprising the CNN-BiLSTM network

	Optimized parameters for the CNN- BiLSTM system
Input dimensions	243 frames, 54 coefficients
Batch size	10
Number of epochs	320
Dropout rate	0.2
Number of filters of each convolutional hidden layers	[12, 24, 48, 96, 192, 384]
Filter size dimension	3×3
Learning rate	0.0003
LSTM units	Variable
Hamming window size	25 ms
Optimization algorithm	Adam

A statistical evaluation of each parameter's contribution to the classification task would further enhance the empirical foundation of our study. This analysis could provide valuable insights into the relative importance of these parameters, strengthening the reliability of our conclusions.

The normalization of the extracted parameters MFCC and PNCC is a crucial step to ensure their robustness and reduce their sensitivity to speaker variability and recording conditions. In our application, we employed variance normalization (or standardization), which is typically performed to scale the coefficients to a standard range. This step ensures that coefficients with larger variances do not dominate the feature vector. The normalized coefficient $\hat{c}_i$ is computed as (Berouti et al., 1979):

$$\hat{c}_i = \frac{c_i - u_i}{\sigma_i} \tag{22}$$

where σ_i is the standard deviation and u_i is the mean of the i -th coefficient across all frames.

Normalization of MFCC and PNCC coefficients is particularly crucial in speech-related applications as it minimizes the impact of speaker variations, recording conditions, and acoustic environments on feature representation. This standard preprocessing technique is widely adopted in tasks such as speech recognition, speaker identification, and speech classification.

The combination of all these valuable acoustic parameters MFCCs, PNCCs, jitter, and shimmer using the multi-stream approach allows to achieve high performance since each provides unique information about the spectral and temporal characteristics of speech signals.

We evaluated the system based on the multi-stream approach using the MEEI database and observed that it achieved impressive classification accuracy values compared to the systems based on a single data stream. The results are presented for Nodules, Paralysis, and Polypoid voice pathologies, as shown in Table 4 below.

The results in Table 4 indicate that the multi-stream approach yielded a high level of accuracy in the pathological speech classification system. Specifically, the concatenated vector comprising MFCC, Jitter, Shimmer, and PNCC (a total of 54 coefficients) achieved an efficient accuracy rate of 85.71%. This outperformed the combined vector containing MFCC-F0-Jitter-Shimmer, and PNCC, which included the fundamental frequency F0. In contrast, the classification based on the use of only the MFCCs, PNCCs, and the coefficients Mel-Spectrograms with a size of 243×50, where 50 represents the frequency bin and 243

Table 4 CNN-BiLSTM network accuracy with various acoustic features for the classification of nodules, paralysis, and polypoid voice pathologies

Acoustic feature	Accuracy (%)
PNCC(39)	61.90
MFCC(39)	66.67
Mel-Spectrogram	71.43
Mel-Spectrogram-Jitter-Shimmer	76.19
Mel-Spectrogram-Jitter-Shimmer-PNCC	71.43
MFCC(39)-F0-Jitter-Shimmer-PNCC(13)	80.95
MFCC(39)-Jitter-Shimmer-PNCC(13)	**85.71**

Bold indicates the highest accuracy

denotes the number of frames produced a lower accuracy rates, except when supplementing the Mel-Spectrogram with Jitter and Shimmer coefficients, the accuracy reach 76.19%. However, this accuracy rate was slightly lower than that achieved by the multi-stream concatenation method based on the MFCC- F0-Jitter-Shimmer-PNCC vector. From the results, we concluded that the coefficients Jitter and Shimmer are well suited in combination with the PNCC and MFCC coefficients than with Mel-Spectrograms parameters. The multi-stream approach notably outperformed the speech disorder classification systems, resulting in a 24% improvement in accuracy value when compared to the original system based solely on PNCC coefficients.

Multiple experiments were carried out with different data partitions to identify the most effective configuration for the test subsets. As indicated in Table 5, the CNN-BiLSTM system trained and tested using the combined acoustic features MFCC-Jitter-Shimmer-PNCC vector of 54 coefficients exhibited the highest performance split configuration (70/30).

3.5 Pathological voice enhancement

To improve the performance of the pathological speech classification system and improve accuracy rates, we have introduced speech enhancement techniques to our datasets such as the MMSE algorithm, Wiener filter, and spectral subtraction.

In the context of the Minimum Mean Square Error (MMSE) algorithm, particularly for noise estimation and subsequent noise reduction in speech or signal processing, dividing the frequency spectrum into sub-bands serves several important roles, the MMSE algorithm can adapt its noise estimation and subsequent processing strategies to better suit the characteristics of speech signals in each frequency band. Also, the MMSE algorithm can selectively enhance speech components while suppressing noise in a more targeted manner, which allows for more precise noise estimation and adaptive processing tailored to the spectral characteristics of speech and noise. This approach investigated in Brijesh Anilbhai and Kinnar (2017) helps to minimize the impact on perceived speech quality. In particular, we have performed the enhancement technique based on the MMSE algorithm considering the noise estimation in the frequency ranges of 0–1.5 kHz, 0–2.5 kHz, and 0–3.5 kHz. These specific frequency ranges chosen for noise reduction were selected based on the understanding that speech information is primarily concentrated in lower and mid-frequency ranges, while noise can be present across a wider spectrum.

For comparative purposes, we have also utilized the Wiener algorithm and Berouti's spectral subtraction method (Strand & Egeberg, 2004).

