N^o d'ordre: 274/2025-C/GE

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

Ministère de L'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie HOUARI BOUMEDIENE

Faculté Génie Electrique

THÈSE DE DOCTORAT

Présentée pour l'obtention du grade de **DOCTEUR**

En : Télécommunications

Spécialité : Télécommunication et Traitement de L'information

Par : BELABBAS Soumeya

Thème

Modélisation acoustique multi-variable pour l'évaluation de la compréhension de la parole pathologique

Soutenue publiquement, le **26/02/2025**, devant le jury composé de :

Mme. F. MERAZKA	Professeur	à l'USTHB, Alger	Présidente
M. D. ADDOU	Maître de Conférences/A	à l'USTHB, Alger	Directeur de thèse
Mme. M. TALHA-KEDIR	Professeur	à l'USTHB, Alger	Examinatrice
Mme. T. MERAZI-MEKSEN	Professeur	à l'USTHB, Alger	Examinatrice
M. D. TEGUIG	Maître de Conférences/A	à l'EMP, Bordj el-Bahri	Examinateur

Abstract

Pathological speech disorders significantly hinder effective communication and quality of life. Traditional diagnostic methods, often subjective and time-consuming, limit the accuracy and efficiency of identifying these conditions. This thesis introduces a novel, automated system for classifying pathological speech, and recognition of dysarthric voice, aimed at enhancing diagnostic precision and timeliness. The system employs a comprehensive approach, combining advanced signal processing techniques, robust feature extraction, and cutting-edge deep learning architectures. By integrating multiple acoustic features—including MFCCs, PNCCs, Mel-spectrograms, Jitter, and Shimmer—the system captures detailed information about the speech signal. Furthermore, speech enhancement techniques, such as MMSE-based noise reduction, are applied to improve input quality, especially in noisy environments.

At the core of this system is a deep learning model combining convolutional neural networks (CNNs) with bidirectional long short-term memory (BiLSTM) networks, effectively modeling both local and temporal dependencies within the speech signal. Rigorous evaluation on a diverse dataset of pathological speech samples demonstrated the system's superior performance compared to existing methods. These results underscore the system's potential to transform the diagnosis and management of speech disorders, ultimately contributing to improved patient outcomes and quality of life.

Keywords: Pathological Speech Disorders, Dysarthric Speech Recognition, Speech Classification, Speech Enhancement, Deep Learning Architectures, MFCC, PNCC, Mel-spectrogram, Jitter, Shimmer.

Dedications

I dedicate this work to my family, who have always supported and encouraged me at every step of my journey. To my beloved parents, for their love, patience, sacrifices, and prayers, without which none of this would have been possible. You are my source of inspiration.

To my dear brothers and sisters, for their constant encouragement and unwavering support throughout this work. This achievement is as much yours as it is mine. Thank you for your love and steadfast support.

To my beloved husband, for his understanding, unfailing support, and constant presence, even in the most challenging moments. Thank you for believing in me every single day.

Finally, to all those who accompanied me, near or far, throughout this adventure, for their encouragement and kindness.

Soumeya Belabbas

Acknowledgment

First and foremost, I express my gratitude to God for the strength, courage, patience, and determination He has granted me throughout these years of study. May His guidance continue to accompany me at every stage of my life. I would like to extend my deepest thanks to my research supervisor, **Dr. ADDOU Djamel**, for his trust, patience, and valuable advice. His scientific guidance and insightful feedback have allowed me to progress and refine this work. His technical expertise and unwavering support have been a constant source of inspiration. I am also deeply grateful to my co-supervisor, Professor **SELOUANI Sid-Ahmed**, for his significant contribution to the conception and development of this project. His technical expertise and thoughtful suggestions have greatly enriched my reflections. My sincerest thanks also go to the members of the jury, particularly the president: Professor **MERAZKA Fatiha**, and examiners: Professor **TALHA-KEDIR Malika**, Professor **MERAZI-MEKSEN Thouraya** and Dr. **TEGUIG Djamel**, for the honor of reviewing this work. I am profoundly grateful for their availability and the valuable time they have dedicated to its evaluation. Finally, I deeply thank my family and all those, near and far, for their unwavering support and understanding throughout this journey. Their presence, encouragement, and love have been essential in overcoming moments of doubt and difficulty. My great thanks also go to Mr. Bouali Yacine who helped me a lot, I am deeply grateful to him. Thank you so much. Thank you all.

Contents

List of Figures

Chapter 01: State of the art: speech disorders and automatic speech recognition (ASR) systems for dysarthric voice

Chapter 02: Methodological framework and implementation system design

Chapter 03: Experiments and evaluations

List of Tables

Chapter 01: State of the art: speech disorders and automatic speech recognition (ASR) systems for dysarthric voice

Chapter 03: Experiments and evaluations

Acronyms

Acc	Accuracy
Adam	Adaptive Moment Estimation
AFE	Advanced Front End
ANN	Artificial Neural Network
AReLU	Adaptive Rectified Linear Unit
ASR	Automatic Speech Recognition
BiLSTM	Bidirectional Long Short-Term Memory
CART	Classification and Regression Tree
CER	Character Error Rate
CIFE	Conditional Information Feature Extraction
CMIM	Conditional Mutual Information Maximization
CNN	Convolutional Neural Network
CTC	Connectionist Temporal Classification
DCT	Discrete Cosine Transform
DFA	Detrended Fluctuation Analysis
DISR	Double Input Symmetrical Relevance
DNN	Deep Neural Network
F0	Fundamental Frequency
FFT	Fast Fourier Transform
GAN	Generative Adversarial Network
GELU	Gaussian Error Linear Unit

GRBAS	Grade Roughness Breathiness Asthenia Scale
GRU	Gated Recurrent Unit
GTSL	Gammatone spectral latitude
HMM	Hidden Markov Model
HNR	Harmonics-to-Noise Ratio
HTK	Hidden Markov Model Toolkit
HUPA	Hospital Universitaire Principe de Asturies
IFFT	Inverse Fast Fourier Transform
JMI	Joint Mutual Information
LDA	Linear Discriminant Analysis
LLTSA	Linear Local Tangent Space Alignment
LSTM	Long Short-Term Memory
MDVP	Multi-Dimensional Voice Program
MEEI	Massachusetts Eye and Ear Infirmary
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multilayer Perceptron
MMHFNet	Multi-Modal High-Frequency Network
MMSE	Minimum Mean Square Error
MSE	Mean Squared Error
NB	Naive Bayes
NHR	Noise-to-Harmonics Ratio
PNCC	Power Normalized Cepstral Coefficients
PSD	Power Spectrum Density
ReLU	Rectified Linear Unit

RF	Random Forest
RNN	Recurrent Neural Network
SELU	Scaled Exponential Linear Unit
SinRU	Sinus Rectified Unit
SNR	Signal-to-Noise Ratio
STFT	Short-time Fourier Transform
SVD	Saarbruecken Voice Database
SVM	Support Vector Machine
TDNN	Time-Delay Neural Networks
UASpeech	Universal Access Speech
VAD	Voice Activity Detection
WER	Word Error Rate

General introduction

Speech is a fundamental means of human communication, and any impairment in this ability can profoundly impact an individual's quality of life. Pathological speech, which can arise from various neurological, physiological, or psychological disorders, manifests in diverse ways including dysarthria, dysphonia, and stuttering all of which can significantly affect speech intelligibility and comprehensibility. Early and accurate detection of these disorders is crucial for timely intervention and effective treatment, yet existing diagnostic methods face notable limitations.

Traditional approaches often rely on subjective clinical evaluations, which are not only time-consuming but also prone to variability among practitioners. Objective measures, particularly acoustic analysis, have shown promise as alternatives, though manual analysis remains labor-intensive and requires specialized expertise. These challenges underline the need for a reliable, automated system capable of classifying pathological speech effectively and efficiently.

To address these limitations, this thesis proposes an automated, deep learning-based system for the classification and recognition of pathological speech, aiming to improve both the detection and characterization of voice disorders, and the recognition of dysarthric words. The system leverages a multi-stream approach, combining acoustic features such as MFCCs, PNCCs, Mel-spectrograms, Jitter, and Shimmer to capture comprehensive information about the speech signal within an innovative multi-stream architecture. Additionally, speech enhancement techniques are employed to improve the quality of the input signal, particularly in the presence of noise or distortion. At its core, the system features a CNN-BiLSTM architecture based on convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM) layers allowing for the modeling of complex vocal characteristics. This approach captures detailed spectral and prosodic nuances while enhancing adaptability across various speech pathologies, including vocal cord nodules, paralysis, and polypoid lesions, as well as conditions influenced by factors like gastric reflux and vocal strain.

The methodological framework is validated using the Massachusetts Eye and Ear Infirmary (MEEI) database, a prominent resource in pathological speech research, and the UASpeech database for dysarthric speech. Evaluation spans various aspects, including the impact of feature integration strategies, speech enhancement techniques, and optimized

classification parameters. Comparative analysis with prior systems further highlights the proposed model's effectiveness in achieving high recognition accuracy and computational efficiency.

This thesis aims to achieve several objectives: investigating the effectiveness of acoustic features in characterizing pathological speech and recognition of dysarthric words; examining noise reduction and spectral filtering impacts on recognition accuracy; developing an effective CNN-BiLSTM architecture for pathological speech recognition; and assessing system performance against state-of-the-art methods.

The contributions of this research include:

- **Advanced Feature Extraction:** Integrating multiple acoustic features to provide a comprehensive representation of pathological speech, enhancing recognition accuracy.
- **Robust Speech Enhancement:** Applying speech enhancement techniques to improve input signal quality, leading to more reliable recognition.
- **Powerful Deep Learning Model:** Implementing a CNN-BiLSTM architecture that effectively captures complex speech patterns, outperforming traditional DNN-based approaches.
- **Comprehensive Evaluation:** A rigorous assessment on diverse datasets, demonstrating the system's effectiveness and generalizability.

Thesis Structure

This research aims to revolutionize the diagnosis and management of speech disorders by developing a robust and efficient classification system. The primary goal is to improve patient outcomes by addressing the limitations of traditional diagnostic methods, which often rely on subjective clinical assessments. The thesis is structured as follows:

- **Chapter 1:** Provides an overview of speech production, characteristics of pathological speech, acoustic analysis techniques, and recent advancements in the field.
- **Chapter 2:** Details the data collection process, speech enhancement techniques, feature extraction methods, and the development of the CNN-BiLSTM model.
- **Chapter 3:** Presents the experimental setup, evaluation metrics, and a comprehensive analysis of the system's performance across various datasets.

Finally, the thesis concludes with a discussion of the key findings, the study's limitations, and potential future directions, including exploring advanced deep learning architectures, leveraging larger and more diverse datasets, and developing real-time applications.

Chapter 1

State of the art: speech disorders and automatic speech recognition (ASR) systems for dysarthric voice

1.1 Introduction

Acoustic modeling is a field of study that involves the analysis and representation of speech signals using mathematical models. It plays a crucial role in understanding and classifying pathological speech, which refers to speech that deviates from normal in terms of articulation, fluency, or voice quality. By analyzing the acoustic properties of speech, researchers and clinicians can gain insights into the underlying physiological and neurological processes that contribute to speech disorders.

This chapter will delve deeper into the human vocal apparatus, the different types of speech disorders, assessment methodologies, and the emerging role of deep neural networks in vocal classification. By understanding these fundamental aspects, we can lay a solid foundation for exploring the latest research and clinical applications in the field of speech disorders. We will also discuss the current state-of-the-art in this rapidly evolving field.

1.2 Speech production

The human vocal tract is a complex system of organs responsible for speech production, comprising three main components (Fig. 1.1):

- **Larynx:** Also known as the voice box, the larynx is located at the top of the trachea and houses the vocal folds. These folds are two muscular structures that vibrate to create vocal sounds. The tension and vibration of the vocal folds control the pitch and quality of the voice.
- **Pharynx:** The pharynx is a muscular tube connecting the larynx to the nasal and oral cavities. It plays a key role in shaping the sounds generated by the larynx.
- **Articulators:** The articulators consist of the muscles and structures in the mouth and throat that shape and modify sounds produced by the larynx and pharynx. These include the tongue, lips, teeth, and palate.

Speech is the result of sounds produced through the human phonatory system [1]. This system comprises an exciter (Lungs and Vocal Cords) as the source of the sound and a set of resonators (Vocal Cavities), which modify the sound as it propagates (Fig. 1.1). These cavities provide both articulation points for different sounds and modes of articulation (such as occlusive or fricative) [2].

The production of a speech occurs in three essential phases:

- **Air generation**, involving the lungs, diaphragm, and thoracic muscles;

- **Vibration of this air by the vocal cords**, the larynx. Unvoiced sounds, such as [t], occur when the vocal cords do not vibrate, while voiced sounds, like [d], arise from the vibration of the cords;
- **Resonance of this vibration in the vocal cavities** (pharyngeal, oral, labial, and nasal), which contributes to the timbre of the voice. These cavities acting as acoustic resonators shape the sound making each voice unique. It is in these cavities that the consonants and vowels form.

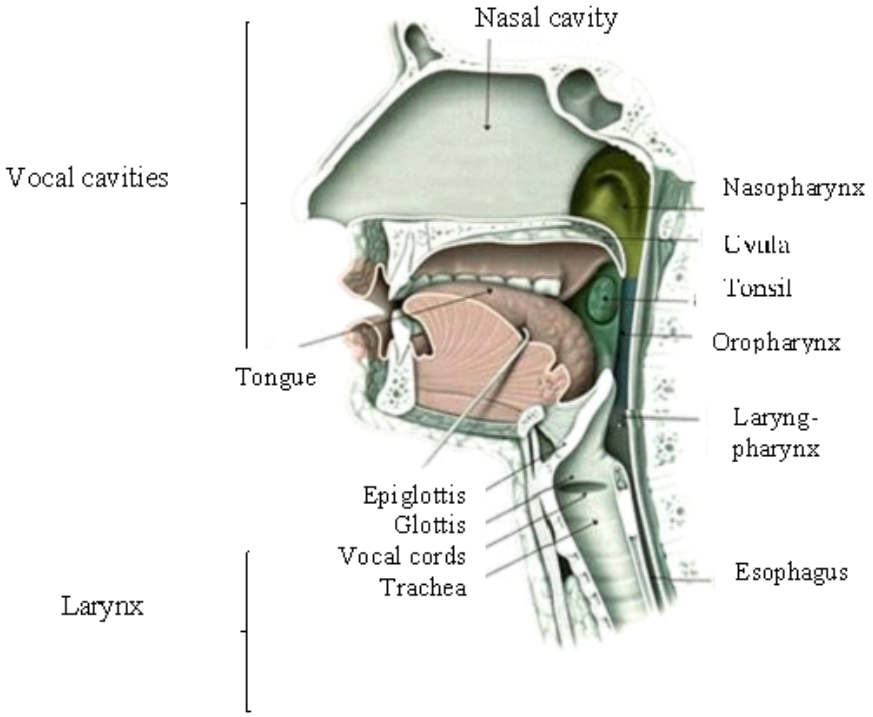

Fig 1.1: Simplified diagram of the human phonatory system [3].

1.2.1 Larynx

The larynx is an unpaired organ [4] located in the anterior part of the neck, above the trachea and in front of the esophagus. It extends the trachea towards the back of the mouth. The cartilages within the larynx articulate through muscles, controlling its opening and closing to facilitate both phonation and respiration. Inside the larynx are two vocal folds that form the "Vocal Cords" (Fig. 1.2) [5].

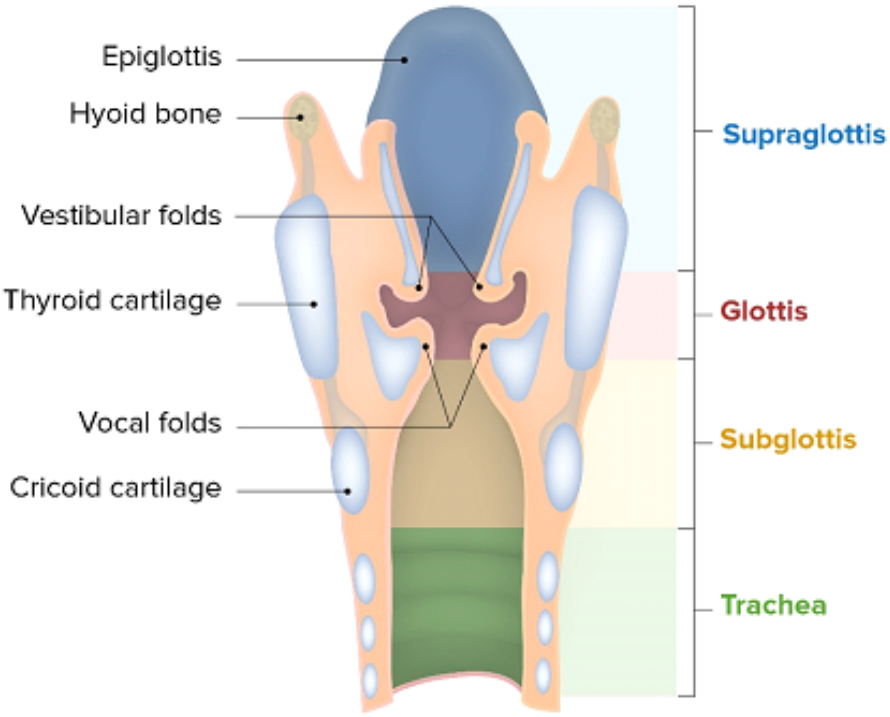

Fig 1.2: Diagram of the larynx.

When these vocal cords vibrate, they produce all voiced sounds, particularly vowels [6]. Air from the lungs, at a specific pressure and flow, passes between the vocal folds, causing them to vibrate and generate what is known as the "laryngeal" vocal sound. Due to the continuous mechanical stress on the vocal folds, this area is prone to pathologies, especially in individuals with professions involving heavy voice use (e.g., teachers, singers). Conditions such as dysphonia may develop early (around 35–40 years old), and lifestyle factors like smoking and alcohol consumption can exacerbate these issues, potentially leading to more severe conditions, including cancer.

1.2.2 Vocal cavities

The sound emerging from the larynx is not yet fully formed speech. To become comprehensible speech, it must pass through the supraglottic vocal cavities (pharyngeal, oral, labial, and nasal), which modify the sound into vowels and consonants. These cavities take on specific shapes based on the movements of the lower jaw, lips, and tongue (Fig. 1.1). The uvula or velum, a muscular portion between the pharyngeal and nasal cavities, opens to allow air to escape through the nose for nasal sounds [7]. The oral cavity contains several key articulation points: the tongue, teeth, alveoli, hard palate and

soft palate. These areas allow for the differentiation of various consonants and contribute to the overall clarity and distinctiveness of speech sounds.

1.3 Pathological speech characteristics

The voice, like any other part of the human body, can be affected by various pathologies. These are disorders characterized by difficulties in producing clear, well-articulated and modulated sounds. Speech disorder. Speech disorders can impact the ability to produce sequences of syllables that form understandable words, ultimately leading to communication difficulties.
Speech disorders can be broadly categorized into two main groups based on their underlying cause [8]:

- **Functional or neurological disorders:** These disorders result from improper or excessive voice use without underlying anatomical abnormalities. They are often linked to behavioral, emotional, or psychological factors; Examples include:
  - Dysphonia
  - Aphonia
  - Dysarthria associated with neurological disorders like Parkinson's or Alzheimer's disease
- **Organic or physiological disorders:** These disorders are caused by physical abnormalities or diseases affecting the vocal apparatus. These abnormalities may can be congenital (present at birth), caused by disease, infection, injury, or trauma. Examples include:
  - Vocal fold nodules
  - Polyps
  - Laryngeal cancer

Here are some of the most common types of pathological speech:

1.3.1 Dysarthria

Dysarthria is a neuromotor speech disorder resulting from damage or dysfunction in the brain or spinal cord areas responsible for controlling speech production [9, 10]. This damage can affect nerves or neural centers, leading to abnormalities in:

- Speed
- Strength
- Precision
- Amplitude (volume)
- Tone
- Duration

These abnormalities manifest as difficulties in controlling the tongue, vocal cords, and respiratory muscles. This can negatively impact articulation, voice quality, and speech rate. Importantly, individuals with dysarthria typically retain the content of their spoken language, allowing them to communicate through writing and understand both spoken and written language.