Spectral subtraction is a commonly used method in speech enhancement due to its simplicity (Kaladharan, 2014). The Berouti algorithm extends spectral subtraction by subtraction of not only amplitude spectra but also power spectra. The power spectra of the estimated clean signal, the noisy signal, and the noise are represented as Ps, Px, and Pd,

respectively. The spectral subtraction estimator can be defined as follows (Brijesh Anilbhai & Kinnar, 2017):

$$\widehat{P}_s = \left(P_x^{\sigma} - \beta P_d^{\sigma}\right)^{1/\sigma} \tag{23}$$

σ and β are the two parameters allowing the overestimation of the power spectrum of the noise signal and raising the power spectrum before subtraction respectively.

The Wiener filter is a method designed to minimize the average squared error between the desired signal and the estimated signal by leveraging the spectral characteristics of both the desired signal and the noise (Lim & Oppenheim, 1979). It calculates a gain for each frequency bin based on the ratio of the estimated signal power to the estimated noise power.

Table 5 Corpus splitting techniques employed in the CNN-BiLSTM system utilizing MFCC-Jitter-Shimmer-PNCC features for the first subset

Technique splitting	90/10	80/20	70/30	60/40	50/50	Three-fold
Training + validation	90%	80%	70%	60%	50%	67%
Test	10%	20%	30%	40%	50%	33%
Accuracy (%)	80	85.71	**86.67**	75.61	76.47	68.91

Bold indicates the highest accuracy

This factor can be expressed as:

$$G(k) = \frac{X(k)}{X(k) + D(k)} \tag{24}$$

where G(k) represents the wiener gain calculated for the k frequency bin, S(k) represents the estimated power spectral density of the clean speech and N(k) denotes the estimated power spectral density of the noise. This formula can be derived considering the signal x and the noise d as uncorrelated and stationary signals. The SNR is defined by Picone (1993):

$$SNR = \frac{X(k)}{\widehat{X}(k)} \tag{25}$$

This definition can be integrated into the Wiener filter equation as follows:

$$G(k) = \left[1 + \frac{1}{SNR}\right]^{-1} \tag{26}$$

The fixed frequency response at all frequencies and the requirement to estimate the power spectral density of the clean signal and noise prior to filtering is the drawback of the Wiener filter.

Table 6 Accuracy comparison of CNN-BiLSTM network using various voice signals in the classification of Nodules, Paralysis, and Polypoid voice pathologies

Signals	Accuracy (%)
Signal based on multi-stream approach	85.71
Enhanced signal	
Wiener filter	85.71
Spectral Subtraction (Berouti Algorithm)	90.48
MMSE 15	90.48
MMSE 25	90.48
MMSE 35	**95.24**

Bold indicates the highest accuracy

We conducted tests to assess the performance of our system, which is based on the multi-stream paradigm consisting of the MFCC- Jitter- Shimmer-PNCC combined vector. Table 6 displays the accuracy results of the CNN-BiLSTM system when various speech enhancement methods were applied to a dataset involving nodules, paralysis, and polypoid pathologies.

Our findings indicate that the wiener filter maintains a recognition rate of 85.71%, while the spectral subtraction-based Berouti algorithm demonstrates a 4.77% improvement in accuracy. Moreover, the MMSE-based enhancement, with noise estimation over different frequency ranges, such as (0–1.5 kHz), (0–2.5 kHz), and (0–3.5 kHz) referred to as MMSE 15, MMSE 25, and MMSE 35 respectively, results in high recognition rates. MMSE 15 and MMSE 25 achieve a recognition rate of 90.48%. However, the best recognition rate is achieved when using MMSE 35, showing an improvement of 9.53%. This frequency range contains information highly suitable for dysphonia classification (Pouchoulin et al., 2007).

3.6 Effect of activation function

Activation functions play a crucial role in speech classification as they help capture intricate patterns in the data. They are applied to the output of the network layer neuron. To understand their impact on the classification of pathological speech, we evaluated various standard activation functions, including GELU (Gaussian Error Linear Unit) (Lee, 2023), SELU (Scaled Exponential Linear Unit) (Klambauer et al., 2017), AReLU (Asymmetric Rectified Linear Unit) (Mediratta et al., 2021), and SinRU (a combination of the RELU activation function and the sinus periodic function) (Chaiani et al., 2022).

Table 7 displays these activation functions along with their corresponding mathematical equations. We used these activation functions to train our application systems, as outlined in Table 8.

The comparison of accuracy rates between the methods MMSE 15, MMSE 25, and MMSE 35 incorporated to

Table 7 Different activation functions

Activation function	Expression
Asymmetric Rectified Linear Unit (AReLU)	$AReLU(x,\alpha,\beta) = \begin{cases} C(\alpha)x & if\, x < 0 \\ (1+\sigma(\beta))x & ifx \geq 0 \end{cases}$
Scaled Exponential Linear Unit (SELU)	$SELU(x) = \begin{cases} \lambda x & if\, x > 0 \\ \lambda\alpha(e^x - 1) & ifx \leq 0 \end{cases}$ $\alpha = 1.67\ and\ \beta = 1.05$
Gaussian Error Linear Unit (GELU)	$GELU(x) = \frac{x}{2}(1 + tanh(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)))$
Sinus Rectified Linear Unit (SinRU)	$SinRU = \begin{cases} 0 & if\, x \leq 0 \\ x + \sin(x) & if\, x > 0 \end{cases}$

Table 8 Pathology classification rate (%) using original and enhanced signals with different activation functions for three pathologies classes: Nodules, Paralysis, and Polypoid

Activation function	ReLU	SELU	GELU	SinRU	AReLU
Signal based on multi-stream approach	85.71	76.19	**90.48**	80.95	85.71
MMSE 15	**90.48**	80.95	**90.48**	85.71	80.95
MMSE 25	**90.48**	80.95	80.95	80.95	80.95
MMSE 35	**95.24**	80.95	90.48	76.19	**95.24**

Bold indicates the highest accuracy

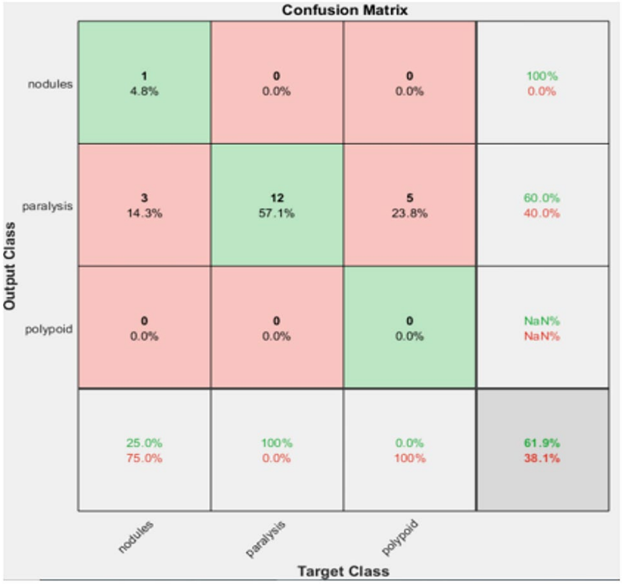


Fig. 6 Confusion matrix of original system-based MFCC coefficients

the system based on the multi-stream approach revealed overall improvements. Notably, The ReLU (Hara et al., 2015), SELU, and SinRU activation functions enhanced the MMSE 15 signal-based accuracy by 4.77% when compared to the original signal. A similar enhancement (Accuracy rate = 4.77%) was observed with the ReLU and SELU activation functions in the case of MMSE 25 signal-based accuracy. However, the MMSE 35 signal-based accuracy saw a substantial 9.53% improvement when employing the ReLU and AReLU activation functions, with a 4.77% boost achieved using the SELU activation function. In the context of assessing activation functions, we observed that the use of the ReLU and AReLU activation functions led to a significant improvement in the accuracy rate of the MMSE 35 enhancement compared to the signal based on the multi-stream approach. Figures 6 and 7 provide the confusion matrix details for the system using only PNCC coefficients and the system integrating the multi-stream approach based on MFCC-Jitter-Shimmer-PNCC

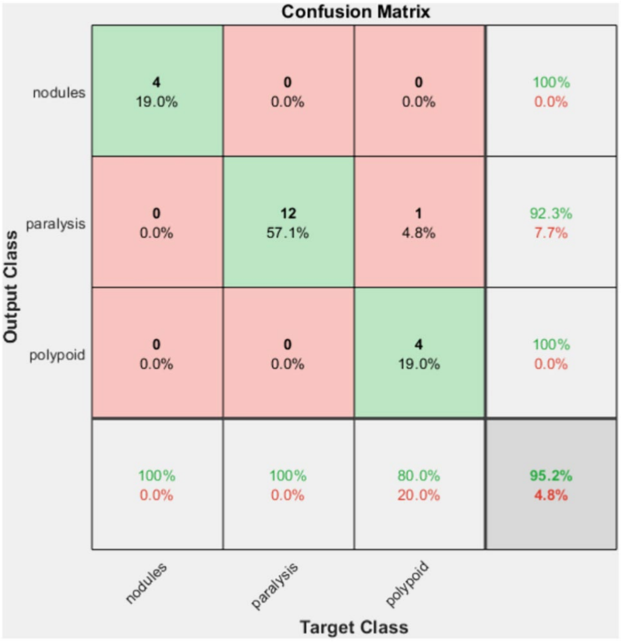


Fig. 7 Confusion matrix of the improved system based on the two techniques: multi-stream approach and speech enhancement MMSE 35

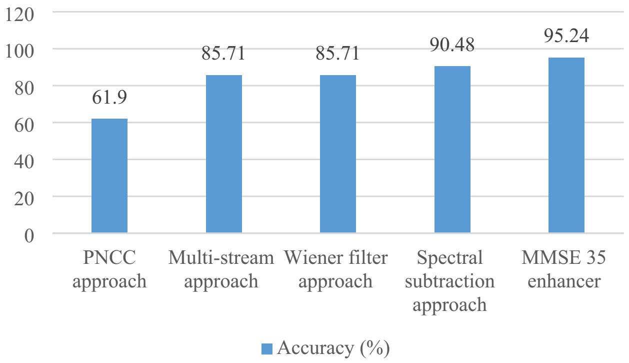


Fig. 8 Accuracy comparison of the different approaches performed by the CNN-BiLSTM classifier using three pathological classes: Nodules, Paralysis, and Polypoid

coefficients, and the enhancement method based on MMSE 35. The results underscore the substantial improvement brought about by the multi-stream approach and the speech enhancement technique, particularly for the paralysis and polypoid classes, where the improvement amounted to 33% over the original system relying solely on PNCC coefficients, as illustrated in Fig. 8.

To demonstrate the effect of the activation function AReLU and the efficiency of our proposed system in classifying voice disorders, we conducted a series of additional experiments using the second subset of data, which

Table 9 Corpus splitting techniques used in the CNN-LSTM system based on MFCC-Jitter-Shimmer-PNCC features for the second subset

Technique splitting	90/10	80/20	70/30	60/40	50/50	Three-fold
Training + validation	90%	80%	70%	60%	50%	67%
Test	10%	20%	30%	40%	50%	33%
Accuracy (%)	37.5	**52.94**	48	41.18	40.48	29.47

Bold indicates the highest accuracy

Table 10 Comparison of voice disorders classification performance using ReLU and AReLU activation functions and the MMSE 35 enhancement technique for three pathologies classes: A–P squeezing, Gastric reflux, and Ventricular compression (mild)

Method	Accuracy (%)
Signal based on multi-stream approach (MFCC-Jitter-Shimmer-PNCC)	52.94
MMSE 35 & ReLU	58.82
MMSE 35 & AReLU	**64.71**

Bold indicates the highest accuracy

combines the pathologies: A–P squeezing, Gastric reflux, and mild Ventricular compression.