1.3.2 Dysphonia

Dysphonia is characterized by an alteration in vocal quality, affecting clarity, volume, pitch, or resonance of the voice. It can manifest in several ways, including a hoarse, husky, weak, or shaky voice. In severe cases, complete voice loss (aphonia) may occur. An example of dysphonia is recurrent vocal fold paralysis, where one vocal cord remains paralyzed.

1.3.3 Dyslalia

Dyslalia is a speech articulation disorder characterized by difficulties in correctly producing specific sounds or sound groups in speech. This can include substitutions, omissions, distortions, or additions of sounds in speech. While dyslalia is common in young children learning to speak, it can persist in some individuals if left untreated. Dyslalia can have functional or organic origins:

- **Functional origins:** Stuttering and sigmatism are examples of functional dyslalia.
- **Organic origins:** Cleft palate is an example of organic dyslalia.

1.3.3.1 Stuttering

Stuttering is a fluency disorder characterized by involuntary interruptions in the flow of speech. These interruptions can include sound repetitions, prolongations, blocks, or inappropriate pauses. Stuttering can make communication difficult and frustrating for individuals affected.

1.3.3.2 Sigmatism

Sigmatism refers to the incorrect articulation of consonants, particularly fricatives (sounds produced by forcing air through a narrow channel). It's one of the most common speech articulation disorders in children, as this sound requires precise in articulation. Different types of sigmatism exist, depending on the cause:

- **Nasal sigmatism:** Incorrect tongue position prevents air from exiting through the mouth.
- **Dorsal sigmatism:** Excessive tongue raising during sound production.
- **Occlusive sigmatism:** Systematic replacement of fricatives with closest-sounding occlusive consonants (e.g., replacing "s" with "t").
- **Voiced/Unvoiced confusion:** Non-vibration of vocal cords in voiced consonants, often seen in children with hearing impairments.

1.3.3.3 Cleft Palates

Cleft palates, also known as cleft lips and palates, are congenital malformations characterized by openings or splits in the upper lip, the palate, or both. These malformations occur during early embryonic development. Cleft lip results from the failure of facial tissue to fuse properly, causing a gap in the upper lip. Cleft palate refers to a gap in the roof of the mouth, creating a communication between the nose and mouth. Cleft palates can significantly impact making it difficult to produce labial consonants (cleft lips) and oral sounds due to nasalization (cleft palates).
Cleft palate refers to a gap in the roof of the mouth, creating an communication between the nose and mouth. Cleft palates can significantly impact articulation, making it difficult to produce labial consonants (cleft lip) and oral sounds due to nasalization (cleft palate).

1.3.4 Dysprosody

Dysprosody refers to the weakening or complete loss of prosodic features (rhythm, intonation and voice timbre). Speech with dysprosody lacks melody, has a monotonous tone, and may be delivered very slow.

1.3.5 Parkinson's disease

Parkinson's disease is a chronic neurodegenerative condition that primarily affects the motor system. It is caused by the progressive loss of dopamine-producing neurons in

a region of the brain called the substantia nigra. This loss leads to a variety of motor symptoms, including:

- **Tremor:** Shaking of a limb or other body part.
- **Rigidity:** Stiffness and resistance to movement.
- **Bradykinesia:** Slowness of movement.
- **Postural instability:** Difficulty maintaining balance.

In addition to motor symptoms, Parkinson's disease can also cause non-motor symptoms, such as:

- Speech and swallowing difficulties
- Cognitive impairment
- Depression
- Anxiety
- Sleep disturbances

While there is no cure for Parkinson's disease, treatments are available to manage symptoms and improve quality of life [11, 12].

1.3.6 Total laryngectomy

Total laryngectomy is a surgical procedure that removes the larynx, often used to treat advanced laryngeal cancer. It results in the loss of natural voice and requires individuals to learn alternative communication methods, such as esophageal speech or electrolarynx.

1.4 Acoustic modeling fundamentals

1.4.1 Definition and principles

Acoustic modeling mathematically represents speech signals for analysis, understanding, and classification. It acts as a bridge between raw audio signals and higher-level linguistic information in speech processing. Acoustic models capture distinctive speech sound features, enabling systems to recognize and classify speech, including pathological speech, which deviates from normal due to disorders.

The primary goal is to accurately represent speech production and its acoustic manifestations. This involves modeling vocal tract sound generation and analyzing how speech pathologies affect these sounds. By understanding both typical and atypical speech acoustics, acoustic models assist in diagnosing and assessing speech disorders.

1.4.2 Key features of pathological speech

Acoustic signals possess key features crucial for analyzing both healthy and pathological speech. These include:

- **Pitch (Fundamental frequency)**: Vocal fold vibration frequency. Pathologies can disrupt pitch, leading to deviations like monotone or irregular pitch.
- **Intensity (Loudness)**: Signal energy perceived as loudness. Vocal cord dysfunction can weaken intensity, while other conditions may cause excessive loudness.
- **Formant frequencies**: Resonant frequencies defining vowel sounds (F1 and F2 are crucial). Pathological speech often exhibits abnormal formants due to articulation or vocal tract issues.
- **Duration and rhythm**: Speech disorders can alter rhythm, pauses, or prolonged sounds, impacting temporal aspects. These features are essential for identifying dysfluencies or prosodic issues.

1.4.3 Common acoustic analysis techniques

To analyze these key features, various acoustic analysis techniques are employed [13]:

- **Spectrograms:** Visual representations of frequency content over time. They reveal disruptions in normal frequency patterns, irregular voicing, or abnormal resonance in pathological speech.
- **Cepstral analysis:** Separates source (vocal cords) and filter (vocal tract) components. MFCCs are widely used for speech recognition and effectively distinguish subtle speech differences.
- **Linear predictive coding (LPC):** Models the vocal tract as a series of filters and represents formants. LPC analyzes acoustic properties of pathological speech, providing insights into vocal tract shaping [14].

These techniques form the basis of acoustic modeling, enabling detailed analysis and classification of pathological speech. They are indispensable tools for clinicians and researchers working to improve diagnostic accuracy and develop assistive technologies for individuals with speech impairments.

1.4.4 Integration of acoustic analysis techniques with speech pathology knowledge

Acoustic modeling, as discussed previously, involves the mathematical representation and analysis of speech signals. When integrated with speech pathology knowledge, it provides a powerful tool for understanding and assessing speech disorders. By combining the technical capabilities of acoustic analysis with the clinical expertise of speech pathologists, researchers can gain deeper insights into the underlying causes of speech impairments and develop more effective diagnostic and therapeutic approaches [15].
We have explored various applications of acoustic modeling in speech pathology:

- **Identifying specific acoustic features:** Acoustic modeling can identify unique acoustic features associated with different speech disorders. For example, studies have shown that individuals with stuttering exhibit specific patterns of dysfluencies, such as repetitions and prolongations, which can be quantified using acoustic analysis techniques.
- **Quantifying the severity of speech impairments:** Acoustic modeling can help quantify the severity of speech impairments by measuring objective parameters such as pitch variability, intensity range, and formant dispersion.
- **Monitoring treatment progress:** Acoustic modeling can be employed to monitor the effectiveness of speech therapy interventions by comparing pre- and post-treatment acoustic measurements.
- **Predicting speech outcomes:** Acoustic modeling can be used to predict future speech outcomes, such as the likelihood of speech recovery or the need for ongoing therapy.

1.5 Recent advances in acoustic modeling for voice disorder diagnosis and classification

Pathological speech classification refers to an automatic speech processing system that categorizes and labels the speech of individuals suffering from voice disorders. This research area and its applications in vocal technology hold significant clinical importance [16].

Several tools and methods have been developed to assist speakers with voice pathologies. Notable progress has been made in improving speech intelligibility and automatically evaluating voice disorders based on various modern acoustic modeling techniques.

Among the most recent and significant works that have had the greatest impact on scientific research in the field of automated speech processing for speakers with voice disorders from various perspectives, we can cite the following:

- In [17], the authors developed a voice disorder classification system using a two-stage structure. The first stage involved a speech enhancement technique based on the minimum mean square error (MMSE). In the second stage, a CNN-LSTM network, equipped with the SinRU activation function, was implemented.
- A new method called deep multimodal and multi-layer hybrid fusion network (MMHFNet) was introduced in [18] to extract deep features. The authors conducted experiments using a deep learning algorithm based on the LSTM model, utilizing the Saarbruecken SVD vocal database with complete and balanced samples.
- The study described in [19] aims to introduce a hybrid feature vector defined as input and a multi-model composed of CNN and LSTM to diagnose various voice disorders and increase classification accuracy. Two types of fusion models (feature fusion and decision-level fusion) are used to enhance the classification accuracy of the multi-models. Experimental results showed that both fusion models successfully classified pathological data.
- In [20], a multi-model architecture is also proposed, which is a coupled machine learning algorithm (CNN-RNN) used for the classification of healthy and pathological audio samples, along with a two-level cascade architecture that enables the accurate identification of pathological voices from the input dataset by incorporating gender (male or female) information and manually extracted features.
- A study in [21] introduced a new approach to categorize four voice disorders (functional dysphonia, neoplasm, phonotrauma, and vocal paralysis). Instead of using a single vowel, this approach employs continuous speech in Mandarin. The researchers first transformed the acoustic data into Mel frequency cepstral coefficients, then used a bidirectional long short-term memory (BiLSTM) network to capture the sequential characteristics of the signal. Experimental results demonstrated that this proposed framework produced significant improvements in classification accuracy compared to systems that only use a single vowel.
- Gammatone spectral latitude (GTSL) coefficients were proposed in [22] to improve the performance of the voice classification system for healthy, neuromuscular, and structural voice disorders. The proposed features, based on traditional machine learning methods, achieved an average accuracy of 99.6% using the Massachusetts Eye and Ear Infirmary (MEEI) database. The accuracies in other databases, such as

Saarbruecken Voice Database (SVD) and Hospital Universitario Principe de Asturias (HUPA), were 89.9% and 97.4%, respectively.

- In [23], researchers proposed a transfer learning framework, OpenL3-SVM, pre-trained for automatic multi-class voice disorder recognition. The framework combines a pre-trained OpenL3 convolutional neural network with a support vector machine (SVM) classifier. The Mel spectrum of the given vocal signal is first extracted and then entered into the OpenL3 network to obtain high-level feature integration. Given the effects of high-dimensional redundant and negative features, model overfitting occurs easily. Therefore, the linear local tangent space alignment (LLTSA) method is used to reduce feature dimensionality. Finally, the obtained dimensionality-reduced features are used to train the SVM for voice disorder classification. A 5-fold cross-validation is used to verify classification performance. Tests on the VOICED dataset demonstrated that the proposed method achieved values of 99.46%, 99.64%, 98.92%, and 99.64% for ACC, SEN, SPE, and F1 metrics, respectively.

1.6 Evaluation of intelligibility and comprehension of pathological speech

Intelligibility is a crucial indicator of speech impairment in pathological speech, often used by speech therapists for diagnosis and therapy assessment [24, 25]. Intelligibility refers to the accuracy with which a listener can recover a speaker's acoustic signal [26]. This inherently involves human listener participation, whether experts (speech therapists) or naïve listeners (students). Common intelligibility measurement procedures involve listeners assessing different vocal materials (e.g., word lists, sentences, passages) using various methods (e.g., visual analog scale, orthographic transcription, multiple-choice tasks).

Comprehension refers to a listener's ability to interpret the meaning of a spoken message, regardless of phonetic or lexical accuracy. Evaluating pathological speech comprehension determines a person's ability to understand others' speech, crucial for diagnosing specific disorders, planning therapeutic interventions, and tracking patient progress.

Evaluating intelligibility and comprehension of pathological speech relies on a combination of subjective, objective, and automated techniques. Each approach has advantages and limitations, and combining them allows for a more comprehensive and accurate evaluation. Technological advancements continue to enhance these methods, offering new possibilities for better evaluation and treatment of speech disorders with more precise

and efficient tools. Automatic speech recognition, computer-assisted acoustic analysis, and machine learning techniques provide more objective evaluations of intelligibility. For speech comprehension, technologies like eye-tracking, computer-assisted behavioral responses, and functional neuroimaging offer detailed insights into comprehension abilities and underlying brain processes. These innovations pave the way for more precise diagnoses and better-targeted therapeutic interventions.

We have explored various approaches for evaluating vocal quality:

1.6.1 Perceptual evaluation

Based on human judgment and the listener's ability to assess voice quality. In designing a perceptual evaluation protocol, the selection of the jury is crucial. Jury quality is assessed in terms of reliability, which refers to the reproducibility of judgments between listeners (inter-listener variability) and by the same listener during multiple listening sessions (intra-listener variability). To improve reliability, several listening sessions are organized, during which voices are presented in random order. Context control is essential as it influences perception: a moderately dysphonic voice may appear more impaired after a normal voice than after a severely dysphonic voice. Conducted under properly controlled experimental conditions, perceptual analysis is easy to implement, accessible to any clinician, and inexpensive. However, it carries several intrinsic biases that render it imperfect or even insufficient. Many factors influencing perceptual judgment cannot be entirely controlled, including the listener's emotional state during evaluation, their aesthetic values, native language and/or dialect, and how they interpret the measurement scale, among others. In practice, a perceptual analysis of voice quality is performed by a panel of specialists (voice therapists, phoniatrists, and speech therapists), who provide an analytical description of vocal characteristics using the GRBAS scale [27]. The GRBAS is a perceptual scale based on the evaluation of five acoustic parameters: overall grade of dysphonia (Grade), degree of voice roughness (Roughness), breathiness (Breathiness), vocal weakness or asthenia (Aesthenia), and vocal strain (Strain). Each parameter is rated on a four-point scale of severity: 0 (normal), 1 (slightly impaired), 2 (impaired), and 3 (severely impaired). This scale can be applied during the production of a sustained vowel, a sentence, or a typically read text.

1.6.2 Instrumental evaluation

Instrumental analysis is designed to qualify and, most importantly, quantify dysphonias through acoustic and/or aerodynamic measurements. These measurements are most

frequently performed on a sustained vowel, typically /a/, using various sensors designed to record and study multiple parameters of speech production. Often, it is necessary to combine different complementary measurements to account for the multidimensional aspect of vocal production. Acoustic measurements (e.g., frequency and amplitude, jitter and shimmer, spectral analysis) reveal audible characteristics of dysphonia, including primarily fundamental frequency and intensity measurements, their stability, and the analysis of the emitted sound's spectrum. Aerodynamic measurements, while not direct measurements of the voice, help assess the biomechanical characteristics of the pneumo-phonatory system. These include primarily airflow, pressure, and glottal efficiency measurements. In the evaluation of dysphonias, the multiparametric instrumental approach, made possible by the existence of reliable expert tools, offers a complementary approach to perceptual analysis, which is inherently considered the "gold standard" since voice is first and foremost a perceptual phenomenon, both from the speaker's perspective (the speaking subject) and from the evaluator's perspective (the clinician as well as the speaker's surroundings).

1.7 Advancements in dysarthric speech recognition systems ASR

Improving speech recognition for individuals with dysarthria is a crucial area of research, addressing the challenges posed by the complex acoustic and articulatory variations of dysarthric speech. Developing effective rehabilitation strategies for dysarthric speakers is essential to enhance their communication skills and promote their social integration.

Currently, automatic speech recognition (ASR) emerges as one of the most promising tools in this field. Table. 1.1 summarizes the approaches and results achieved by various researchers over the past decade [28].

Table 1.1: Progress in dysarthric speech recognition: a decade of research.

Author	**Methodology**	**Results**
Rudzicz [29]	Noise reduction-based transformation, consonant desonorization, tempo morphing, and frequency morphing	Increase in recognition accuracy from 72.7% to 87.9%

Shahamiri et Salim [30]	Proposed a multi-network speech recognizer for dysarthria (DM-NSR) using a multi-view, multi-learner approach, called multi-network artificial neural networks.	Improvement in accuracy by 24.67% compared to a single-network speech recognizer.
Takashima et al. [31]	Use of the pre-trained Convolutional Bottleneck Network (CBN) to extract acoustic features from dysarthric speech and train the ASR. The use of a Convolutional Restricted Boltzmann Machine for pre-training and to prevent overfitting.	Better word recognition performance compared to the convolutional network without pre-training.
Bhat et al. [32]	Comparison of the performance of different acoustic feature parameters in SRP based on DNN-HMM and GMM-HMM.	Use of incremental data effectively improving the accuracy of SRP for dysarthric speech.
Kim et al. [33]	Use of KL-HMM to model the variations in dysarthric speech and speaker adaptation based on "L2-norm" regularization.	Reduction of speaker confusion and better recognition performance compared to the traditional method.

Yu et al. [34]	Acoustic models based on TDNN, LSTM-RNN, and advanced variants, with learning of hidden unit contributions (LHUC) to adapt to acoustic variation, and semi-supervised complementary autoencoder.	Overall word recognition accuracy of 69.4% on the test set of 16 UASpeech speakers.
Xiong et al. [35]	Non-linear approach to modify speech rate, with two methods: modifying dysarthric speech to bring it closer to typical speech, and modifying typical speech to bring it closer to dysarthric speech.	Improvement of absolute accuracy by nearly 7% with the second approach, more effective for speakers with moderate to severe dysarthria, tested on the UASpeech database.
Wu et al. [36]	Contrastive learning framework to capture the acoustic variations of dysarthric speech, with exploration of data augmentation strategies.	Robust recognition results for dysarthric speech.
Zaidi et al. [37]	Use of DNN and integration of CNN and LSTM to improve ASR accuracy for dysarthric speech.	Improvement in accuracy by 11% and 32% compared to the LSTM and GMM-HMM models, respectively

Shahamiri [38]	Dysarthria-specific ASR system (Speech Vision) learning to recognize the shape of words spoken by dysarthric speakers, with visual data augmentation.	Improvement in accuracy by 67%, particularly for severely dysarthric speech.
Hu et al. [39]	Application of Neural Architecture Search (NAS) machine learning to optimize the hyperparameters of Time-Delay Neural Networks (TDNN-Fs).	Improvement of ASR performance for dysarthric speech on the UASpeech database.
Almadhor, et al. [40]	Spatio-temporal ASR system based on Spatial CNN and multi-head attention Transformer to visually extract the acoustic features from dysarthric speech.	The best word recognition accuracy achieved to 90.75%, 61.52%, 69.98% and 36.91% of low level dysarthric speech, mild level dysarthric speech, high level dysarthric speech and very high level dysarthric speech, respectively.

Wang et al. [41]	Comparative study of various data augmentation approaches to improve the robustness of pre-trained ASR model fine-tuning to dysarthric speech. These include: a) conventional speaker-independent perturbation of impaired speech; b) speaker-dependent speed perturbation, or GAN-based adversarial perturbation of normal, control speech based on their time alignment against parallel dysarthric speech; c) novel Spectral basis GAN-based adversarial data augmentation operating on non-parallel data.	Experiments conducted on the UASpeech corpus suggest GAN-based data augmentation consistently outperforms fine-tuned Wav2vec2.0 and HuBERT models using no data augmentation and speed perturbation across different data expansion. After cross-system outputs rescoring, the best system produced the lowest published WER of 16.53% (46.47% on very low intelligibility) on UASpeech.