This subset was carefully balanced with an equal number of files from male and female speakers. We compared the baseline system based on the multi-stream approach and the CNN-BiLSTM model, with those employing the enhancement technique based on the MMSE 35 utilizing the ReLU and AReLU as activation functions. The performance of these systems was evaluated and the results are outlined in Table 10.

To ensure the validity of our approach, we first explored various data-splitting configurations on the second subset of our system. The best classification rate was reached when using (80/20) split configuration, the results are presented in Table 9. The improvement of the performance of the baseline system was performed and the best accuracy achieved was 52.94% compared to other splitting.

Then, we proceeded to enhance the baseline system's performance, resulting in the best accuracy rate of 64.71% as shown in Table 10.

The results obtained in Table 10 demonstrate that both the incorporation of the enhancement technique and utilization of the AReLU function lead to notable improvements in the severity level classification system. In direct comparison to the system based on the multi-stream approach, the MMSE 35 signals exhibited an accuracy improvement of 5.88% for the ReLU function and an even more substantial 11.77% improvement for the AReLU function as shown in Fig. 9.

3.7 Assessment of the proposed approach

To evaluate the performance of the proposed system, we have conducted an experiment using the same experimental conditions as those established in the work presented in Hamdi et al. (2018) based on the MEEI database. Four classes were tested: (Nodules, Spasmodic, Polypoid, and normal). The performance of our system based on the CNN-BiLSTM network using the enhancement technique and the multi-stream approach and those of the system of Hamdi et al. (2018) are shown in Table 11. As indicated in Hamdi et al. (2018), the system is based on the hidden Markov model associated with a Gaussian mixing density (HMM-GM), and the accuracy rate increased to 94.44% using the combination vector MFCC_HNR_NHR_DFA_F0. As depicted in Table 11, our system based on the multi-stream approach only achieved an accuracy rate of 88.57% using the CNN model. However, after applying the enhancement technique, the accuracy increased to 91.67%. Then, this latter achieved an optimal value of 100% when concatenating the CNN network with the BiLSTM layer and improved the

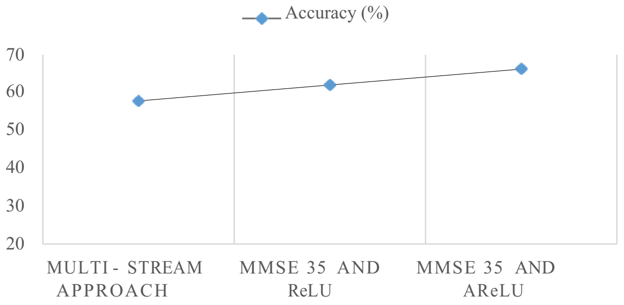


Fig. 9 Comparison of accuracy among different approaches for the combined CNN-BiLSTM system in classifying three pathological classes: A–P squeezing, Gastric reflux, and Mild Ventricular compression

Table 11 Comparison of the performance of our system with those of HAMDI Rabeh et al. for multi-class classification of speech pathologies

System	Accuracy (%)
Our system (CNN network + multi-stream)	88.57
Our system (CNN network + multi-stream approach + MMSE 35)	91.67
Our system (CNN-BiLSTM network + multi-stream approach + MMSE 35)	**100**
System proposed in Hamdi et al. (2018)	94.44

Bold indicates the highest accuracy

performance by 5.56% compared to the system presented in Hamdi et al. (2018).

4 Discussion

The system proposed in this study represents a significant advancement in the field of pathological speech classification, demonstrating its effectiveness when compared to existing state-of-the-art systems. The architecture is based on a three-stage framework that is proposed to perform a precise classification of voice disorders. The first step compromises a robust acoustic analysis based on the multi-stream approach that combines crucial vocal parameters, such as MFCC, PNCC, and coefficients representing variations in frequency and amplitude (Jitter and Shimmer). This combination of acoustic features allows to achieve optimal performance, since each acoustic parameter provide unique information about the spectral and temporal characteristics of speech signals.

The second critical step involves noise reduction techniques to improve the efficiency and intelligibility of the pathological speech signal. This noise reduction process enhances the overall robustness of the voice disorders classification systems.

The classification system for voice disorders is improved by leveraging a feature extraction architecture that combines a Convolutional Neural Network (CNN) and a Bidirectional Long Short-Term Memory network (BiLSTM). This powerful feature extractor captures local patterns and sequential information within the acoustic features, enabling robust classification of voice disorders.

The integration of the multi-stream approach in our study significantly improved the system's classification accuracy. As demonstrated in Table 4, this proposed paradigm proved its effectiveness and robustness, elevating the recognition rate to 85.71%. Employing a split configuration (70/30) for the train and test dataset further improved the model's performance to 86.67%. The speech enhancement module, rooted in signal processing, emerged as a crucial denoising method, contributing to the enhancement of our pathological speech classification system. Various approaches, including Wiener filtering, spectral subtraction, and the minimum-mean square error (MMSE) enhancer were explored. The most notable recognition rate of 95.24% for classifying Nodules, Paralysis, and Polypoid pathologies was achieved through the MMSE approach over the frequency range of 0–3.5 kHz.