1.8 Conclusion

This chapter has laid out the theoretical foundations of acoustic modeling for the assessment of pathological speech. It covered key concepts such as the crucial role of acoustic modeling in understanding and classifying pathological speech, the diverse acoustic features associated with different speech disorders, recent technological advancements that enhance voice disorder diagnosis and classification, and the ongoing challenges and opportunities in evaluating intelligibility and comprehension of pathological speech. By integrating acoustic modeling with the expertise of speech pathology, researchers and clinicians can deepen their understanding of speech impairments, enabling the development of more accurate diagnostic tools and more effective therapeutic interventions. As technology continues to advance, we can expect further breakthroughs in acoustic modeling, expanding its applications in speech pathology and improving the quality of life for

individuals with speech disorders.

Chapter 2

Methodological framework and implementation system design

2.1 Introduction

Pathological voice classification can be framed as a pattern recognition problem. This perspective enables us to leverage established methods and algorithms from the field of pattern recognition to achieve our goal. By treating speech signals as patterns, we can develop systems that take these signals as input and output the corresponding pathology category.

In this chapter, we will present the detailed steps involved in constructing such a system and explore various approaches to optimize its performance. We will discuss the methodological framework that guides our research and the implementation strategies we employ to build a robust and effective classification system.

2.2 Data collection and preprocessing

The speech signal serves as the raw material for analysis in a pathological voice recognition system. The speech signal database must be collected efficiently, considering equipment and recording environment. These factors are crucial for the application, requiring high-quality signals.

Utilizing a standard database is essential for recognized progress in speech signal processing and its applications, including speaker recognition and pathological voice classification [42]. Standard data enables comparison of results and assessment of the proposed method's effectiveness relative to state-of-the-art techniques. In our application, we used the MEEI database.

To improve the characteristics of pathological speech and bring them closer to those of healthy speech, we applied several techniques to enhance the quality and intelligibility of the speech signal. These techniques will be detailed in the following paragraphs.

2.2.1 MEEI database: description and characteristics

The Massachusetts Eye and Ear Infirmary (MEEI) database [43] is a widely used resource in pathological speech analysis. Developed by the MEEI Voice and Speech Lab, it contains over 1400 audio samples of sustained vowel /a/ and the first part of the Rainbow Passage. Commercialized by Kay Elemetrics, it was recorded in two different environments. Normal samples were recorded at 50 kHz, while pathological samples were recorded at either 25 kHz or 50 kHz.

Despite its widespread use, the MEEI database has limitations, such as the varying environments and sampling frequencies for normal and pathological voices.

The database includes assessments of voice state using stroboscopy, acoustic aerodynamic measurements, and physical examinations of the neck and mouth (provided by Kay Elemetrics). By observing changes in vocal muscles, many vocal pathologies can be studied.

For our analysis, we filtered the CD's file names to include only samples with three specific diseases. Files with multiple pathologies or missing MDVP parameters were excluded. We selected 53 available samples from normal speakers.

2.2.2 Techniques for enhancing pathological speech

Speech enhancement refers to algorithms that improve quality, reduce listening fatigue from noisy speech, increase intelligibility, and enhance voice communication system performance [44]. While "noise reduction" has a broader scope, speech enhancement specifically focuses on detecting and removing unwanted noise in audio to improve clarity, intelligibility, or the overall listening experience. The flowchart in Fig. 2.1 illustrates the basic steps of a speech enhancement system including:

- **Noise estimation:** Estimating the noise power spectrum density (PSD) from the noisy speech signal.
- **Speech presence probability:** Determining the probability of speech presence in each frequency bin using a three-step decision gain factor.
- **Gain computation:** Calculating the gain function based on the estimated noise PSD and speech presence probability.
- **Filtering:** Applying the gain function to the noisy speech signal to obtain the enhanced speech signal.
- **Smoothing:** Applying optimal smoothing techniques to reduce noise artifacts and improve the quality of the enhanced speech signal.

Intelligibility is a key aspect of speech quality. High-quality speech ensures good intelligibility, while unintelligible speech is not considered high quality. Speech enhancement has been applied to improve the robustness of pathological speech recognition systems, as a positive correlation exists between intelligibility scores of pathological and noisy speeches [46].

Various speech enhancement techniques, such as spectral subtraction, Wiener filtering, MMSE, and wavelet-based denoising, have been used to improve the classification of vocal pathologies.

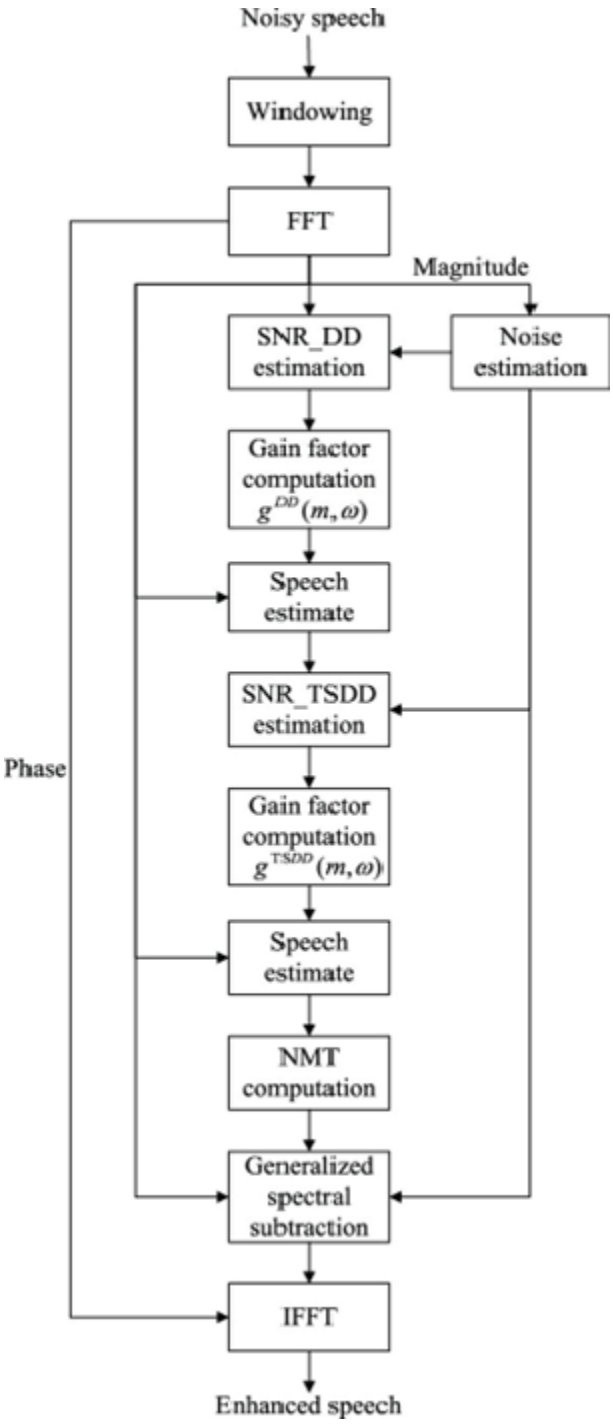

Fig 2.1: Basic steps of a speech enhancement system with three-step decision gain factor and optimal smoothing [45].

Spectral subtraction is a traditional method for enhancing speech degraded by stationary additive background noise [47, 48]. It's a non-parametric method requiring noise spectrum estimation. A common issue with spectral subtraction is insufficient noise attenuation during silent periods.

Wiener filtering [49] is an alternative to spectral subtraction. It's a linear filter that recovers the original speech signal from noisy signals by minimizing mean squared error (MSE) between estimated and original signals. Wavelet-based denoising [50–52] decomposes noisy signals into wavelets and applies thresholding to suppress noise. The wavelet transform decomposes the signal into sub-bands, and noise reduction is performed through hard or soft thresholding. A drawback of this method is its tendency to distort some useful components of the original speech.

Other methods adapt the statistical model of a recognition module to identify noise characteristics before speech enhancement. One such method is the minimum mean squared error (MMSE) technique, which uses a statistical model to estimate and minimize errors between the actual speech signal and the estimated signal [53].

In this work, we used three enhancement methods spectral subtraction, Wiener filtering, and the MMSE-based enhancer to conduct our experiments and compare the results to improve the performance of the pathological speech classification system.

2.2.2.1 Spectral subtraction

Spectral subtraction is one of the earliest and most significant methods for speech enhancement, primarily due to its simplicity. It was introduced in the late 1970s by Boll [48] and later generalized and improved by Berouti [54].

a) Boll's Algorithm

Let $x(t)$ be a noisy signal, composed of a clean signal $s(t)$, corrupted by additive noise $n(t)$, which is uncorrelated with $s(t)$. The noisy signal is expressed as:

$$x(t) = s(t) + n(t) \tag{2.1}$$

The enhanced signal $\hat{s}(t)$ is an estimate of $s(t)$. Assuming the speech signal is quasi-stationary over analysis windows of 20 to 30 ms, the short-time Fourier transform (STFT) of $x(t)$ is given by:

$$X(\omega) = S(\omega) + N(\omega) \tag{2.2}$$

where ω is the angular frequency, and $X(\omega)$, $S(\omega)$ et $N(\omega)$ denote the spectra of $x(t)$, $s(t)$ et $n(t)$ respectively. The spectrum $X(\omega)$ can be expressed in polar form as follows:

$$X(\omega) = |X(\omega)|e^{i\Phi x(\omega)} \tag{2.3}$$

Here, $|X(\omega)|$ and $\Phi x(\omega)$ represent the amplitude and phase of $X(\omega)$, respectively. The noise spectrum can similarly be expressed as $N(\omega) = |N(\omega)|e^{i\Phi n(\omega)}$, where $|N(\omega)|$ is

unknown and is replaced by its average value, estimated during the period of no vocal activity.

Spectral subtraction assumes that noise does not affect the signal's phase, so only the short-term spectral amplitude of the noise is considered. The estimated clean signal spectrum is:

$$\hat{S}(\omega) = \left[|X(\omega)| - |\widetilde{N}(\omega)|\right] e^{i\Phi x(\omega)} \tag{2.4}$$

where $|\widetilde{N}(\omega)|$ is the estimated amplitude spectrum of the noise. The enhanced signal is obtained by computing the inverse Fourier transform of $\hat{S}(\omega)$.

As the amplitude $|\hat{S}(\omega)| = |X(\omega)| - |\widetilde{N}(\omega)|$ may become negative due to noise estimation errors, a common solution is half-wave rectification:

$$|\hat{S}(\omega)| = \begin{cases} |X(\omega)| - |\widetilde{N}(\omega)| & \text{if } |X(\omega)| > |\widetilde{N}(\omega)| \\ 0 & \text{if } |X(\omega)| \leq |\widetilde{N}(\omega)| \end{cases} \tag{2.5}$$

The amplitude spectral subtraction algorithm (2.5) can be extended to power spectral subtraction by multiplying $X(\omega)$ by its conjugate $X^*(\omega)$, we have :

$$|X(\omega)|^2 = |S(\omega)|^2 + |N(\omega)|^2 + 2.Re\left\{S^*(\omega)N(\omega)\right\} \tag{2.6}$$

Assuming that $n(t)$ and $s(t)$ are zero-mean and uncorrelated, the terms $E\left\{S(\omega)N^*(\omega)\right\}$ and $E\left\{S^*(\omega)N(\omega)\right\}$ become zero. Therfore, the estimated power spectrum of the clean signal is:

$$|\hat{S}(\omega)|^2 = |X(\omega)|^2 - |\widetilde{N}(\omega)|^2 \tag{2.7}$$

The power spectral subtraction algorithm can be expressed as:

$$|\hat{S}(\omega)|^2 = H^2(\omega)|X(\omega)|^2 \tag{2.8}$$

Where the supression gain $H(\omega)$ [43] is given by:

$$H(\omega) = \sqrt{1 - \frac{|\widetilde{N}(\omega)|^2}{|X(\omega)|^2}} \tag{2.9}$$

b) Berouti's Algorithm

The Berouti algorithm extends spectral subtraction by considering the subtraction of both amplitude and power spectra, or more generally, any power of the short-term amplitude spectrum. Let P_x, P_s, and P_n represent the power spectra of the estimated clean signal, the noisy signal, and the noise signal, respectively, Berouti introduced two parameters

into the spectral subtraction estimator:

$$\hat{P}_x = (P_s^\gamma - \alpha P_n^\gamma)^{1/\gamma} \tag{2.10}$$

The parameter α controls the overestimation of the noise spectrum, while γ controls the power of the spectrum before subtraction. For α values between 3 and 6, and with $\gamma = 1$, the algorithm corresponds to power subtraction; for $\gamma = 2$, it corresponds to amplitude subtraction.

2.2.2.2 Wiener filtering

A training model of the convolution type is considered, expressed as:

$$x(t) = (h * s)(t) + n(t) \tag{2.11}$$

Where $x(t)$ is the observed signal, $h(t)$ is the system's impulse response, and $n(t)$ is the observation noise. The goal is to estimate $s(t)$ using a linear of the form:

$$\hat{s}(t) = (\omega * x)(t) \tag{2.12}$$

To minimize the mean squared error J, we aim to minimize:

$$J = E\left[(s(t) - \hat{s})^2\right] \tag{2.13}$$

This leads to the classical wiener filter expression in the frequency domain [44]:

$$W(\omega) = \frac{H^*(\omega)\Gamma_s(\omega)}{|H(\omega)|^2\Gamma_s(\omega) + \Gamma_n(\omega)} \tag{2.14}$$

Where $\Gamma_s(\omega)$ and $\Gamma_n(\omega)$ represent the power spectral densities of the signal and noise, respectively.

2.2.2.3 Enhancement technique based on MMSE

The minimum mean squared error (MMSE) algorithm is a widely adopted technique for speech enhancement. Its primary goal is to minimize the mean squared error (MMSE) by estimating the clean speech signal using a gain function [55]:

$$\hat{X}_{(n,k)} = S_{(n,k)}.G_{(n,k)} \tag{2.15}$$

The a priori $\varepsilon_{(n,k)}$ and the a posteriori SNR $\gamma_{(n,k)}$ are used to calculate the gain function $G_{(\varepsilon,\nu)}$, which attenuates noise while preserving the speech signal:

$$G_{(\varepsilon,\nu)} = \frac{1}{1+\varepsilon_{(n,k)}}.exp\left\{2\int \gamma_{(n,k)} udu\right\} \tag{2.16}$$

The MMSE algorithm in our system employs voice activity detection (VAD) for noise estimation, providing accurate noise reduction while maintaining speech quality.

2.3 Multi-stream framework for speech analysis

The speech signal is characterized by a high degree of redundancy. However, it is essential to note that not all information contained in the speech signal is considered useful for specific applications. Therefore, effective speech processing must focus on extracting relevant information pertinent to the intended application. In the field of pathological speech processing, the quality of the voice as perceived by a listener can be affected by various diseases. Acoustic analysis serves as the most useful tool for diagnosing such conditions. This approach involves representing the acoustic signal through parameters or features specifically designed for automatic speech recognition and classification applications. The main objective is to extract relevant information from the speech signal while excluding non-informative parts, as the inclusion of such parts can lead to memory overload and deteriorate recognition and classification performance.

Cepstral coefficients are commonly used in automatic speech recognition due to their ability to effectively represent the signal, particularly by decorrelating features. However, they suffer from several limitations. In this work, we hypothesize that the vocal signal under pathological conditions shares similarities with speech signals recorded in noisy environments or those with added noise.

Cepstral coefficients are sensitive to signal acquisition conditions and the surrounding acoustic environment, leading to a robustness issue that can degrade the performance of speech recognition systems. To improve performance in the domain of pathological speech recognition, many studies have explored the use of multiple parameters derived from independent sources. The multivariable approach processes information from different perspectives, allowing it to be represented in multiple ways. Therefore, integrating multiple sources of information using a multivariable technique has the potential to enhance the robustness and performance of speech recognition systems.

Multivariable acoustic analysis involves the fusion of several sources of information, addressing the shortcomings of traditional methods that rely on a single source. Informa-

tion that may be lost during the extraction of a specific set of parameters can be recovered through another acoustic analysis. This method is effective only if the various parameters provide complementary, diverse information about the speech characteristics. Integrating highly correlated information is less likely to offer any benefit, as redundant information from different features does not enhance the system's performance.

In [56], new acoustic parameters were introduced to better describe the vocal signal, assisting in classifying pathological voices. These parameters include the fundamental frequency (F0), Harmonics-to-Noise Ratio (HNR), Noise-to-Harmonics Ratio (NHR), and Detrended Fluctuation Analysis (DFA). Classification was performed on two pathological databases, Saarbruecken and MEEI Voice, using HTK classifiers. Two types of classification were conducted: binary classification for normal and pathological voices, and a four-category classification for male and female speakers with conditions such as spasmodic dysphonia, polyps, nodules, and normal voice. The effects of these parameters, when combined with MFCC coefficients, Delta, Delta-Delta, and Energy, were studied, and the results demonstrated the effectiveness of this approach.

Another study [57] investigated acoustic parameters and feature selection methods to improve the classification of dysarthric speech. Depending on the severity of the impairment, four types of acoustic features—prosodic, spectral, cepstral, and vocal quality—were used, along with seven feature selection methods: Interaction Cap (ICAP), Conditional Information Feature Extraction (CIFE), Conditional Mutual Information Maximization (CMIM), Double Input Symmetrical Relevance (DISR), Joint Mutual Information (JMI), Conditional Redundancy (Condred), and Relief. Six classification algorithms were employed: Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), Classification and Regression Tree (CART), Naive Bayes (NB), and Random Forest (RF). The results demonstrated that the best classification accuracy was achieved when all prosodic acoustic features of dysarthric speech were combined.

By exploiting various speech preprocessing techniques and combining different parameters based on human auditory perception into a single data vector, we propose a new acoustic front-end that characterizes pathological speech signals. This front-end offers the advantages of computational simplicity, feasibility, and improved performance. The parameters include cepstral coefficients (MFCC, PNCC, Mel-spectrogram), parameters that reflect frequency and amplitude variations (Jitter and Shimmer), and the prosodic parameter F0.

2.3.1 Extraction of acoustic parameters

The extraction of acoustic parameters involves two primary steps:

Step1: Formatting of the speech signal

This step prepares the speech signal for subsequent analysis by applying signal processing operations.

- **Sampling:** The analog speech signal is digitized by sampling and quantizing each sample. According to Shannon's theorem, information loss is minimized if the sampling frequency (f_e) is at least twice the maximum frequency (f_{max}) of the signal.
- **Pre-emphasis:** A high-pass filter is applied to amplify high frequencies. This first-order filter has the transfer function:

$$(z) = 1 - \alpha z^{-1} \tag{2.17}$$

  Where α is a weighting coefficient between 0.9 and 1.
- **Frame segmentation:** The continuous speech signal is divided into short, overlapping frames (approximately 20–30 ms with 10–15 ms overlap).
- **Windowing:** A weighting window (often Hamming) is applied to reduce edge effects caused by segmentation. The Hamming window is given by:

$$h(n) = \begin{cases} 0.54 - 0.46 * \cos\left(2\pi \dfrac{n}{N-1}\right) & \text{if} \quad 0 \leq n \leq N-1 \\ 0 & \text{else} \end{cases} \tag{2.18}$$

  where N is the window size in samples.

Step2: Calculation of acoustic coefficients

Once the speech signal is formatted, acoustic coefficients are extracted. These coefficients represent various characteristics of the speech signal, such as pitch, intensity, and spectral features.