During our study, we emphasized the impact of the activation function on the voice disorders classification system's efficiency. Rectified Linear Unit (ReLU), a widely used activation function, is simple and easily implementable in various classifier models, including CNN models. Additionally, the Asymmetric Rectified Linear Unit (AReLU) yielded satisfactory results. For instance, in the classification of three pathological voices A–P squeezing, Gastric reflux, and Ventricular compression, which represent the second and most severe dataset of the MEEI database compared to the first dataset, the highest accuracy achieved was 64.71% when using the AReLU activation function, as indicated in Table 10. The results obtained can be considered satisfactory since the multiclass identification is a complicated task compared to the binary classification task. For instance, the multi-model CNN-LSTM proposed in Ankışhan and İnam (2021) reached an accuracy rate of 99.58% using the SVD database to classify pathological and healthy voices (binary classification), while our system based on more complex task, which consists of identifying three classes of pathological voice reached significates results. However, the system developed by Ksibi et al. (2023) and based on multi-model CNN-RNN and two-level cascaded architecture reached accuracy rate of 88.83% for binary classification.

For the same context, researchers in Guedes et al. (2019) used the German Saarbrücken Voice Database to develop a system classifying four classes: dysphonia, laryngitis, paralysis of vocal cords, and healthy voices. Two models were developed based on Long-Short-Term-Memory and Convolutional Neural Network for classification of extracted embeddings and comparison of the best results, using cross-validation. The results showed an improvement of 40% f1-score for the four classes, 66% f1-score for Dysphonia x Healthy, 67% for Laryngitis x healthy and 80% for Paralysis x Healthy. By comparison, our system reached the best performance of 100% for classifying four classes: Nodules, Spasmodic, Polypoid, and normal voice as presented in Table 10. Furthermore, our results found for the classification of three classes.

In Harar et al. (2019), the system presented achieved an accuracy of 68.08% in voice pathology detection (binary task) using the SVD database and a 1D-CNN-LSTM architecture with segmented raw signals as inputs. (Wu et al., 2018) used spectrograms features to identify organic dysphonia (binary task) by a network composed of 2D-CNN and fully connected layers pretrained by a convolutional deep belief network DBN. The accuracies achieved are 71% and 77% with and without the pretraining process respectively.

The gammatone spectral latitude (GTSL) coefficients were proposed in Zhou et al. (2022) to improve the performance of the classification system between healthy, neuromuscular and structural voices. The proposed features achieved an average accuracy of 99.6% in the Massachusetts Eye and Ear Infirmary (MEEI) database. The accuracies in other databases, Saarbruecken Voice Database (SVD) (Pützer & Barry, 2007) and Hospital Universitario Principe de Asturias (HUPA) were 89.9% and 97.4% respectively. On a more complex task, which consists of classifying four classes: Nodules, Spasmodic, Polypoid, and normal voice, our proposed system achieved 100% accuracy.

The work investigated in Deli et al. (2022) used phase space reconstruction and convolution neural network to classify the normal and pathological voice. The phase space information of normal and pathological voice is reconstructed using delay time and embedding dimension, the one-dimensional signal is converted to a two-dimensional matrix, and the reconstructed trajectory graph sample of the signal is generated. The trajectory graph samples are used as the input of the VGG-like convolutional neural network, and the graphical features are extracted to achieve a classification of normal and pathological voices. The average accuracy rates obtained are 96.04% and 92.27% for normal, vocal fold paralysis, and vocal fold non-paralysis voice in the MEEI database and SVD database respectively. By comparing the performance of our system based on the classification of three disordered classes achieving an accuracy rate of 95.24% and those of the systems demonstrated in Deli et al. (2022) we conclude that the proposed three-stage architecture of this study is effective in performing a multiclass identification of voice disorders.

The ultimate goal of this study is to provide an efficient system for people suffering from speech disorders. This study utilized the MEEI Voice Disorders database, providing a diverse range of disorders speech samples. However, variability in recording environments and conditions could potentially influence the classification results. Evaluating the system on additional databases with different recording characteristics would be beneficial for assessing generalizability and robustness. In our future work, we aim to incorporate well-established databases like the Saarbruecken Voice Database SVD into the evaluation process to strengthen the validity of our findings and provide a more comprehensive assessment of the proposed classification system's performance across diverse recording conditions.

5 Conclusion

In conclusion, this study presents a comprehensive approach to enhance the automatic classification of voice disorders through a three-stage voice pathology classification system.

The initial contribution involves a pre-processing step that leverages the Minimum Mean Square Error (MMSE) enhancer, to improve speech quality and intelligibility.

After, a multi-stream approach is introduced, combining various acoustic vectors to capture diverse patterns and characteristics in the vocal signal. This approach, grounded in robust acoustic parameters motivated by auditory processing, yields efficient results.

Finally, to outperform the results, we suggested a system that combines two deep learning architectures CNN and BiLSTM to improve the accuracy of pathological voice classification.

Following the results obtained, the multi-stream approach used in this work improved the accuracy rate of the voice classification system based on the PNCC features by 24% when using the efficient combined vector MFCC-Jitter-Shimmer-PNCC. Besides, the enhancement techniques adopted in this research contributed effectively to the improvement of the system performance, particularly the MMSE technique with the range of frequency [0–3.5 kHz]. This metric allows improving the accuracy rates with a factor of 10%.

The experimental results confirm that the three-stage architecture proposed in this work underscores the effectiveness and robustness of the proposed methods in performing multi-classes speech disorder identification, demonstrating its adaptability to the complexities of this task of voice disorders classification compared to the related works including different methods.

The developed model in this research serves as a solution for improving pathological speech and addressing the challenges posed by this issue. The main objective of this work is to showcase the resilience of pathological speech classification systems through the preformation of diverse approaches aimed at enhancing voice signal quality and intelligibility via denoising metrics, developing a robust acoustic speech analysis interface using the multi-stream method, and integrating the efficient deep neural architecture CNN-BiLSTM.

References

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *International conference on engineering and technology (ICET)* (pp. 1–6). https://doi.org/10.1109/ICEngTechnol.2017.8308186.

Alhussein, M., & Muhammad, G. (2019). Automatic voice pathology monitoring using parallel deep models for smart healthcare. *IEEE Access, 7*, 46474–46479. https://doi.org/10.1109/ACCESS.2019.2905597

Amara, F., Fezari, M., & Bourouba, H. (2016). An improved GMM-SVM system based on distance metric for voice pathology detection. *An International Journal of Applied Mathematics & Information Sciences, 10*(3), 1061–1070. https://doi.org/10.18576/amis/100324

American Speech-Language-Hearing Association. (1993). Definitions of communication disorders and variations [relevant paper]. Retrieved from https://www.asha.org/policy/rp1993-00208/.

Ankışhan, H., & İnam, S. C. (2021). Voice pathology detection by using the deep network architecture. *Applied Soft Computing, 106*, 107310. https://doi.org/10.1016/j.asoc.2021.107310

Bailly, L., Bernardoni, N. H., Müller, F., Rohlfs, A. K., & Hess, M. (2014). Ventricular-fold dynamics in human phonation. *Journal of Speech, Language, and Hearing Research, 57*(4), 1219–1242. https://doi.org/10.1044/2014_JSLHR-S-12-0418

Behrman, A., Dahl, L. D., Abramson, A. L., & Schutte, H. K. (2003). Anterior-posterior and medial compression of the supraglottis: Signs of nonorganic dysphonia or normal postures? *Journal of*

Voice, 17(3), 403–410. https://doi.org/10.1067/s0892-1997(03)00018-3

Berouti, M., Schwartz, R., & Makoul, J. (1979). Enhancement of speech corrupted by additive noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. https://doi.org/10.1109/ICASSP.1979.1170788

Brijesh Anilbhai, S., & Kinnar, V. (2017). Spectral subtraction and MMSE: A hybrid approach for speech enhancement. *International Research Journal of Engineering and Technology (IRJET), 4*(4), 2340–2343.

Brockmann, M., Drinnan, M. J., Storck, C., & Carding, P. N. (2011). Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. *Journal of Voice, 25*(1), 44–53. https://doi.org/10.1016/j.jvoice.2009.07.002

Brockmann-Bauser, M. (2012). *Improving jitter and shimmer measurements in normal voices*. Phd Thesis of Newcastle University. http://theses.ncl.ac.uk/jspui/handle/10443/1472.

Carding, P., Bos-Clark, M., Fu, S., Gillivan-Murphy, P., Jones, S. M., & Walton, C. (2016). Evaluating the efficacy of voice therapy for functional, organic, and neurological voice disorders. *National Library of Medicine, 42*(2), 201–217. https://doi.org/10.1111/coa.12765

Chaiani, M., Selouani, S. A., Boudraa, M., & Sidi Yakoub, M. (2022). Voice disorder classification using speech enhancement and deep learning models. *Biocybernetics and Biomedical Engineering, 42*, 463–480. https://doi.org/10.1016/j.bbe.2022.03.002

Chung, D. S., Wettroth, C., Hallett, M., & Maurer, C. W. (2018). Functional speech and voice disorders: Case series and literature review. *Movement Disorders Clinical Practices, 5*(3), 312–316. https://doi.org/10.1002/mdc3.12609

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spokens entences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*(4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420

Deli, F., Xuehui, Z., Dandan, C., & Weiping, H. (2022). Pathological voice detection based on phase reconstitution and convolutional neural network. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2022.08.028

Disordered Voice Database. (1994). Version 1.03 (CD-ROM), MEEI, Voice and Speech Lab, Kay Elemetrics Corp, Boston, MA, USA.

Duffy, J. R. (2019). *Motor speech disorders: Substrates, differential diagnosis, and management*, 4th Ed. Retrieved from https://shop.elsevier.com/books/motor-speech-disorders/duffy/978-0-323-53054-5.

El Emary, I. M. M., Fezari, M., & Amara, F. (2014). Towards developing a voice pathologies detection system. *Journal of Communications Technology and Electronics, 59*, 1280–1288. https://doi.org/10.1134/S1064226914110059

Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 33*(2), 443–445. https://doi.org/10.1109/TASSP.1985.1164550

Farhadipour, A., Veisi, H., Asgari, M., & Keyvanrad, M. A. (2018). Dysarthric speaker identification with different degrees of dysarthria severity using deep belief networks. *ETRI Journal, 40*(5), 643–652. https://doi.org/10.4218/etrij.2017-0260

Gholamalinezhad, H., & Khosravi, H. (2020). Pooling methods in deep neural networks, a review. https://doi.org/10.48550/arXiv.2009.07485.

Guedes, V., Teixeira, F., Oliveira, A., Fernandes, J., Silva, L., Junior, A., & Teixeira, J. P. (2019). Transfer learning with audioset to voice pathologies identification in continuous speech. *Procedia Computer Science, 164*, 662–669. https://doi.org/10.1016/j.procs.2019.12.233

Gupta, V. K., Bhowmick, A., Mahesh, C., & Saran, S. N. (2011). Speech enhancement using MMSE estimation and spectral subtraction methods. In *International conference on devices and communications (ICDeCom)* (pp. 1–5). https://doi.org/10.1109/ICDECOM.2011.5738532.

Hamdi, R., Hajji, S., & Cherif, A. (2018). Voice pathology recognition and classification using noise related features. *International Journal of Advanced Computer Science and Applications (IJACSA), 9*(11), 82–87. https://doi.org/10.14569/IJACSA.2018.091112

Hara, K., Saito, D., Shouno, H. (2015). Analysis of function of rectified linear unit used in deep learning. In *International joint conference on neural networks (IJCNN)* (pp. 1–8). https://doi.org/10.1109/IJCNN.2015.7280578.