2.3.1.1 Mel-frequency cepstral coefficients (MFCC)

Mel-frequency cepstral coefficients (MFCCs) are widely used in speech processing for feature extraction and classification [58]. They are particularly effective for tasks involving human speech, as they are based on the human auditory system's perception of sound. MFCCs are calculated using the following step (Fig. 2.2):

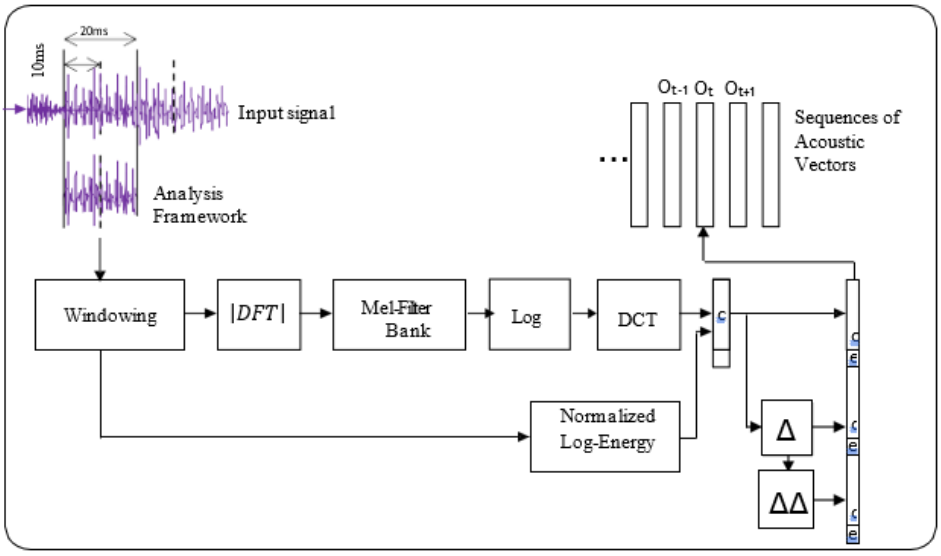

Fig 2.2: Acoustic analysis module using MFCC representation.

1. **Fourier transform:** The speech signal is transformed into the frequency domain using the Fourier transform.
2. **Mel-frequency filtering:** The spectrum is filtered using a bank of Mel-frequency filters, which are triangular filters spaced logarithmically in frequency.
3. **Logarithm:** The energy of each filter bank is calculated and converted to the logarithmic domain.
4. **Discrete cosine transform (DCT):** A DCT is applied to the logarithm of the filter bank energies to obtain the MFCC coefficients.
5. **Cepstral coefficients:** The first few MFCC coefficients (typically 12-13) are retained, representing the most important spectral characteristics of the speech signal.

The Mel scale is a perceptual scale that approximates the human auditory system's response to frequency. It is more linear at low frequencies and more logarithmic at high frequencies, reflecting how humans perceive pitch.

Dynamic features can be added to MFCCs by calculating their first and second derivatives (delta and delta-delta coefficients). These features capture the temporal dynamics of the speech signal and can be useful for tasks such as speaker identification and emotion recognition.

In summary, MFCCs are a powerful feature extraction technique that has been successfully applied to various speech processing tasks, including pathological speech classification. They provide a compact and informative representation of the spectral characteristics of speech signals, making them well-suited for machine learning algorithms.

2.3.1.2 Mel-spectrogram coefficients

Mel-spectrogram coefficients are representations of the variation in signal frequencies over time [59], where the frequencies are transformed according to the Mel scale, which is the perceptual scale of frequencies based on human hearing techniques. The MFCCs and Mel-spectrograms follow the same calculation steps, except for the integration of the DCT transformation in the MFCC coefficients [60].
The Mel-spectrogram is calculated using the following steps (Fig. 2.3):

1. **Short-time fourier transform (STFT):** The speech signal is divided into short frames and the Fourier transform is applied to each frame to obtain the spectrum.
2. **Mel-frequency filtering:** The spectrum is filtered using a bank of Mel-frequency filters, which are triangular filters spaced logarithmically in frequency.
3. **Logarithm:** The energy of each filter bank is calculated and converted to the logarithmic domain.
4. **Normalization:** The log-mel spectrum is often normalized to account for variations in overall signal energy.

Mel-spectrograms provide a visual representation of the spectral characteristics of a speech signal over time. They are commonly used for tasks such as speech recognition, speaker identification, and audio analysis.

Mel-spectrograms can be used directly as features for machine learning models, or they can be further processed to extract Mel-frequency cepstral coefficients (MFCCs). MFCCs are a more compact representation of the Mel-spectrogram, and they are often used for tasks that require a lower-dimensional feature space.

In summary, Mel-spectrogram coefficients are a powerful tool for analyzing speech signals. They provide a rich representation of the spectral and temporal characteristics of speech, making them suitable for a variety of applications.

2.3.1.3 Power normalized cepstral coefficients (PNCC)

Power Normalized Cepstral Coefficients (PNCC) are another type of acoustic feature that can be extracted from speech signals. They are similar to MFCCs but offer certain advantages in terms of robustness and discriminative power. Experimental results in [61] demonstrate that PNCC feature vectors provide substantial improvements in recognition accuracy in the presence of various types of noise and in reverberant environments, with a slightly higher computational cost than conventional MFCC processing. PNCC processing also offers better recognition accuracy in noisy environments compared to techniques such

as Vector Taylor Series (VTS) [62] and ETSI Advanced Front End (AFE) [63], while requiring much less computation.

The calculation of PNCCs involves the following steps (Fig. 2.3):

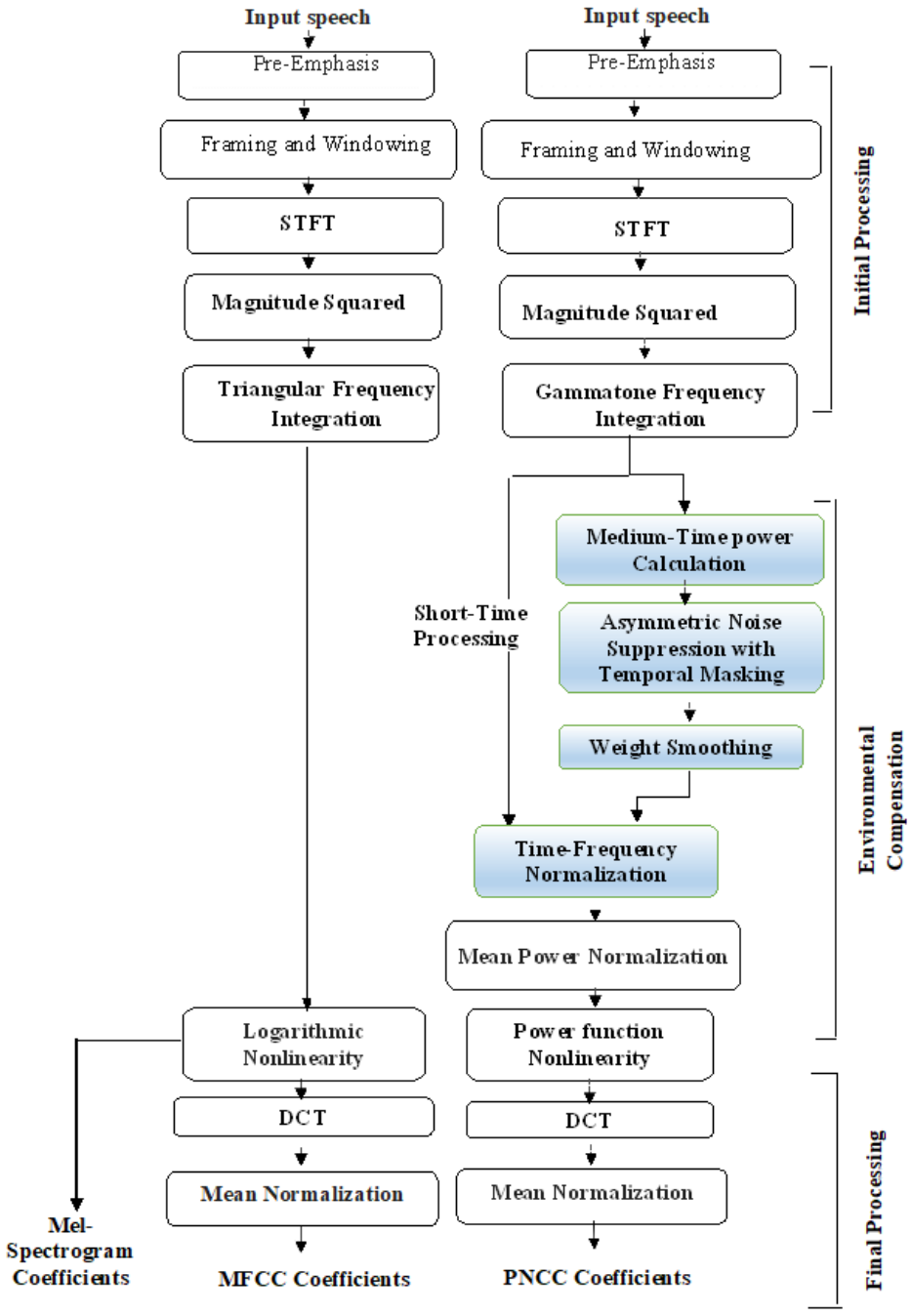

Fig 2.3: Acoustic analysis module using MFCC representation.

1. **Short-time fourier transform (STFT):** The speech signal is divided into short frames and the Fourier transform is applied to each frame to obtain the spectrum.
2. **Power spectrum:** The magnitude squared of the spectrum is calculated to obtain the power spectrum.
3. **Cepstral analysis:** The cepstrum is obtained by applying the inverse Fourier transform to the logarithm of the power spectrum.
4. **Normalization:** The cepstral coefficients are normalized to account for variations in overall signal energy.

PNCCs are normalized to have zero mean and unit variance, which can improve their robustness to variations in speaker characteristics and recording conditions. They also tend to be more discriminative than MFCCs for certain tasks, such as speaker identification and emotion recognition. PNCCs can be used as features for machine learning models, or they can be combined with other features, such as MFCCs and delta coefficients, to improve the performance of classification tasks.

Fig. 2.3 represents the structure of the PNCC approach compared to MFCCs and Mel-spectrograms. As seen in the figure, the initial processing steps of PNCC are similar to those in MFCC and Mel-spectrogram analysis. A pre-emphasis filter and a short-time Fourier transform (STFT) using Hamming windows are performed. However, frequency analysis is conducted using gammatone filters [64], as shown in Fig. 2.4.

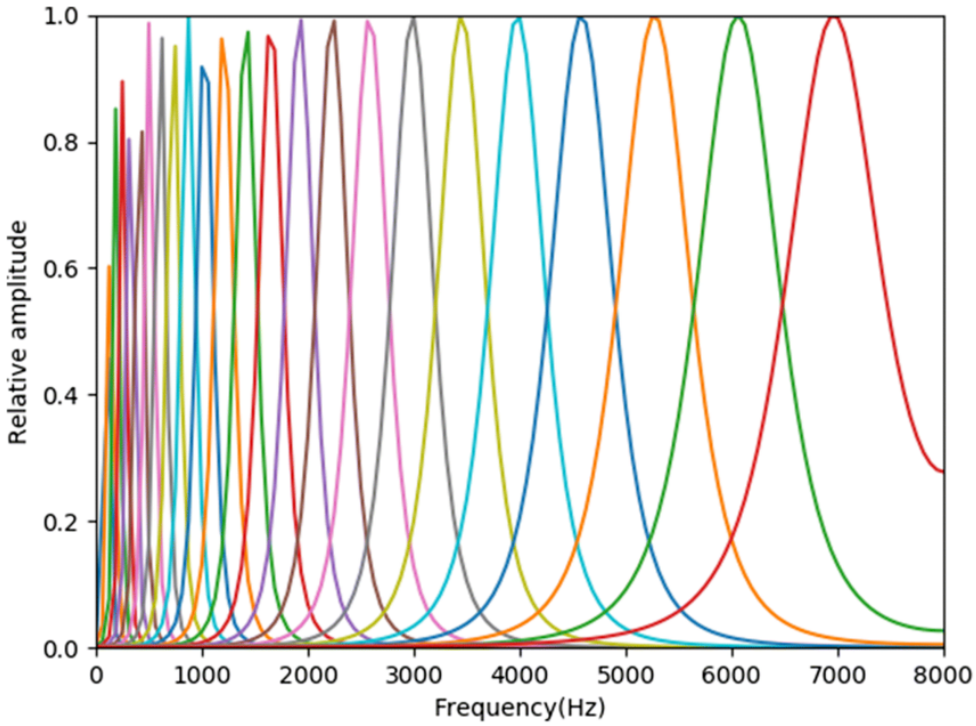

Fig 2.4: Representation of the frequency response of a gammatone filter bank.

This is followed by a series of nonlinear operations performed using longer-term temporal analysis, which provides better performance for noise modeling since the power associated with most background noise conditions changes more slowly than the instantaneous power associated with speech. Subsequently, noise suppression is applied through an asymmetric approach using a nonlinear asymmetric filter. The output of the filter (the detected noise) is subtracted from the input (the average power), resulting in a noise-free signal.

The combined effects of asymmetric noise suppression and temporal masking are represented for a specific time interval and frequency bin as a transfer function. The smoothing of the transfer function across frequency is achieved by calculating the mean operation over the channel index of this transfer function.

In PNCC processing, the power-law nonlinearity results in a processing response that is affected by changes in absolute power. To minimize the impact of amplitude scaling in PNCC, an average power normalization step is invoked. The final steps of PNCC processing are also similar to MFCC processing, except for the carefully chosen power-law nonlinearity with an exponent of 1/15 [61], which provides a reasonably good fit to physiological data while optimizing recognition accuracy in the presence of noise.

In summary, PNCCs are a valuable alternative to MFCCs for acoustic feature extraction. They offer certain advantages in terms of robustness and discriminative power,

making them suitable for a variety of speech processing applications.

2.3.2 Prosodic features

Prosodic parameters, such as fundamental frequency (F0), jitter, and shimmer, play a crucial role in the acoustic analysis of pathological speech. They allow for the early detection of vocal disorders by identifying subtle anomalies in the voice. These measurements quantify the severity of pathologies, helping to diagnose and differentiate specific vocal conditions. They are also essential for assessing the effectiveness of treatments and monitoring patient progress over time. Furthermore, prosodic analysis enables the personalization of therapeutic interventions and deepens the understanding of underlying pathological mechanisms.

2.3.2.1 Fundamental frequency (F0)

The fundamental frequency, denoted as F0, represents the number of vibrations per second of the vocal cords. It occurs because, when pronouncing certain sounds, such as [b], [d], or [z], the vocal cords vibrate at a specific quasi-periodic frequency. This acoustic parameter, also known as pitch, typically varies from 80 to 200 Hz for male voices, 150 to 450 Hz for female voices, and 350 to 600 Hz for children's voices [65]. F0 is generally measured in Hertz (Hz) and serves as a key indicator for analyzing intonation, voice tone, and speech melody.

2.3.2.2 Perturbation of F0 (Jitter)

Jitter represents the cycle-to-cycle variation of F0 within a frame of the signal. It is calculated as the average difference in F0 between two consecutive cycles of vibrations (Fig. 2.5). Jitter is primarily affected by insufficient control of the vibration of the vocal cords [66].

Jitter values can be measured according to different parameters, such as relative absolute perturbation and relative average perturbation (RAP) and period perturbation quotient (PPQ5) [67].

Absolute jitter is the cycle-to-cycle variation in the fundamental frequency, which is the average absolute difference between consecutive periods, expressed as follows:

$$Jitter(Absolu) = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \tag{2.19}$$

Where T_i represents the length of the extracted glottal periods and N is the number of these extracted glottal periods.

Relative jitter, or local jitter, is the average absolute difference between consecutive periods, divided by the mean period, and expressed as a percentage:

$$Jitter(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100 \tag{2.20}$$

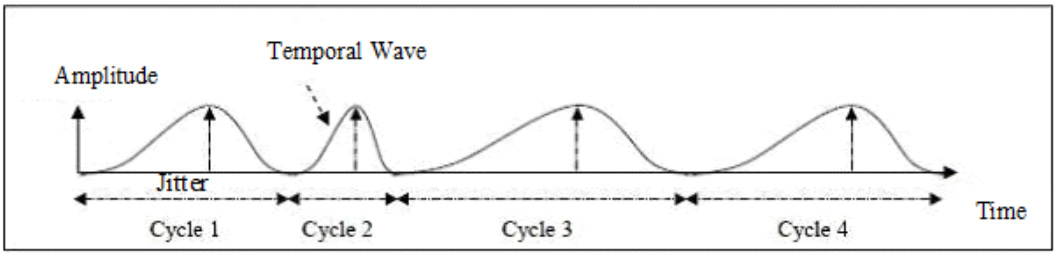

Fig 2.5: Example of fundamental frequency periodicity.

In the case of vocal pathology, jitter increases, and a value greater than 1.04% corresponds to pathological speech [68]. In other words, the normal/pathology threshold is set at 1.04%.

2.3.2.3 Intensity perturbation (Shimmer)

Shimmer refers to the cycle-to-cycle variation in intensity within a frame of the signal. It is calculated as the average difference in amplitude between two consecutive cycles of vibrations (Fig. 2.6) [69]. Similar to the Jitter factor, the Shimmer factor is a good indicator for exploring the stability of intensity. It allows the normalization of the mean Shimmer by comparing it to the average amplitude.

The mean absolute Shimmer, expressed in dB, represents the variability of peak-to-peak amplitude in decibels [70]. It is the average absolute logarithm base 10 of the difference between the amplitude of consecutive periods, multiplied by 20:

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20 * log\left(\frac{A_{i+1}}{A_i}\right)\right| \tag{2.21}$$

Where A_i represents the extracted peak-to-peak amplitude data and N is the number of extracted fundamental frequency periods.

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

$$Shimmer(Relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100 \tag{2.22}$$

In the case of voice pathology, Shimmer increases, and a value greater than 3.81% corresponds to pathological speech. In decibels, this threshold corresponds to 0.35 dB.

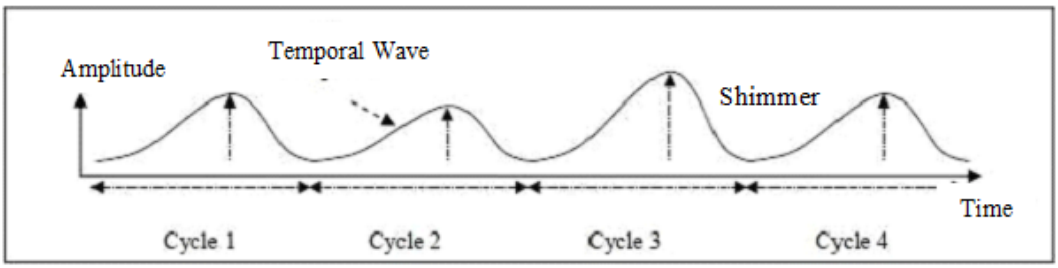

Fig 2.6: Example of signal amplitude instability.

2.4 Normalization of parameters

Normalization is a crucial step in ensuring the robustness of extracted MFCC and PNCC parameters. It reduces sensitivity to speaker variability and recording conditions. In our application, we used variance normalization to scale coefficients to a standard range. This ensures coefficients with greater variances do not dominate the feature vector. The normalized coefficient $\hat{c}_i$ is calculated as follows [71]:

$$\hat{c}_i = \frac{c_i - u_i}{\sigma_i} \tag{2.23}$$

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

This standard preprocessing technique is widely adopted for tasks such as speech recognition, speech classification, and speaker identification.

2.5 Deep learning techniques

Deep learning is a type of machine learning that mimics how humans acquire certain types of knowledge. It's a crucial component of data science, encompassing statistics and predictive modeling. Deep learning is particularly useful for scientists working with large amounts of data, as it can accelerate and simplify the processes of collection, analysis, and interpretation.

Artificial neural networks (ANNs) are software and/or hardware systems that function similarly to neurons in the human brain. They are a variety of deep learning technologies. Typically, a neural network relies on a large number of processors operating in parallel and organized into layers. The first layer receives raw input information, while subsequent layers receive output information from the previous layer. The final layer produces the system's results.

Deep Neural Networks (DNNs) are multilayer perceptrons (MLPs) with more than three layers as depicted in Fig. 2.7 [72,73]. The general structure of an MLP is organized as follows:

- **Input layer:** Receives input data.
- **Hidden layers:** Intermediate layers that apply activation functions to weighted sums of inputs [74].
- **Output layer:** Produces the final output, with the number of neurons depending on the task.

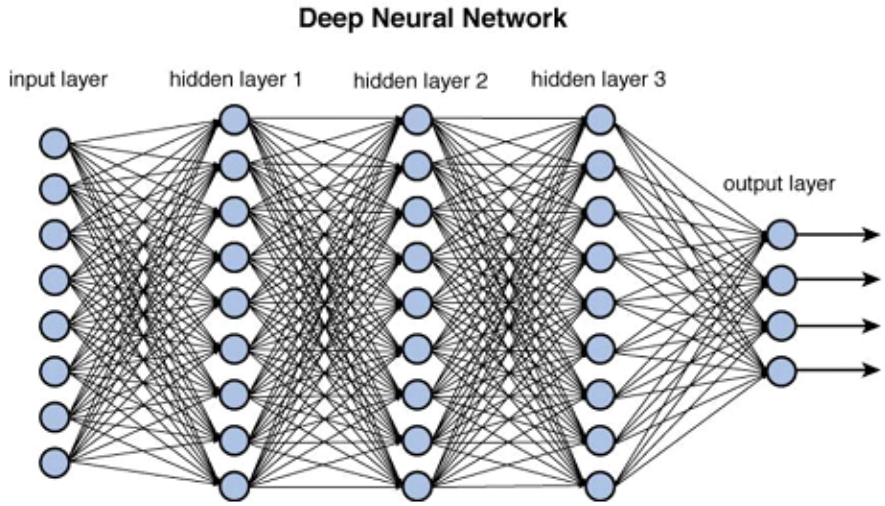

Fig 2.7: Deep network architecture with multiple layers.

MLPs use a supervised learning technique called backpropagation for training:

- **Forward propagation:** Data passes through the network, with each neuron applying an activation function to the weighted sum of its inputs [75].
- **Backpropagation:** The error between predicted and actual output is calculated, and weights are adjusted to minimize this error.

Various deep learning techniques can be applied to pathological speech classification. In our application, we proposed the following approaches:

2.5.1 Convolutional neural networks (CNN)

CNNs are a type of ANN where connections are arranged to perform a convolution operation. They are inspired by the visual cortex and are used in various fields, including computer vision, natural language processing, and recommendation systems [76]. CNNs implicitly perform feature extraction through convolutional layers.

- **General architecture and functioning:** A CNN consists of convolutional layers, pooling layers, activation functions, and fully connected layers [77].
    - **Convolutional layer:** Extracts local features by applying filters to the input data.

- **Pooling layer:** Reduces the spatial dimensions of features while retaining important information (e.g., max pooling).
- **ReLU layer:** Introduces nonlinearity using the ReLU activation function.
- **Fully connected layer:** Transforms features into a one-dimensional vector for classification or regression [78].

- **Advantages for pathological voice analysis:** Pathological Voice Analysis involves identifying and characterizing anomalies in the voice caused by vocal disorders or pathologies. Convolutional Neural Networks (CNNs) offer several advantages in this specific field, particularly for the detection, diagnosis, and classification of vocal disorders. Here are some key benefits:
  - **Automatic extraction of pathology-specific features:** CNNs can automatically extract complex acoustic features from pathological voice recordings, such as irregularities in frequencies, disturbances in the sound spectrum, and variations in amplitude. This allows for the detection of subtle signs of vocal pathologies.
  - **Consideration of local variations in the signal:** CNNs effectively capture local variations in the temporal representations of vocal signals. Pathological anomalies can manifest as irregularities at specific moments in the signal, and CNNs are well-suited to detect these localized variations.
  - **Invariance to irrelevant variations:** CNNs are invariant to local translations, which is beneficial when pathological features appear at different temporal positions in the vocal signal. This allows the model to focus on essential aspects of the pathologies, regardless of minor irrelevant variations.
  - **Reduction of model complexity:** CNNs reduce model complexity by using shared filters, which is particularly useful when training data is limited, as is often the case with pathological voice databases. This helps avoid overfitting while maintaining high performance.
  - **Robustness to artifacts and noise:** Pathological voice recordings may contain artifacts or noise due to capturing equipment or suboptimal recording conditions. CNNs can filter out noise and focus on relevant features for pathology, increasing diagnostic reliability.
  - **Enhanced computer-aided diagnostic tools:** Integrating CNNs into computer-aided diagnostic (CAD) systems for pathological voice analysis can improve the accuracy and speed of diagnoses, providing clinicians with powerful tools to detect and monitor vocal disorders.

These advantages make CNNs a particularly effective choice for pathological voice analysis, contributing to more accurate diagnoses, a better understanding of vocal disorders, and more tailored treatments.

2.5.2 Bidirectional long short-term memory networks (BiLSTM)

A Bidirectional Long Short-Term Memory (BiLSTM) network is an advanced architecture of recurrent neural networks (RNNs) that processes sequences of data by considering information from both the past and the future within an input sequence. This makes it particularly well-suited for speech recognition tasks, where the context of a spoken word can significantly impact its interpretation [79].

- **General architecture and functioning:** This BiLSTM network consists of LSTM cells, which are the fundamental building blocks of the network. Each LSTM cell includes an input gate, a forget gate, an output gate, and a cell state (memory cell). These gates modify the cell state, allowing it to carry information through long sequences without losing relevance, as illustrated in Fig. 2.8.

  The BiLSTM network captures both short-term and long-term contextual dependencies of the input sequence (both preceding and succeeding), making it particularly powerful for various sequence processing tasks. According to the equations below, the forward LSTM processes the input sequence in the order from the first to the last instance, producing an output sequence $(\overrightarrow{h}_1, \overrightarrow{h}_2, \ldots, \overrightarrow{h}_t)$. Meanwhile, the backward LSTM does the opposite by reversing the order, producing an output sequence $(\overleftarrow{h}_1, \overleftarrow{h}_2, \ldots, \overleftarrow{h}_t)$ The forward and backward paths operate in parallel, with each maintaining separate weights and biases.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{2.24}$$

$$g_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{2.25}$$

$$c_t = f_t \circ c_{t_1} + i_t \circ g_t \tag{2.26}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{2.27}$$

$$h_t = o_t \circ \tanh(c_t) \tag{2.28}$$

$$z_t = softmax(W_{hz}h_t + b_z) \tag{2.29}$$

  Where W_{xi} and W_{hi} are weight matrices, b_i are bias vectors, and $\circ$ denotes the element-wise product. Similarly, i_t is the output of the forget gate, which deter-

mines which information from the previous cell state should be forgotten and retained. The forget gate addresses the vanishing gradient problem by ensuring that the problematic component of the product contains elements close to one [80].

Here, f_t is the output of the input gate, which determines the important information from c_{t1} and the input modulation after their multiplication. The activation cell, as given by equation (2.26), updates the cell state by adding the important information from the input data and the retained information from the previous cell state. Finally, o_t is the output of the output gate, which controls the hidden state of the next cell h_t, as defined by equation (2.28).

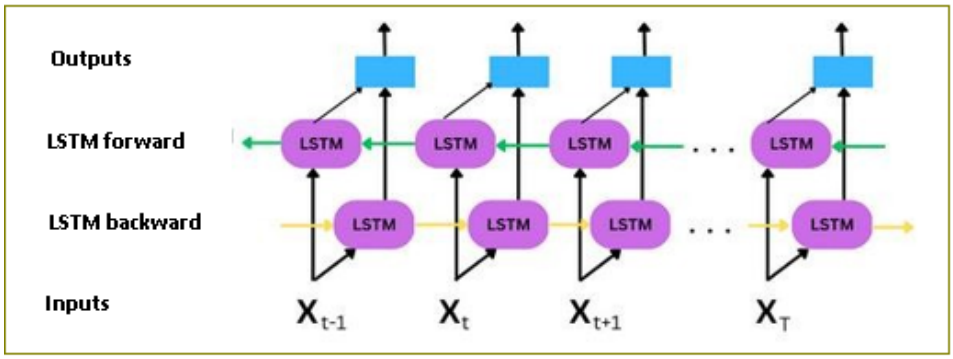

Fig 2.8: Architecture of a bidirectional LSTM layer.

- **Advantages for temporal analysis of pathological speech:** BiLSTM networks offer several significant advantages for the temporal analysis of speech. Their ability to capture temporal dependencies in both directions before and after a given point in a sequence makes them particularly well-suited for this complex task. Here are some of the main advantages of BiLSTM networks for temporal speech analysis:
  - **Capture of long-term dependencies in both directions:** BiLSTMs analyze vocal data while considering both the past and future context of a sequence. This is crucial for pathological speech analysis, where anomalies may depend not only on preceding sounds but also on those that follow. This comprehensive view allows for a better understanding of subtle changes in speech, such as irregularities in rhythm, tone, or intensity.
  - **Robustness to speech variations:** Vocal pathologies can lead to unpredictable variations in rhythm, tone, or voice quality. BiLSTMs are capable of handling these variations by utilizing the complete context to stabilize the analysis and accurately interpret anomalies.

- **Improved diagnostic accuracy:** By analyzing sequences in both directions, BiLSTMs enhance diagnostic accuracy by reducing the risks of false positives (detecting non- pathological anomalies) and false negatives (failing to detect actual pathologies).
- **Optimization of detection and analysis systems:** BiLSTMs can automate the process of detecting and analyzing vocal pathologies, providing powerful tools for clinicians and researchers while reducing the need for intensive manual analysis.

2.5.3 Hybrid CNN-BiLSTM architecture

The proposed approach in our study is based on a combination of two connectionist models, namely the Convolutional Neural Network (CNN) and the Bidirectional Long Short-Term Memory (BiLSTM) network. The aim is to leverage the advantages of both networks. This hybrid model is extensively studied for the application of speech recognition systems. In [81], the hybrid model CNN-BiLSTM enhances the system performance compared to those using CNN and HMM architectures [82].

- **Integration and functioning of the hybrid CNN-BiLSTM model:** In the proposed model, based on a hybrid BiLSTM and CNN network for the detection of vocal pathologies, the different layers of the network have specific roles in processing the input data for the relevant classification of pathological speech. Here is a brief overview of the role of each layer:
  - **Input layer:** The input layer receives the data in image format of speech features.
  - **Convolutional layer:** The convolutional layer applies a set of filters to the input features to extract relevant local characteristics for classification. The filters are learned through the training of the network.
  - **BiLSTM layer:** The bidirectional Long Short-Term Memory (BiLSTM) layer processes the output of the convolutional layer after its dimension is reduced by the max-pooling layer and extracts important temporal information for the detection of vocal pathologies.
  - **Fully connected layer:** The fully connected layer receives the output from the BiLSTM layer and performs classification based on the learned features. This layer is typically followed by a Softmax layer.
  - **Output layer:** This layer provides the final prediction and may use a Softmax activation function to produce a probability distribution over the possible

classes

The combination of these layers in a hybrid CNN-BiLSTM network enables the model to effectively capture the spatial and temporal features of the input data for accurate identification of vocal pathologies. The network is trained using a labeled dataset of pathological voice audio recordings from the MEEI database, allowing it to learn to recognize patterns in the data that indicate different types of voice disorders. Fig. 2.9 illustrates the architecture of the proposed system, based on CNN-BiLSTM.

- **Hyperparameter optimization:** The optimization of hyperparameters in a CNN-BiLSTM network is a crucial step for improving the model's performance, especially for complex tasks like pathological voice analysis. This optimization aims to adjust the network parameters to maximize accuracy, minimize loss, and potentially reduce computation time, while avoiding overfitting. In our application, we optimized the hyperparameters of the CNN by integrating six convolutional layers, with the number of filters in the first convolutional layer set to 12, of size 3x3, and this number increases by a factor of 2. The max-pooling filter is of size 2x2. A single BiLSTM layer is used with a variable number of neurons. The activation function is one of the optimization parameters adjusted to study its impact on the performance of the learning procedure.

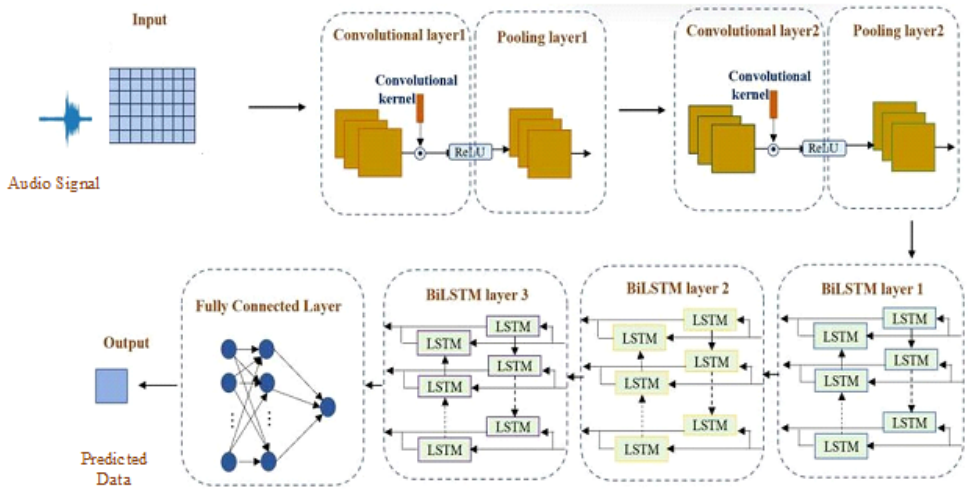

Fig 2.9: Architecture of the proposed CNN-BiLSTM system.

2.6 Pathological speech classification pipeline

2.6.1 System design

The creation of a pathological speech classification system involves two main steps: training and testing. During the training phase, each pathological class is represented by a model that defines the boundaries for each class. In the testing phase, the system uses these models to make decisions for new input data. Fig. 2.10 illustrates the different steps of the process.

a) Data preprocessing

Data preprocessing is a crucial step for ensuring accurate classification of pathological speech. The main steps include:

1. **Data loading:** Input data is loaded as two-dimensional images, representing the number of frames and the dimensions of the acoustic vector, to allow efficient manipulation in machine learning-compatible formats.
2. **Data labeling:** Each acoustic vector is assigned a class label that corresponds to the type of pathology it represents.
3. **Data normalization:** The values in the data matrices are normalized between 0 and 1, ensuring standardized input for more stable and faster model convergence.

Data splitting: Data is split into training and testing sets with an 80:20 ratio. The training set is further divided into a training subset and a validation subset to evaluate the model's performance during training.

b) Model architecture

Our system combines Convolutional Neural Networks (CNN) with Bidirectional Long Short-Term Memory (BiLSTM) networks. The architecture includes layers for convolution, pooling, fully connected operations, and dropout. A dropout rate of 0.2 is applied, which randomly deactivates 20% of the neurons during training. This technique prevents the model from over-relying on specific neurons, improving its generalization capability when handling new data.

c) Model training

During training, the CNN-BiLSTM model uses backpropagation in conjunction with the Adam optimizer to minimize the loss function and adjust network weights. Data passes through successive layers applying nonlinear transformations, such as ReLU, and this process is repeated over multiple epochs until the model converges. Convergence is achieved when the training loss sufficiently decreases, allowing the model to make accurate predictions on unseen data.

2.6.2 Training and optimization

Enhancing the performance of the CNN-BiLSTM model for pathological speech classification is further achieved through speech enhancement techniques. These methods help clean and strengthen speech signals, improving feature extraction and ultimately model performance. Techniques such as spectral subtraction, Wiener filtering, and Minimum Mean Square Error (MMSE) are applied to raw speech signals.

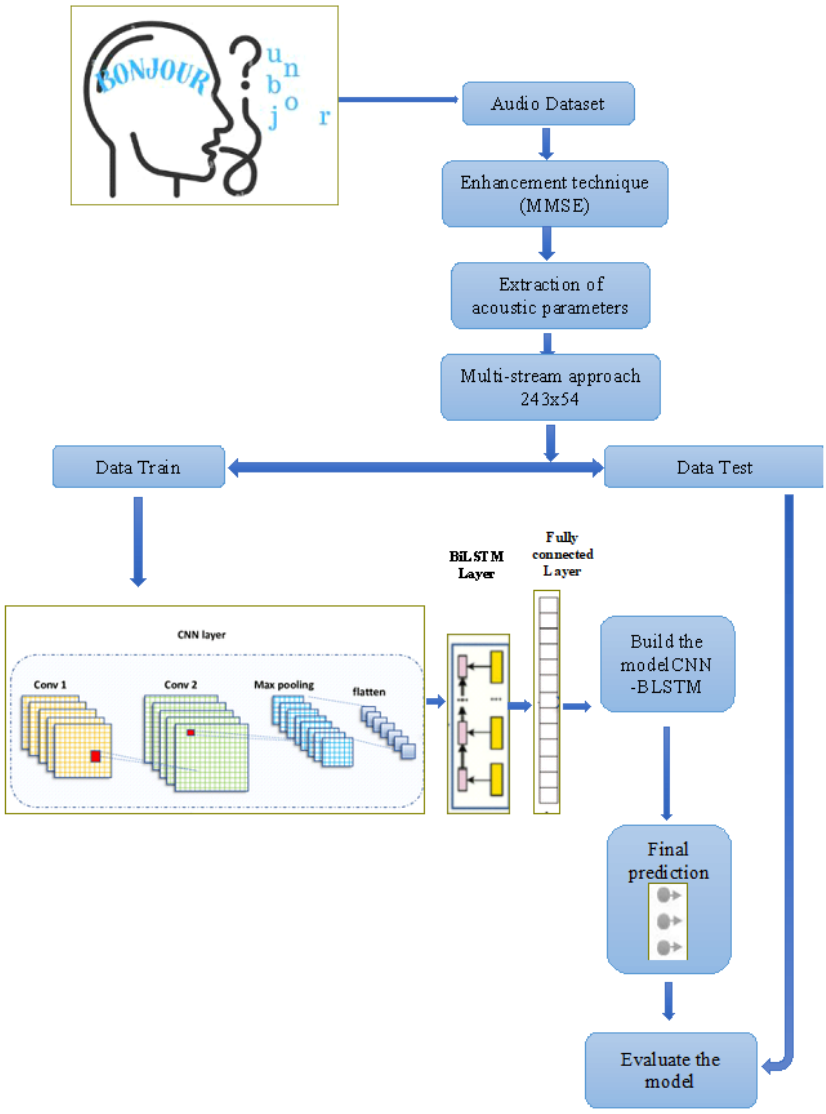

Fig 2.10: Flowchart representing the proposed diagnosis system architecture.

Key factors that influence the effectiveness of these techniques include:

- **Spectral subtraction:** The over-subtraction factor controls how much of the noise

spectrum is removed. An excessive value can distort speech, while a low value may leave residual noise, making it crucial to find an optimal balance.

- **Wiener filtering:** The smoothing factor affects temporal smoothing in noise spectrum estimation. Higher smoothing leads to a smoother signal but slower noise adaptation, whereas lower smoothing allows faster noise adjustment but may introduce distortions.
- **MMSE algorithm:** The smoothing parameter determines the estimation of the a priori signal-to-noise ratio (SNR). Higher values create more stable estimations but respond slower to rapid noise changes, while lower values react faster but risk instability.

Optimizing these parameters for each technique enhances the signal quality, resulting in clearer feature extraction and improved classification accuracy for different types of vocal pathologies, thus providing more reliable diagnostic support.

2.6.3 Integration of multi-stream features

Integrating multi-stream features into the voice classification system, particularly for pathological speech analysis, boosts the model's accuracy and robustness by utilizing various representations of the vocal signal. Each feature set captures unique characteristics of the signal, such as pitch, frequency, and energy, allowing for a more comprehensive model. In our system, the combined feature vector includes Mel-Frequency Cepstral Coefficients (MFCC), Power-Normalized Cepstral Coefficients (PNCC), and prosodic features like Jitter and Shimmer. The matrices are concatenated by calculating the maximum number of frames among the feature sets. Zeros are padded to the matrices with fewer frames, ensuring uniform dimensionality for all features.