Harar, P., Alonso-Hernandezy, J. B., Mekyska, J., Galaz, Z., Burget, R., & Smekal, Z. (2019). Voice pathology detection using deep learning: a preliminary study. In *International conference and workshop on bioinspired intelligence (IWOBI)* (pp. 1–4). https://doi.org/10.1109/IWOBI.2017.7985525.

Hossain, M. S., & Muhammad, G. (2016). Healthcare big data voice pathology assessment framework. *IEEE Access, 4*, 7806–7815. https://doi.org/10.1109/ACCESS.2016.2626316

Janbakhshi, P., Kodrasi. I. (2022a). Adversarial-free speaker identity-invariant representation learning for automatic dysarthric speech classification. In *Proceedings of the annual conference of the international speech communication (Interspeech)* (pp. 2138–2142). https://doi.org/10.21437/Interspeech.2022-402.

Janbakhshi, P., Kodrasi. (2022b). Experimental investigation on stft phase representations for deep learning-based dysarthric speech detection. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6477–6481). https://doi.org/10.48550/arXiv.2110.03283.

Jayaraman, D. K., & Das, J. M. (2023). *Dysarthria. StatPearls [Internet]*. StatPearls Publishing.

Joshy, A. A., & Rajan, R. (2021). Automated dysarthria severity classification using deep learning frameworks. In *28th European signal processing conference (EUSIPCO)* (pp. 116–120). https://doi.org/10.23919/Eusipco47968.2020.9287741.

Kadi, K. L., Selouani, S. A., Boudraa, B., & Boudraa, M. (2016). Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. *Biocybernetics and Biomedical Engineering, 36*, 233–247. https://doi.org/10.1016/j.bbe.2015.11.004

Kaladharan, N. (2014). Speech enhancement by spectral subtraction method. *International Journal of Computer Applications, 96*(13), 45–48. https://doi.org/10.5120/16858-6739

Karkos, P. D., & McCormick, M. (2009). The etiology of vocal fold nodules in adults. *Current Opinion in Otolaryngology & Head and Neck Surgery, 17*(6), 420–423. https://doi.org/10.1097/MOO.0b013e328331a7f8

Kent, R. D., & Kim, Y. (2008). Acoustic analysis of speech. In *The handbook of clinical linguistics*(pp. 360–380). https://doi.org/10.1002/9781444301007.ch22

Kim, C., & Stern, R. M. (2016). Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(7), 1315–1329. https://doi.org/10.1109/TASLP.2016.2545928

Kishore, P. (2011). *Speech technology: A practical introduction, topic: spectrogram, cepstrum, and mel frequency analysis*. Retrieved from https://www.cs.brandeis.edu/~cs136a/CS136a_docs/KishorePrahallad_CMU_mfcc.pdf.

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-normalizing neural networks. In *31st conference on neural information processing systems (NIPS)* (pp. 972–981). https://doi.org/10.48550/arXiv.1706.02515.

Ksibi, A., Hakami, N. A., Alturki, N., Asiri, M. M., Zakariah, M., & Ayadi, M. (2023). Voice pathology detection using a two-level classifier based on combined CNN–RNN architecture. *Sustainability, 15*(4), 3204. https://doi.org/10.3390/su15043204

Kulkarni, D. S., Deshmukh, R. R., & Shrishrimal, P. (2016). A review of speech signal enhancement techniques. *International Journal of Computer Applications, 139*(14), 23–26. https://doi.org/10.5120/ijca2016909507

Lee, M. (2023). GELU activation function in deep learning: A comprehensive mathematical analysis and performance. https://doi.org/10.48550/arXiv.2305.12073.

Lim, J. S., & Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE, 12*, 197–210. https://doi.org/10.1109/PROC.1979.11540

Loizou, P. C. (2007). *Speech enhancement: Theory and practice*. CRC Press. https://doi.org/10.1201/9781420015836

Mayle, A., Mou, Z., Bunescu, R., Mirshekarian, S., Xu, L., & Liu, C. (2019). Diagnosing dysarthria with long short-term memory networks. In *Proceedings of the annual conference of the international speech communication (Interspeech)* (pp. 4514–4518). https://doi.org/10.21437/Interspeech.2019-2903.

Mediratta, I., Saha, S., Mathur, S. (2021). LipARELU: ARELU networks aided by Lipschitz Acceleration. In *International joint conference on neural networks (IJCNN)* (pp. 1–8). https://doi.org/10.1109/IJCNN52387.2021.9533853.

Mohammed, H. M. A., Omergolu, A. N., & Oral, E. A. (2023). MMHFNet: Multi-modal and multi-layer hybrid fusion network for voice pathology detection. *Expert Systems and Applications, 223*, 119790. https://doi.org/10.1016/j.eswa.2023.119790

Narendra, N. P., Schuller, B., & Alku, P. (2021). The detection of Parkinson's disease from speech using voice source information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 1925–1936. https://doi.org/10.1109/TASLP.2021.3078364

Peng, X., Xu, H., Liu, J., Wang, J., & He, C. (2023). Voice disorder classification using convolutional neural network based on deep transfer learning. *Scientific Reports, 13*, 7264. https://doi.org/10.1038/s41598-023-34461-9

Picone, J. W. (1993). Signal modeling techniques in speech recognition. *Proceedings of the IEEE, 81*(9), 1215–1247. https://doi.org/10.1109/5.237532

Pouchoulin, G., Fredouille, C., Bonastre, J. F., Ghio, A., & Giovanni, A. (2007). Frequency study for the characterization of the dysphonic voices. In *Interspeech*. https://doi.org/10.21437/Interspeech.2007-386

Pützer, M., & Barry, W. J. (2007). Saarbruecken Voice Database. *Institut für Phonetik. Universität des Saarlandes*. Retrieved from https://stimmdb.coli.uni-saarland.de/help_en.php4.