2.7 Conclusion

Throughout this chapter, we have explored the methodological framework and implementation details of our pathological speech classification system, focusing on several key components. First, we addressed data collection and preprocessing, utilizing the MEEI database and applying speech enhancement techniques to improve data quality. We then explored acoustic feature extraction, including multi-stream features such as MFCCs, PNCCs, and prosodic parameters. The development and training of a hybrid CNN-BiLSTM model for classification were also detailed, followed by the integration of acoustic features into a unified representation for input to the model. By combining these

elements, we established a robust and effective system for classifying pathological speech. The next chapter will focus on evaluating the system and analyzing its performance.

Chapter 3

Experiments and evaluations

3.1 Introduction

This chapter will delve into the evaluation of our proposed pathological speech classification system. We will discuss the datasets used and examine how the proposed technique enhances the robustness of the system through various experiments and a detailed analysis of the results. The evaluation begins with the deep learning system combining CNN and BiLSTM, followed by an exploration of how the fusion of acoustic parameters using a multi-stream approach improves system performance. Additionally, we will assess the impact of speech enhancement techniques and various activation functions in optimizing accuracy. Finally, the performance of our proposed system will be compared with a previously developed DNN-based system, and we will evaluate the system's ability to classify pathological versus normal speech.

3.2 Dataset

The empirical experiments were conducted using the MEEI database, as described in the previous chapter. Two datasets were selected from MEEI database:

1. **Dataset 1**, outlined in Table. 3.1, encompasses various vocal pathologies that adversely affect the vocal cords, leading to voice difficulties. This dataset includes

    a. **Vocal cord nodules:** Non-cancerous growths caused by vocal strain or misuse [83].

    b. **Vocal cord paralysis:** Impairment of vocal cord movement due to nerve damage [84].

    c. **Polypoid lesions:** Benign growths on the vocal cords caused by vocal abuse or trauma [85].

Table 3.1: Distribution of female and male records with average age for the dataset1.

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

2. **Dataset 2** includes anteroposterior compressions, gastric reflux, and mild ventricular compression. This dataset has an equal distribution of male and female speaker

files, as detailed in Table. 3.2. It is important to note that the participant pool is evenly distributed by gender, and age effects are neutralized, ensuring that all groups share the same average age.

a. **Gastric reflux** can lead to voice disorders when stomach acid enters the esophagus and larynx, causing irritation and inflammation [86].
b. **Anteroposterior compression** typically occurs due to poor closure of the vocal cords, resulting in a sensation of compression during phonation [87].
c. **Ventricular compression disorder** involves the compression of the ventricular folds during phonation, often resulting in a strained or breathy voice quality [88].

Table 3.2: Distribution of female and male records with average age for the dataset2.

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	12	41.25 ± 12.83	13	46.46 ± 13.40	25

These disorders affect the vocal mechanisms differently, leading to various changes in voice quality and potentially resulting in dysarthria. The severity of dysarthria depends on the underlying disorder and the individual's overall health. Severe dysarthria, such as that caused by anteroposterior or ventricular compression, can significantly impact speech intelligibility due to pronounced changes in voice quality [89].

3.3 Experimental setup

3.3.1 Experimental protocol

The experimental protocol for evaluating the pathological speech classification system involved three phases:

1. **Baseline system development:** A pathological speech classification system was established using a deep learning model (CNN-BiLSTM). The system incorporated acoustic parameters such as MFCC, PNCC, and Mel-spectrogram coefficients.

2. **Multi-variable acoustic analysis:** To enhance robustness, we explored a multi-variable acoustic analysis approach that combined various data streams, including MFCC, PNCC, Mel-spectrograms, fundamental frequency (Fo), Jitter, and Shimmer.
3. **Speech enhancement:** To improve speech quality and intelligibility, we applied speech enhancement techniques such as Wiener filtering, spectral subtraction, and the MMSE algorithm. The MMSE algorithm was implemented considering different frequency ranges (0–1.5 kHz, 0–2.5 kHz, 0–3.5 kHz).

To validate and evaluate the three-phase system architecture, we conducted experiments by comparing it to a previously developed DNN-based pathological speech recognition system. The comparison results were analyzed to assess the effectiveness of the proposed system.

Additionally, we evaluated the proposed system's ability to differentiate between pathological and normal speech by examining a set of three pathological classes and one normal class.

3.3.2 Evaluation metrics

To evaluate pathological speech understanding, researchers often focus on the following areas:

- **Objective measures:**
  - **Speech analysis tools:** Use software to analyze acoustic features (e.g., pitch, loudness, speech rate).
  - **Automated speech recognition (ASR):** Evaluate how well ASR systems transcribe and understand pathological speech.
- **Subjective measures:**
  - **Listener perception studies:** Gather data from speech-language pathologists and naive listeners on their understanding of pathological speech.
  - **Comprehension tests:** Administer standardized tests to assess listener understanding of spoken phrases or sentences.
- **Clinical assessments:** Use tools like the Western Aphasia Battery (WAB) or Boston Diagnostic Aphasia Examination (BDAE) to assess understanding in clinical settings.

- **Machine learning approaches:** Employ machine learning algorithms to classify and predict understanding levels based on speech features. Explore neural network models for improved speech recognition in varied conditions.
- **Impact of interventions:** Assess how different speech therapy techniques affect the understanding of pathological speech.
- **Cultural and linguistic considerations:** Examine how dialects and languages influence the comprehension of pathological speech.

To evaluate our proposed model, we used objective measures and machine learning approaches. We employed speech analysis tools, ASR systems, and accuracy metrics to assess the model's performance. Additionally, we conducted subjective evaluations with individuals suffering from dysarthria to gather concrete feedback.

The evaluation experiments were conducted using the MEEI pathological voice database. Subsets were created to assess the effectiveness of the deep learning approach, multivariable acoustic analysis, and speech enhancement techniques for classifying pathological voices. A DNN-based system was developed for comparison.

The classification model was evaluated using the "accuracy" metric. Accuracy represents the percentage of correct predictions. High accuracy indicates the model's ability to correctly predict test sample classes.

$$accuracy = \frac{\text{number of correct predictions}}{\text{total number of predictions}} \tag{3.1}$$

Confusion matrices were also used as an evaluation metric. These matrices were generated for systems using the original input signals and signals enhanced with the MMSE algorithm.

3.3.3 Hyperparameter tuning

Hyperparameter tuning is a critical step in optimizing the performance of the CNN-BiLSTM model. By carefully selecting the appropriate hyperparameters, we can improve the model's ability to generalize to new data and achieve better classification results.

The combined CNN-BiLSTM model and the DNN system [90] were implemented using MATLAB software. Different architectures and features were explored to find optimal configurations. Table. 3.3 describes the hyperparameters used for both systems.

These hyperparameters were chosen based on empirical experimentation and common practices in deep learning. The specific values may vary depending on the dataset and the complexity of the task. By carefully tuning these hyperparameters, we were able to opti-

mize the performance of our CNN-BiLSTM model for pathological speech classification. The experiments were carried out using the MATLAB R2021a software.

Table 3.3: Hyperparameter configuration

	CNN-BiLSTM	DNN
Input dimensions	243 frames, 54 coefficients	243 frames, 54 coefficients
Batch size	10	10
Number of epochs	320	320
Dropout rate	0.2	0.2
Number of filters of each convolutional hidden layer	[12, 24, 48, 96, 192,384]	-
Convolution filter size	3x3	-
Learning rate	0.0003	0.0003
LSTM units	Variable	-
Hamming window size	25 ms	25 ms
Optimization algorithm	Adam	Adam
Number of neurons in the hidden layers of the DNN	-	[100, 100]

3.3.4 Cross-validation

Cross-validation is a crucial technique in machine learning used to assess model's performance and ensure its ability to generalize to unseen data. In k-fold cross-validation, the dataset is divided into k equal-sized subsets or "folds.". The model is then trained and evaluated k times.

Here is a description of the K-fold cross-validation strategy used in our application:

1. **Data splitting:** The dataset is randomly divided into k equally-sized folds (subsets). Common choices for k are 5 or 10, but this can vary based on the size of the dataset.

2. **Model training and validation:**
    - In each iteration:
        - **Training set:** The model is trained on k-1 folds.
        - **Validation set:** The model is validated on the remaining fold.
    - This process is repeated k times, with each fold serving as the validation set exactly once.
3. **Performance averaging:** After k iterations, the performance metrics (e.g., accuracy, F1 score) from each fold are averaged to obtain a robust estimate of the model's generalization performance

Advantages of k-fold cross-validation

- **Reduced bias:** By using multiple training-validation splits, k-fold cross-validation provides a more unbiased estimate of model performance compared to a single train-test split.
- **Improved generalization:** Training the model on different subsets of the data helps prevent overfitting to a specific training set, leading to better generalization performance on unseen data.

This approach ensures a more reliable evaluation of the model's performance and helps to identify potential overfitting issues.

3.4 Results analysis

3.4.1 Comparison of acoustic feature vectors

We developed a pathological speech classification system based on a deep learning model that combined a convolutional neural network (CNN) and a bidirectional long short-term memory network (BiLSTM).
During the acoustic analysis phase, we tested the system's performance using various acoustic features:

- **Mel-frequency cepstral coefficients (MFCC):** Calculated every 10 ms over a 25 ms Hamming window.
- **Power-normalized cepstral coefficients (PNCC):** Calculated using the same parameters as MFCC.

- **Mel-spectrogram coefficients:** Computed over a 25 ms Hamming window with a 10 ms shift, resulting in a spectrogram with dimensions of 243x50.

For each feature vector, the energy of the frame was added to the 12-dimensional static coefficients to provide information on signal intensity. Additionally, the first and second derivatives (deltas and delta-deltas) were calculated and added to the static vectors, resulting in 39-dimensional feature vectors (MFCC_E_D_A and PNCC_E_D_A). The training and testing phases of our system used a simple split technique (80% training, 20% testing) due to its simplicity and efficiency.

The results in Table. 3.4 demonstrate the effectiveness of Mel-spectrogram coefficients in classifying nodules, paralysis, and polypoid pathologies. Compared to MFCC and PNCC vectors, Mel-spectrogram coefficients achieved a 5-10% improvement in accuracy.

Table 3.4: Accuracy (%) of the CNN-BiLSTM network with various acoustic features

Acoustic Vector	Accuracy (%)
PNCC_E_D_A (39)	61.90
MFCC_E_D_A (39)	66.67
Mel-Spectrogram (50)	**71.43**

3.4.2 Leveraging multi-variable feature fusion

To improve the performance of our pathological speech classification system, we employed a multi-stream approach. This approach involves fusing various acoustic parameters representative of the speech signal, inspired by human hearing. We specifically analyzed the vocal spectrum image (Mel-spectrogram), cepstral coefficients (MFCC and PNCC), and prosodic aspects such as the fundamental frequency (Fo), jitter, and shimmer.

The selection of acoustic parameters significantly influences the performance of pathological classification systems. Each range of acoustic measurements was chosen for its specific contributions:

- **MFCC and Mel-spectrogram images:** Capture the spectral envelope of speech signals, effectively representing spectral characteristics.
- **PNCC:** Capture perceptually relevant aspects of speech signals and demonstrate robustness in noisy environments and voice disorders.

- **Jitter and Shimmer:** Provide information on the stability, regularity, and vibratory characteristics of the vocal cords, aiding in detecting pathological voice disorders.

We evaluated systems based on the multi-stream approach and observed that they achieved impressive classification accuracy values compared to single-stream data systems. The results for the dataset1 are presented in Table. 3.5 below.

Table 3.5: Accuracy (%) of the CNN-BiLSTM network using the multi-variable acoustic analysis approach

Acoustic parameter	**Accuracy (%)**
Mel-Spectrogram-Jitter-Shimmer	76.19
Mel-Spectrogram-Jitter-Shimmer-PNCC	71.43
MFCC -F0-Jitter-Shimmer-PNCC	**80.95**
MFCC-Jitter-Shimmer-PNCC	85.71

Building on the results from Table. 3.4, we concluded that the highest performance for pathological speech classification was achieved using Mel-spectrogram images. The first system developed under the multi-variable acoustic analysis framework integrates 50-dimensional Mel-spectrogram parameters with acoustic measures Jitter and Shimmer, which quantify variability in the fundamental frequency and amplitude of the voice, respectively. This system is denoted as Mel-spectrogram-Jitter-Shimmer (52), where 52 represents the dimensionality of the acoustic vector.

Next, we introduced a second data stream consisting of 13 PNCC coefficients, integrating it with the first system. The resulting multi-stream acoustic vector is denoted as Mel-spectrogram-Jitter-Shimmer-PNCC (65).

A third system for pathological speech classification was developed, creating a more comprehensive set of parameters. The first stream consists of 39 MFCC coefficients, including the energy component and their first and second derivatives. The second stream incorporates Fo, Jitter, and Shimmer coefficients, while the third stream adds 13 PNCC coefficients. This new acoustic vector is noted as MFCC_F0_Jitter_Shimmer_PNCC (55). The choice to concatenate the MFCC and PNCC coefficients is inspired by experiments conducted in [91], which demonstrated that integrating PNCC into the MFCC-based acoustic vector improves the system's accuracy, particularly for low SNR values.

To compare performance, we conducted experiments with two variants of the acoustic frame (with and without the fundamental frequency Fo) to assess its impact. The

best classification accuracy scores served as the baseline for evaluating other acoustic vectors. The acoustic frame excluding the fundamental frequency is denoted as MFCC_Jitter_Shimmer_PNCC (54).

The results presented in Table. 3.5 demonstrate that Jitter and Shimmer coefficients yield better classification accuracy when combined with PNCC and MFCC coefficients rather than with Mel-spectrogram images. For example, replacing the Mel-spectrogram components in the Mel-spectrogram-Jitter-Shimmer-PNCC acoustic vector with MFCC improved classification accuracy by 14%, while also reducing the dimensionality from 65 to 54 coefficients.

Moreover, comparing results with and without the fundamental frequency Fo revealed that Fo introduces redundant information, and reducing the dimensionality to 54 helped preserve important information while enhancing accuracy. Specifically, the MFCC-Jitter-Shimmer-PNCC acoustic vector improved accuracy by 5% compared to the system that included the fundamental frequency. Removing Fo also reduced computational costs and storage requirements while maintaining high classification accuracy.

In summary, the multi-variable approach, particularly the MFCC-Jitter-Shimmer-PNCC combination, provides a significant improvement in classification accuracy compared to single-stream features. This demonstrates the value of integrating diverse acoustic information for robust pathological speech classification.

3.4.3 Impact of segmentation techniques

To evaluate the impact of data segmentation techniques, we compared a simple segmentation approach with a threefold cross-validation method.

- **Simple segmentation:** The dataset was divided into training (70%) and testing (30%) sets.
- **Threefold cross-validation:** The dataset was divided into three equal parts. One part was used for testing, while the remaining two parts were combined for training. This process was repeated three times, with each part being used for testing once.

In threefold top segmentation, the upper part is reserved for testing, while in threefold middle segmentation, the middle part is reserved for testing. Lastly, in threefold bottom segmentation, the lower part is reserved for testing, with the remaining two parts used for training. In our application, we calculated the average accuracy for the three methods: threefold top, middle, and bottom.

The results in Table. 3.6 demonstrate that the simple segmentation technique achieved the best performance for the dataset1. While threefold cross-validation provides a more

robust evaluation, the difference in accuracy compared to the simple technique was not significant in this case.

It's important to note that the choice of segmentation technique may vary depending on the dataset size and the specific requirements of the classification task. In some cases, more complex cross-validation strategies may be necessary to ensure unbiased evaluation.

Table 3.6: Accuracy (%) of the CNN-BiLSTM system with MFCC-Jitter-Shimmer-PNCC for the dataset1

Segmentation Techniques	Training + validation	90	80	70	60	50	67
	Test	10	20	30	40	50	33
Accuracy (%)		80.00	85.71	**86.67**	75.61	76.47	68.91

3.4.4 Impact of speech enhancement techniques

To further enhance the performance of the pathological speech classification system, we integrated speech enhancement techniques into the multi-stream acoustic vector. These techniques included the MMSE algorithm, Wiener filtering, and spectral subtraction.

The MMSE algorithm is particularly effective in estimating and reducing background noise while preserving important speech characteristics. To explore the impact of frequency range on noise reduction, we implemented MMSE for different frequency bands: 0–1.5 kHz (MMSE 15), 0–2.5 kHz (MMSE 25), and 0–3.5 kHz (MMSE 35).

The rationale for restricting these frequency bands is grounded in a study [92], which determined that segmenting the frequency bands particularly the 0–3.5 kHz range, encompassing critical vocal information such as formants better facilitates the classification of voice disorders compared to using the entire frequency spectrum.

Table. 3.7 below presents a comparison of various enhancement techniques, including the MMSE approach, Berouti's algorithm based on spectral subtraction, and Wiener filtering, focusing on noise reduction, speech signal distortion, presence of artifacts, and algorithm complexity.

The results in Table. 3.8 demonstrate that the MMSE algorithm, particularly MMSE 35, significantly improves classification accuracy. Berouti's algorithm also shows moderate improvement. Wiener filtering maintains a similar accuracy level as the baseline system without enhancement.

These findings highlight the importance of speech enhancement techniques in improving the robustness and accuracy of pathological speech classification systems. The choice

of enhancement technique depends on the specific characteristics of the dataset and the desired trade-off between noise reduction and signal distortion.

Table 3.7: Comparison of speech enhancement techniques

Method	Noise reduction	Signal distortion	Improved SNR	Artifacts	Complexity
MMSE	High	Low	High	Low	Medium
Berouti algorithm	Medium	Medium	Medium	Moderate to High	Low
Wiener filtering	High	Low to Medium	High	Low	High

Table 3.8: Accuracy (%) of the CNN-BiLSTM network with speech enhancement

Enhancement Technique	Accuracy (%)
None	85.71
Wiener filter	85.71
Berouti Algorithm	90.48
MMSE 15	90.48
MMSE 25	90.48
MMSE 35	**95.24**

3.4.5 Impact of activation functions

Activation functions play a crucial role in determining the performance of neural networks, allowing them to learn complex data patterns. After receiving input, the activation function decides whether the neuron should be activated and pass the signal to the next layer. To explore the impact of different activation functions, listed in Table. 3.9, on pathological speech classification, we experimented in addition to ReLU: Gaussian Error Linear Unit (GELU) [93], a smooth version of ReLU with both positive and negative values; Scaled Exponential Linear Unit (SELU) [94], which promotes network self-normalization;

Sinus Rectified Unit (SinRU), a blend of ReLU and sinus functions; and Adaptive Rectified Linear Unit (AReLU), an adaptive variant of ReLU that enhances performance and flexibility.

Table 3.9: Activation functions

Activation function	**Expression**
Asymmetric Rectified Linear Unit (AReLU)	$AReLU(x, \alpha, \beta) = \begin{cases} C(\alpha)x & \text{if } x < 0 \\ (1 + \sigma(\beta))x & \text{if } x \geq 0 \end{cases}$
Scaled Exponential Linear Unit (SELU)	$SELU(x) = \begin{cases} \lambda x & \text{if } x < 0 \\ \lambda\alpha(e^x - 1) & \text{if } x \geq 0 \end{cases}$ $\alpha = 1.67$ 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)	$SinRU = \begin{cases} 0 & \text{if } x \leq 0 \\ x + \sin(x) & \text{if } x > 0 \end{cases}$

Table. 3.10 summarizes the system performances and highlights the best results.

Comparing the accuracy rates of the MMSE 15, MMSE 25, and MMSE 35 methods integrated with the original multi-stream system, we observed significant improvements. The results in Table. 3.10 demonstrate that the choice of activation function can significantly impact classification accuracy. The ReLU and AReLU activation functions consistently outperformed the others, especially when combined with the MMSE 35 enhancement technique.

Table 3.10: Classification accuracy (%) with different activation functions

Activation Function	**Multi-Stream**	**MMSE 15**	**MMSE 25**	**MMSE 35**
ReLU	85.71	**90.48**	90.48	**95.24**
SeLU	76.19	80.95	80.95	80.95
GeLU	**90.48**	**90.48**	80.95	90.48
SinRU	80.95	85.71	80.95	76.19
AReLU	85.71	80.95	80.95	**95.24**

Fig. 3.1 and Fig. 3.2 show confusion matrices for the original system and the system enhanced with MMSE 35. The differences in the confusion matrices highlight the impact of the enhancement on the accuracy and specificity of the classification. The combination of multi-stream features and MMSE 35 significantly improved classification accuracy, particularly for nodules and polypoid lesions. This improvement is 33% over the original system based solely on PNCC coefficients, as illustrated in Fig. 3.3.

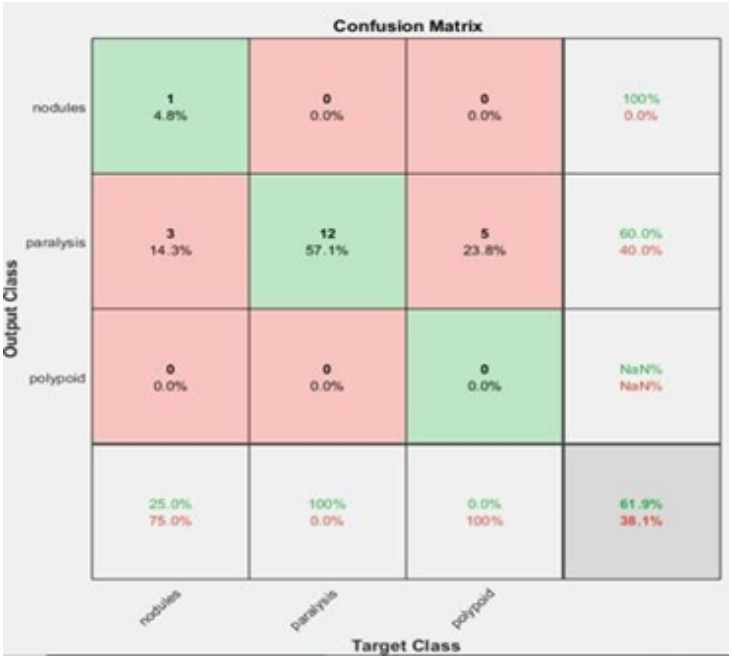

Fig 3.1: Original system confusion matrix.

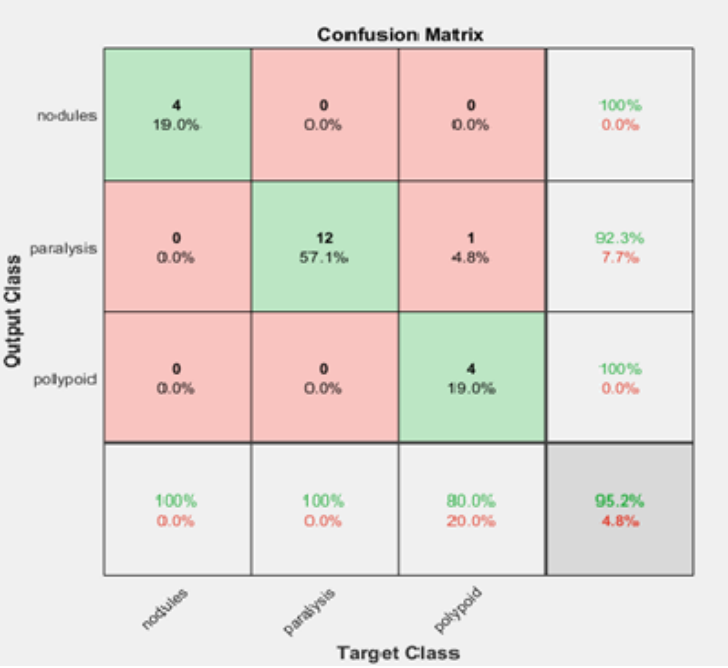

Fig 3.2: Enhanced system confusion matrix

In summury, the choice of activation function is a critical factor in optimizing the performance of the pathological speech classification system. The ReLU and AReLU functions, in combination with the MMSE 35 enhancement technique, provide the best results in terms of accuracy and overall performance.

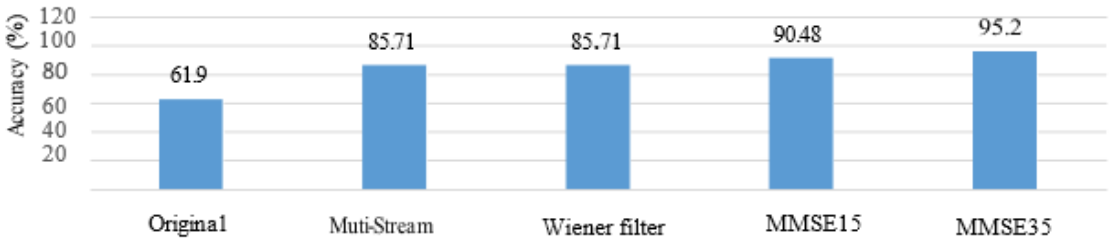

Fig 3.3: Classification accuracy for dataset1 with different enhancement techniques.

3.4.6 Comparison of CNN-BiLSTM and DNN-based systems

To demonstrate the effectiveness of our proposed CNN-BiLSTM system [95], we conducted a comparative study with a DNN-based system. Both systems were evaluated on the same datasets, using various acoustic features and speech enhancement techniques.

Table 3.11: Comparison of CNN-BiLSTM and DNN systems

Technique	**CNN-BiLSTM (%)**	**DNN (%)**
MFCC	**66.67**	61.90
Multi-Stream (MFCC-Jitter-Shimmer-PNCC)	**85.71**	76.19
MMSE 35 + Multi-Stream (MFCC-Jitter- Shimmer-PNCC)	**95.24**	85.71

The comparative accuracy rates for each CNN-BiLSTM and DNN system are presented in Table. 3.11. The learning and testing phases of the DNN system were carried out using a simple split technique (80% for training and 20% for testing) on the Dataset1, which includes nodules, paralysis, and polypoid lesions.

The results in Table. 3.11 clearly demonstrate the superior performance of the CNN-BiLSTM system. This is attributed to the combination of convolutional layers, which are effective for capturing spatial patterns in the speech signal, and bidirectional LSTM layers, which can model long-term temporal dependencies.

Fig. 3.4 shows the classification accuracies for the CNN-BiLSTM and DNN systems using different corpus splitting techniques. The CNN-BiLSTM system consistently outperforms the DNN system across all configurations, further highlighting its effectiveness for pathological speech classification.

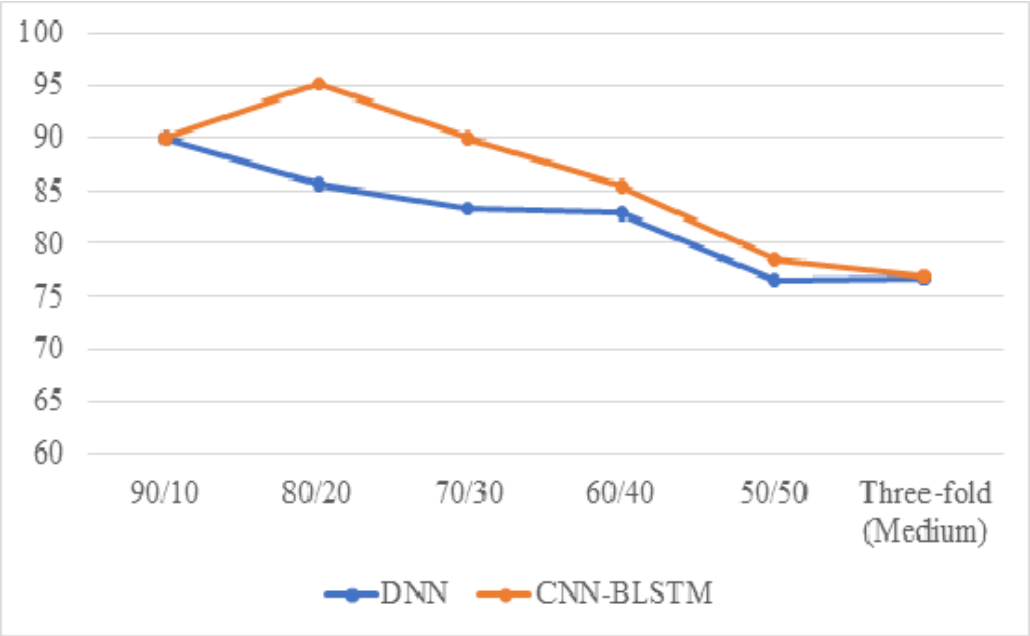

Fig 3.4: Comparison of accuracies (%) for different corpus splitting techniques.

In conclusion, the CNN-BiLSTM system, combined with multi-stream acoustic features and speech enhancement techniques, significantly outperforms the DNN-based system. This demonstrates the benefits of using a more sophisticated deep learning architecture tailored to the specific requirements of pathological speech classification. The CNN serves as the feature extractor, learning local patterns in the acoustic features from the speech signal, it captures important information like phonetic and prosodic features that help differentiate between normal and pathological speech, and the BiLSTM component adds temporal modeling capability by capturing long-range dependencies in the speech sequence, which is particularly important for speech signals with temporal variation. The BiLSTM layer helps the system understand how these features (such as prosody, intonation, or speech rate) evolve over time, improving the recognition of speech patterns that are not purely local but require context over time to be correctly classified.

3.4.7 Performance on different pathologies

To evaluate the generalizability of our system, we conducted additional experiments using a second dataset that included anteroposterior compression, gastric reflux, and mild ventricular compression. This dataset was balanced in terms of gender and age to ensure a more representative sample.

We compared the performance of the CNN-BiLSTM system with different configurations, including the original multi-stream approach, MMSE 35 enhancement, and various activation functions. The Table. 3.12 indicate that the best performance of the original system including the multi-stream approach was achieved with the simple configuration (80/20).

The results shown in Table. 3.13 demonstrate that the combination of MMSE 35 enhancement and the AReLU activation function significantly improves classification accuracy for the second dataset. This improvement is particularly noticeable for the anteroposterior compression and mild ventricular compression classes.

The use of a balanced dataset with equal representation of male and female speakers is crucial for ensuring the generalizability of the classification model. This helps to mitigate biases and improve performance for a wider range of individuals. Finally, the proposed system, incorporating multi-stream features, speech enhancement, and the AReLU activation function, demonstrates robust performance across different pathological speech datasets. This highlights the system's potential for real-world applications in clinical settings.

Table 3.12: Impact of corpus splitting techniques for dataset2

Technique	**Training+Validation**	**Test**	**Accuracy (%)**
90/10	90%	10%	37.50
80/20	80%	20%	**52.94**
70/30	70%	30%	48.00
60/40	60%	40%	41.18
50/50	50%	50%	40.48
Threefold	67%	33%	29.47

Table 3.13: Classification performance for the dataset2

Method	**Accuracy (%)**
Multi-Stream (MFCC-Jitter-Shimmer-PNCC)	52.94
MMSE 35 & ReLU	58.82
MSE 35 & AReLU	**64.71**

3.4.8 Advancing speech recognition for dysarthric individuals

This research investigates the potential of advanced machine learning techniques to improve speech comprehension for individuals with dysarthria. Dysarthria, a motor speech disorder, significantly impacts communication and quality of life. While existing speech recognition systems often struggle to accurately transcribe dysarthric speech, this study aims to contribute to the development of personalized speech assistance systems specifically designed for individuals with this condition.

To this end, the performance of a previously developed model was assessed. This model integrates a CNN-BiLSTM architecture with a multi-stream analysis and incorporates speech enhancement using the Minimum Mean Square Error (MMSE) technique. This combination of features aims to improve the accuracy and robustness of speech recognition in the presence of dysarthric speech characteristics.

3.4.8.1 Customized methodology

a) Dataset

This study utilizes the UASpeech (Universal Access Speech) database, which specializes in dysarthric speech [96]. This database comprises recordings from speakers with varying levels of dysarthria, including control speakers without dysarthria.

- **Data collection:**
  - Recordings were made using seven microphones arranged in a line.
  - Speakers pronounced isolated words displayed on a computer screen.
  - Each speaker read three blocks of words, including digits, letters, commands, and common words.
- **Data characteristics:**
  - Speakers were categorized as "C" (control), "M" (male with dysarthria), or "F" (female with dysarthria).
  - Intelligibility levels for each speaker were assessed.
  - Audio files are named according to a specific format (e.g., F02_B3_C17_M2.wav).
  - Master Label Files (MLF) provide word-level annotations for each audio file.

b) Speaker selection

- For this study, two speakers were selected from each of three intelligibility levels (low, medium, and high) (Table. 3.14) to ensure a balanced representation of dysarthria severity.

c) Developed system

This section details the evaluated speech recognition system. Building upon the previously described CNN-BiLSTM architecture with multi-stream analysis (MFCC-PNCC) and MMSE-based speech enhancement, this system incorporates the CTC loss function for improved performance.

Table 3.14: Characteristics of the speakers used in UASpeech

Speaker	Age	Speech intelligibility
M04	>18	2%
F02	30	29%
F04	18	62%
M09	18	86%

CTC allows for training sequence-to-sequence models without requiring precise frame-level alignment between the input audio and the target transcription. This is crucial for handling the variability and distortions inherent in dysarthric speech, as it enables the model to learn robustly from unaligned sequences. CTC replaces the traditional cross-entropy loss function, allowing the model to directly learn the mapping between the input audio and the corresponding sequence of characters or words [97].

This integrated approach leverages the strengths of deep learning architectures, advanced feature extraction, and robust speech enhancement techniques to enhance the accuracy and reliability of speech recognition for individuals with dysarthria.

d) Hyperparameter tuning

Table. 3.15 outlines the hyperparameters selected to optimize the performance of the CNN-BiLSTM system.

Our experiments are built using Matlab R2024b software. Learning toolbox in Matlab offers a seamless way to define, train, and evaluate models with CTC, which is particularly well-suited for sequential data like speech.

e) Training and test data

To rigorously evaluate the model's performance, three-fold cross-validation was employed. This technique divides the dataset into three subsets:

1. **Training set:** The model is trained on two-thirds of the data.
2. **Validation set:** The model is evaluated on one-third of the data not used for training.

This process is repeated three times, with each subset serving as the validation set once. The performance metrics from each fold are then averaged to obtain a robust estimate of the model's generalization performance.

f) Evaluation metrics

The performance of the speech recognition system was evaluated using the following metrics:

- **Character error rate (CER):**
  CER measures the number of character-level errors (substitutions, insertions, and deletions) in the transcribed text compared to the ground truth [98].

$$CER = \frac{\text{Substitutions } + \text{Insertions } + \text{Suppressions}}{\text{Nombre total de caractère dans la transcription correcte}} \tag{3.2}$$

- **Word error rate (WER):**
  WER measures the number of word-level errors (substitutions, insertions, and deletions) in the transcribed text compared to the ground truth [99].

$$WER = \frac{\text{Substitutions } + \text{Insertions } + \text{Suppressions}}{\text{Nombre total de mots dans la transcription correcte}} \tag{3.3}$$

These metrics provide a comprehensive assessment of the system's accuracy in transcribing dysarthric speech.

Table 3.15: Hyperparameter configuration

Model parameters	**Value**
Number of epochs	280
Batch size	10
Input dimensions	147 frames, 52 coefficients (MFCC_PNCC)
Learning rate	0.0003
Dropout rate	0.2
Number of filters of each convolutional layer	[16, 20, 32]
Number of LSTM units	198
CPU training time	2H 45min

3.4.8.2 Analysis of results

The performance of the dysarthric ASR system was evaluated across four speakers with varying levels of dysarthria, as shown in Table. 3.16.

The results demonstrate a range of accuracies across speakers, with F02 achieving the highest accuracy (87.47%) and M04 achieving the lowest (80.39%). Similarly, Character Error Rate (CER) and Word Error Rate (WER) varied across speakers, with F02 exhibiting the lowest error rates and M04 exhibiting the highest.

Table 3.16: ASR system performance metrics (in %)

Speakers	Accuracy	CER	WER
F02	**87.47**	**10.51**	**12.53**
F04	80.91	15.47	19.09
M04	80.39	15.90	19.61
M09	85.49	12.04	14.51

Factors influencing performance:
Several factors can influence the system's performance on individual speakers.

- Severity of dysarthria: While speaker F02 had lower intelligibility (29%), the system achieved high accuracy. This suggests that intelligibility level alone may not be the sole determinant of system performance.
- Individual speech characteristics: Factors such as articulation patterns, speech rate, and accent can significantly impact the system's ability to accurately transcribe speech.
- Speech variability: Variability in speech production, such as changes in pitch, intonation, and loudness, can also pose challenges for the system.

Despite the variations in performance across speakers, the system demonstrated promising results in recognizing dysarthric speech, indicating the potential of the proposed approach for improving communication for individuals with this condition (Fig. 3.5 & 3.6).

3.4.8.3 Comparison with a Pix2Pix GAN-based system

To evaluate the performance of the proposed CNN-BiLSTM model, its results were compared against a baseline system that utilized a Pix2Pix Generative Adversarial Network (GAN) for speech enhancement.

Pix2Pix GANs, designed for image-to-image translation tasks [100], were applied to transform pathological speech spectrograms into "healthy" spectrograms, thereby enhancing the speech signal before processing by the speech recognition system (Fig. 3.7).

The ASR model leverages a combination of CNN and RNN (specifically GRU) for effective speech recognition. Data preprocessing is crucial for model accuracy and includes steps such as normalization, transcription encoding, and data splitting. The model's architecture employs the CTC alignment technique to map input features to output sequences [101].

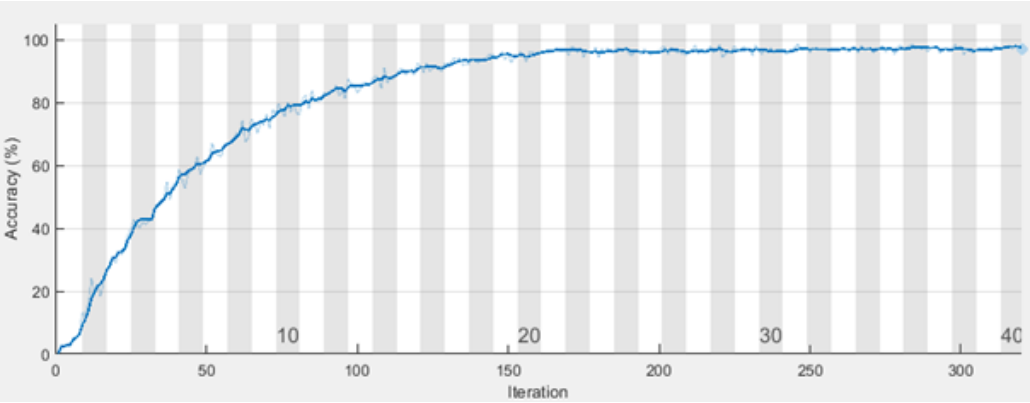

Fig 3.5: The accuracy over epochs for speaker F04.

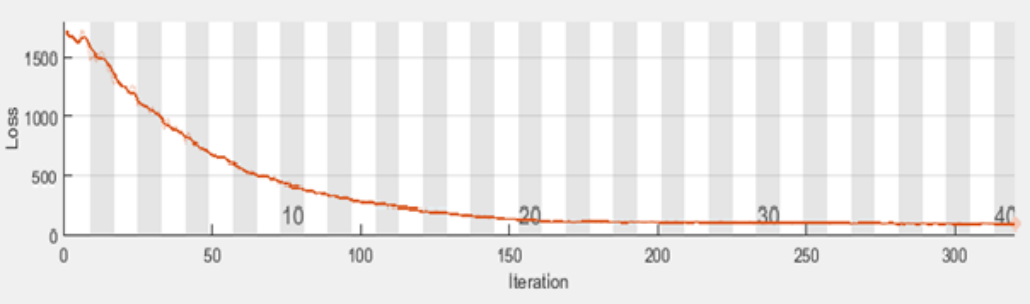

Fig 3.6: The loss over epochs for speaker F04.