Shakeel, A. S., Sahidullah, M. D., Fabrice, H., & Slim, O. (2023). Stuttering detection using speaker representations and self-supervised contextual embeddings. *International Journal of Speech Technology, 26*, 521–530. https://doi.org/10.48550/arXiv.2306.00689

Shakeel, A. S., Sahidullah, M. D., Fabrice, H., & Slim, O. (2021). StutterNet: Stuttering detection using time delay neural network. In *29th European signal processing conference (EUSIPCO)* (pp. 426–430). https://doi.org/10.48550/arXiv.2105.05599.

Souissi, N., & Cherif, A. (2015). Dimensionality reduction for voice disorders identification system based on mel frequency cepstral coefficients and support vector machine. In *7th international conference on modelling, identification and control (ICMIC)* (pp. 1–6). https://doi.org/10.1109/ICMIC.2015.7409479.

Souli, S., Amami, R., & Ben Yahia, S. (2021). A robust pathological voices recognition system based on DCNN and scattering transform. *Applied Acoustics, 177*, 107854. https://doi.org/10.1016/j.apacoust.2020.107854

Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM—a tutorial into long short-term memory recurrent neural networks. https://doi.org/10.48550/arXiv.1909.09586.

Strand, O. M., & Egeberg, A. (2004). Cepstral mean and variance normalization in the model domain. In *Proceedings of the COST/ISCA tutorial and research workshop on robustness issues in conversational interaction,* paper 38.

Sumin, K., Chung, W., & Lee, J. (2021). Acoustic full waveform inversion using discrete cosine transform (DCT). *Journal of Seismic Exploration, 30*, 365–380.

Suresh, M., & Thomas, J. (2023). Review on dysarthric speech severity level classification frameworks. In *International conference on control, communication and computing (ICCC)*. https://doi.org/10.1109/ICCC57789.2023.10165636.

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal acoustic analysis jitter, shimmer and hnr parameters. *Procedia Technology, 9*(5), 1112–1122. https://doi.org/10.1016/j.protcy.2013.12.124

Toutounchi, S. J. S., Eydi, M., Ej Golzari, S., Ghaffari, M. R., & Parvizian, N. (2014). Vocal cord paralysis and its etiologies: A prospective study. *Journal of Cardiovascular and Thoracic Research, 6*(1), 47–50. https://doi.org/10.5681/jcvtr.2014.009

Vaiciukynas, E., Gelzinis, A., Verikas, A., & Bacauskiene, M. (2018). Parkinson's disease detection from speech using convolutional neural networks. In *Smart objects and technologies for social good: Third international conference,* (Vol. 233, pp. 206–215). https://doi.org/10.1007/978-3-319-76111-4_21

Vakil, N., van Zanten, S. V., Kahrilas, P., Dent, J., & Jones, R. (2006). The Montreal definition and classification of gastroesophageal reflux disease: A global evidence-based consensus. *The American Journal of Gastroenterology, 101*(8), 1900–1920. https://doi.org/10.1111/j.1572-0241.2006.00630.x

Vásquez-Correa, J. C., Orozco-Arroyave, J. R., & Nöth, E. (2017). Convolutional neural network to model articulation impairments in patients with Parkinson's disease. In *Proceedings of the annual conference of the international speech communication (Interspeech)* (pp. 314–318). https://doi.org/10.21437/Interspeech.2017-1078.

Wang, S. S., Wang, C. T., Lai, C. C., Tsao, Y., & Fang, S. H. (2022). Continuous speech for improved learning pathological voice disorders. *IEEE Open Journal of Engineering in Medicine and Biology, 3*, 25–33. https://doi.org/10.1109/OJEMB.2022.3151233

Westzner, H. F., Schreiber, S., & Amaro, L. (2005). Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders. *Brazilian Journal of Orthinolaryngology, 71*(5), 582–588. https://doi.org/10.1016/s1808-8694(15)31261-1

Wu, H., Soraghan, J., Lowit, A., & Di-Caterina, G. (2018). A deep learning method for pathological voice detection using convolutional deep belief networks. In *Proceedings of the annual conference of the international speech communication (Interspeech)* (pp. 446–450). https://doi.org/10.21437/Interspeech.2018-1351.

Xiaoyu, L. (2018). Deep convolutional and LSTM neural networks for acoustic modelling in automatic speech recognition. Retrieved from https://cs231n.stanford.edu/reports/2017/pdfs/804.pdf.

Xing Luo, O. (2019). Deep learning for speech enhancement- a study on WaveNet, GANs and general RNN architectures. Retrieved from http://www.divaportal.org/smash/get/diva2:1355369/FULLTEXT01.pdf.

Zabret, M., Hočevar Boltežar, I., & Šereg Bahar, M. (2018). The importance of the occupational vocal load for the occurrence and treatment of organic voice disorders. *National Library of Medicine*. https://doi.org/10.2478/sjph-2018-0003

Zhaoyan, Z. (2016). Mechanics of human voice production and control. *The Journal of Acoustical Society of America, 140*(4), 2614–2635. https://doi.org/10.1121/1.4964509

Zhou, C., Wu, Y., Fan, Z., Zhang, X., Wu, D., & Tao, Z. (2022). Gammatone spectral latitude features extraction for pathological voice detection and classification. *Applied Acoustics, 185*(1), 108417. https://doi.org/10.1016/j.apacoust.2021.108417

Zhuge, P., You, H., Wang, H., Zhang, Y., & Du, H. (2016). An analysis of the effects of voice therapy on patients with early vocal fold polyps. *Journal of Voice, 30*, 698–704. https://doi.org/10.1016/j.jvoice.2015.08.013

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.