Key features and components:

A) Preprocessing Pix2Pix model data (Fig. 3.7)

- **Input:** The Pix2Pix model uses concatenated pairs of images (pathological and

corresponding healthy spectrograms) as input, enabling conditioned transformations [102].

- **Generator:**
  - Employs a U-Net architecture: This architecture excels in image-to-image translation tasks.
  - Consists of an Encoder (extracts features and reduces dimensionality) and a Decoder (reconstructs the image), connected by skip connections, ensuring precise translations while preserving spatial details.
- **Discriminator:**
  - Implements a convolutional PatchGAN architecture (70x70 patch size) to evaluate the authenticity of generated images, focusing on local image details.

This detailed preprocessing ensures the Pix2Pix model effectively transforms pathological spectrograms into enhanced ones, optimizing them for subsequent recognition by the ASR system.

Training of the Pix2Pix model

- **Iterative Process:** The training alternates between updating the generator and the discriminator.
- **Objective:** To train the generator to produce realistic healthy spectrograms from pathological inputs.

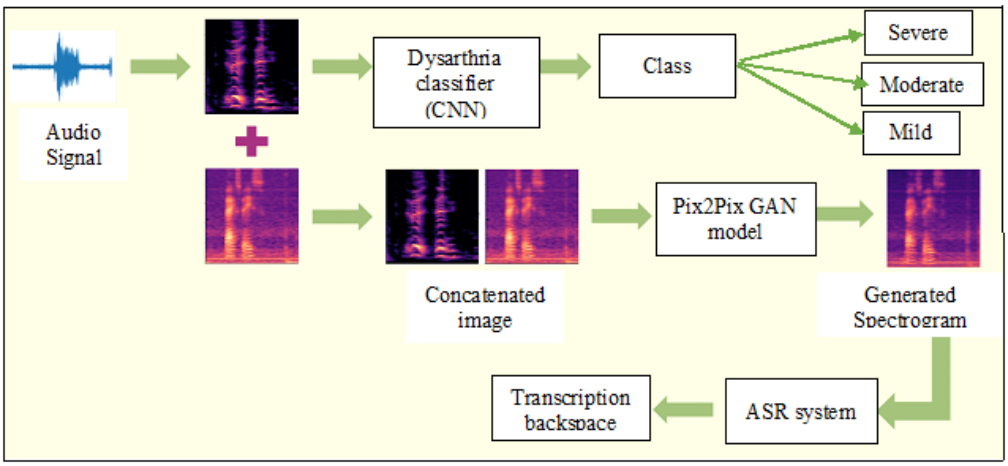

Fig 3.7: Diagram of the Pix2Pix GAN for dysarthric speech recognition.

Optimization and evaluation

- **Optimizer:** Adam optimizer with an initial learning rate of 0.0002.
- **Monitoring:** The losses of both the generator and discriminator are monitored during training.
- **Batch Size:** Set to 1, which is effective for image generation tasks.

Generation of new spectrograms

- **Process:** After training, the model can generate new healthy spectrograms by simply providing a pathological spectrogram as input to the trained generator. The Fig. 3.8 illustrates an example of pathological, healthy, and generated images for speaker F02 from the Pix2Pix model.

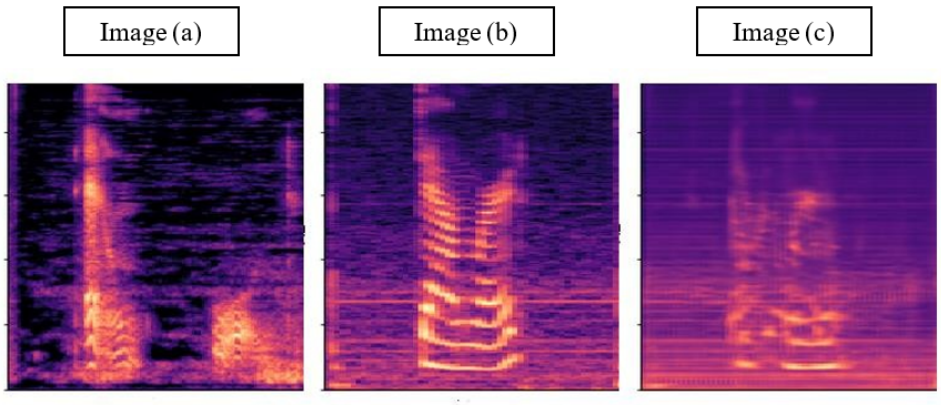

Fig 3.8: An example of pathological image (a), healthy image (b), and generated image (c) for speaker F02 using the Pix2Pix GAN model.

In essence, the Pix2Pix model leverages a U-Net architecture to learn the translation from pathological to healthy spectrograms by analyzing the relationships between paired images. This allows it to generate realistic and informative healthy spectrograms.

B) Data preprocessing

- **Spectrogram handling:**
  - Conversion to grayscale: Simplifies processing and reduces computational cost.
  - Normalization: Standardizes pixel values, improving model learning.
- **Transcription handling:**
  - Conversion to lowercase and numeric representation: Enables model compatibility.
  - Creation of a character dictionary: Defines the set of possible characters.

- **Data splitting:**
  - Division into training, validation, and test sets for model evaluation.
- **Dataset preparation:**
  - Creation of tuples containing file names and corresponding transcriptions.
  - Batching and padding: Ensures consistent input data for the model.

C) ASR model's architecture (Fig. 3.9)

- **Convolutional layers:** Extract spatial features from spectrograms using 3x3 filters.
- **Recurrent layers (GRU):** Capture temporal dependencies within the spectrogram sequences.
- **Dense layer:** Transforms extracted features into a compact representation.
- **Output Layer:** Produces probabilities for each character in the transcription using softmax activation.

D) Hyperparameter tuning

Table. 3.17 outlines the hyperparameters based on empirical experiments to optimize the performance of the ASR system (CNN-RNN).

Table 3.17: Hyperparameter configuration

Model parameters	Value
Number of epochs	200
Batch size	32
Input dimensions	[128, 128]
Learning rate	0.0004
Splitting of training and test data	70/30
Training time (GPU Google Colab) using python 3.11 software	6H

We chose to train our system on the Google Colab GPU using Python due to the significant computational demands required by the model. The input of the system is

a spectrogram image of size 128x128, which involves processing a large amount of data. Training such models on a CPU would have resulted in excessive computation time, making it impractical for efficient training.

E) Comparative results

Table. 3.18 presents the accuracy rates (Acc), Character Error Rate (CER), and Word Error Rate (WER) for various speakers (M04, F02, F04, and M09) across two systems:

1. **Proposed system CNN-BiLSTM + MMSE:** Utilizes a CNN-BiLSTM architecture with Multivariate Analysis (MFCC_PNCC) and speech quality enhancement using the MMSE technique.

2. **Pix2Pix GAN + Baseline ASR System:** Employs a Pix2Pix GAN for image-to-image translation followed by a baseline ASR system.

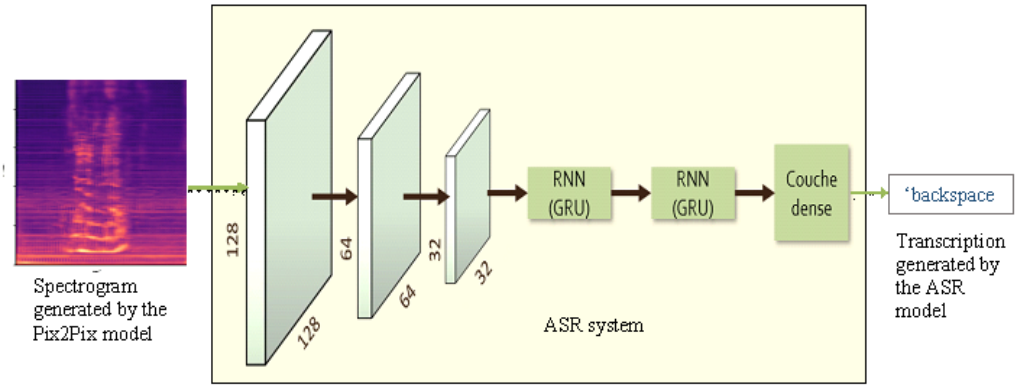

Fig 3.9: Architecture of the ASR model.

Observations:

- **CNN-BiLSTM generally outperforms Pix2Pix GAN:** Notably, for speakers with low intelligibility levels (F02, F04, and M04), the proposed CNN-BiLSTM + MMSE system consistently achieves higher accuracy rates.
- **Speaker-specific performance:**
  - The proposed system excels in improving accuracy for speakers with low intelligibility, particularly speaker M04, showing a 23% precision improvement compared to a CNN-RNN system using the Pix2Pix GAN.
  - For speaker M09, the Pix2Pix GAN + Baseline ASR approach demonstrates superior performance in terms of precision, CER, and WER.

This comparison allows us to retain the following points:

- The combination of CNNs, BiLSTMs, and the MMSE technique within the proposed system effectively enhances speech recognition accuracy for speakers with low intelligibility.
- The choice of the optimal system may depend on the specific characteristics of the speaker and the desired performance metrics.

Table 3.18: Comparison of results between CNN-BiLSTM with MMSE and Pix2Pix GAN with baseline ASR Systems

System	Acc				CER				WER			
	F02	F04	M04	M09	F02	F04	M04	M09	F02	F04	M04	M09
CNN-BiLSTM	**87.47**	**80.91**	**80.39**	85.49	**10.51**	**15.47**	**15.90**	12.04	**12.53**	**19.09**	**19.61**	14.51
Pix2Pix GAN	80.78	70.59	57.98	**92.51**	10.52	19.67	30.81	**04.39**	19.21	29.41	42.02	**7.16**

3.5 Discussion of results

3.5.1 Summary of results and their relevance

This section outlines and examines the key findings from the analysis of the proposed system for pathological speech classification and dysarthric speech recognition.

- **Effectiveness of multi-stream approach and speech enhancement**
  The integration of the multi-stream approach, combining MFCC, Jitter, Shimmer, and PNCC features, significantly improved classification accuracy. As shown in Table 3.5, the proposed approach (MFCC-Jitter-Shimmer-PNCC) achieved an accuracy of 85.71% for the first dataset (nodules, paralysis, and polypoid lesions). This demonstrates the effectiveness of incorporating multiple features to capture different aspects of the pathological voice signal. Using a simple configuration (70/30) for splitting the training and testing data further improved the system's accuracy to 86.67%.

  Furthermore, the speech enhancement module using the MMSE technique in the frequency range of 0 to 3.5 kHz played a crucial role in improving accuracy. The highest classification rate of 95.24% for the first dataset highlights the benefit of denoising techniques in pathological speech analysis.

- **Superiority of CNN-BiLSTM architecture**
  Compared to the DNN-based system, the CNN-BiLSTM architecture achieved superior performance (Table. 3.11). This is attributed to the ability of CNNs to extract local features and BiLSTMs to capture long-term temporal dependencies within the speech signal. DNNs may require a large number of layers to capture similar information, leading to increased computational complexity and overfitting risks. The combination of CNN and BiLSTM often reduces the number of parameters needed while maintaining high performance due to their specialized functionalities in different parts of the classification task.
- **Importance of balanced datasets and AReLU activation function**
  Experiments on the second dataset (anteroposterior compression, gastric reflux, and mild ventricular compression) balanced for gender and age demonstrated the importance of fairness and generalizability. The AReLU activation function provided the best performance within this dataset, further improving accuracy for pathological classification.

3.5.2 Benchmarking against previous work

To evaluate the performance of our proposed system, we conducted a comparative study with the system described in [56]. Both systems used the MEEI database for classification of four pathological classes: nodules, spasmodic dysphonia, polypoid lesions, and normal voice.

Table 3.19: System performance comparison for pathological voice classification

Proposed System	**Accuracy (%)**
CNN + multi-stream	88.57
CNN + multi-stream + MMSE35	91.67
CNN-BiLSTM + multi-stream+ MMSE 35	**100**
System from [56].	94.44

Our system, based on the CNN-BiLSTM architecture, multi-stream approach, and MMSE 35 enhancement [95], significantly outperformed the system proposed in [56]. This demonstrates the effectiveness of our approach for multi-class pathological voice classification (Table. 3.19).

While the system in [56] achieved an accuracy of 94.44% for binary classification, our system's 100% accuracy for multi-class classification is a notable achievement. This highlights the advantages of the CNN-BiLSTM architecture and the multi-stream approach in handling the complexity of multi-class tasks.

Several studies have reported comparable or lower accuracy rates for pathological voice classification. For instance, systems employing HMM-GM [56], CNN-LSTM [103], CNN-RNN [20], or 1D-CNN-LSTM [104] architectures have reported accuracies ranging from 68.08% to 99.58%. These findings underscore the competitive performance of our proposed system.

For more details, our system, based on a multi-stream approach, initially achieved 88.57% accuracy with a CNN network. After applying the MMSE 35 speech enhancement technique, the accuracy improved to 91.67%. Additionally, when using the CNN-BiLSTM combination, the accuracy reached 100%, representing a 6% improvement over the system in [56].

These results are satisfactory, considering that multi-class identification is more complex than binary classification. For example, the multi-model CNN-LSTM proposed in [103] achieved 99.58% accuracy using the SVD database for binary classification of pathological and healthy voices. Similarly, the system in [20], based on a multi-model CNN-RNN with a two-level cascade architecture, achieved 88.83% accuracy for binary classification. In [104], a system using a 1D-CNN-LSTM architecture with segmented raw signals as input reached 68.08% accuracy for binary classification of vocal pathologies using the SVD database.

In [105], researchers used spectrogram features and a network of 2D-CNN and fully connected layers, pre-trained by a deep belief network (DBN), to identify organic dysphonia (binary task). The accuracy rates were 71% and 77%, with and without pre-training, respectively.

In [22], Gammatone Spectral Latitude (GTSL) coefficients were used to classify healthy, neuromuscular, and structural voices, achieving 99.6% accuracy on the MEEI database, and 89.9% and 97.4% accuracy on the SVD and HUPA databases, respectively. In [106], phase space reconstruction and a CNN were used to classify normal and pathological voices, achieving 96.04% accuracy on the MEEI database and 92.27% on the SVD database [107].

Compared to these systems, our three-phase architecture for multi-class voice disorder classification achieved 95.24% accuracy, demonstrating its effectiveness in identifying multiple pathological classes.

To enhance the effectiveness of our proposed system, we integrated the Connectionist Temporal Classification (CTC) approach for dysarthric word recognition. CTC is particu-

larly advantageous for speech recognition tasks as it enables training sequence-to-sequence models without the need for precise alignment between input features and output labels.

To evaluate its robustness, we compared our system with a model incorporating the Pix2Pix GAN, which improves spectrogram quality by mitigating noise and distortions in the speech signal. Results demonstrate the superior performance of our approach, particularly for speakers with low intelligibility levels, achieving a significant 23% improvement.

3.5.3 Limitations and opportunities

This study acknowledges some limitations and areas for improvement:

- **Data Availability:** The limited number of voice recordings for certain speech disorders constrains the quality and quantity of training and evaluation data, potentially affecting model performance. Expanding datasets for diverse speech pathologies is essential for robust development.
- **Challenges in Speech recognition:** Variability in speech, intelligibility issues, and individual user needs present significant challenges. Addressing these complexities is crucial for improving the generalizability and usability of recognition systems.
- **Data and Generalizability Challenges:** The study employed the MEEI and UASpeech databases to design a robust speech disorder classification and dysarthric speech recognition system. However, variability in speech patterns and recording conditions poses challenges to the system's generalizability.

3.6 Conclusion

This chapter presents a comprehensive evaluation of the proposed pathological speech recognition system. We explored the impact of multi-variable acoustc features, speech enhancement techniques, and the CNN-BiLSTM architecture on system performance.

Key findings include:

- **Multi-stream feature fusion:** The combination of MFCC, Jitter, Shimmer, and PNCC features significantly improved classification accuracy.
- **Speech enhancement:** The MMSE technique, particularly MMSE 35, effectively enhanced speech quality and boosted classification performance.
- **CNN-BiLSTM superiority:** The CNN-BiLSTM architecture demonstrated superior performance compared to the DNN-based system and the Pix2Pix GAN model

with baseline ASR, proving its effectiveness for both pathological speech classification and dysarthric speech recognition.

Overall, the proposed system achieved state-of-the-art performance, demonstrating its potential for real-world applications in diagnosing and monitoring speech disorders.

General Conclusions

Pathological speech disorders present considerable obstacles to effective communication and substantially impact quality of life, particularly when stemming from conditions such as dysarthria, dysphonia, and stuttering. Traditional diagnostic methods for these disorders often depend on subjective assessments, which can limit accuracy and efficiency. To address these limitations, this thesis has introduced an innovative, automated system for classifying pathological speech.

The primary objective of this research was to improve the accuracy and efficiency of diagnosing speech disorders that affect intelligibility and comprehensibility. To achieve this, a multi-faceted approach was adopted, incorporating speech enhancement techniques, advanced feature extraction, and a robust deep learning architecture. The proposed system combines convolutional neural networks (CNNs) and bidirectional long short-term memory (BiLSTM) networks, effectively capturing both local and temporal dependencies in the speech signal.

Experimental results demonstrated the efficacy of this system in accurately classifying a range of pathological speech conditions. By integrating multiple acoustic features—such as MFCCs, PNCCs, Mel-spectrograms, Jitter, and Shimmer—the system achieved a comprehensive representation of the speech signal, which led to enhanced classification performance. Additionally, the application of speech enhancement techniques substantially reduced the influence of noise and distortion, further bolstering system accuracy.

Validated using the Massachusetts Eye and Ear Infirmary (MEEI) database, the system demonstrates notable improvements in classification accuracy and computational efficiency compared to previous methods. Extensive experimentation and comparative analysis underscore the effectiveness of the multi-variable approach, revealing that a combination of features yields more accurate and consistent classification results than single-stream approaches.

The integration of the same system into the task of dysarthric speech recognition and its comparison with an advanced system based on the Pix2Pix technique for generating healthy spectrograms demonstrated the effectiveness and improvement of our approach aimed in this work, which allows individuals with dysarthria to communicate more effectively by converting their speech into text with higher accuracy, even if their intelligibility is low.

The contributions of this thesis are multifaceted. They include an advanced feature extraction process, robust speech enhancement techniques, a powerful CNN-BiLSTM model architecture, and a comprehensive evaluation framework that underscores the system's generalizability and performance. These elements collectively provide a scalable model, laying a foundation for future research in automated pathological speech analysis and offering a significant step forward in supporting clinical professionals.

The potential impact of this research is considerable. By providing accurate and timely diagnoses, this system enables early intervention and supports personalized treatment planning. It can also assist healthcare professionals in monitoring disease progression and evaluating therapeutic interventions' effectiveness. Although the system has shown promising results, there remain several opportunities for future research. These include investigating advanced deep learning architectures, such as transformers, to capture even finer nuances in pathological speech patterns and further improve recognition accuracy. Integrating multimodal information—such as visual cues from lip movements—may enhance performance, while developing user-friendly interfaces would facilitate deployment in clinical environments. Expanding the dataset to encompass a wider variety of speech pathologies and conditions could enhance the model's generalization across diverse populations. Additionally, developing real-time processing capabilities would increase the system's applicability in clinical settings, enabling immediate feedback and intervention options for individuals with speech disorders.

In conclusion, this thesis has made significant contributions to the field of speech pathology by contribute a valuable tool for the early and accurate classification of pathological speech and recognition of dysarthric speech, with potential applications in clinical diagnostics and personalized treatment planning. By addressing the challenges inherent in speech disorder assessment and capitalizing on deep learning innovations, this thesis paves the way for more accessible, reliable, and automated speech analysis solutions that can ultimately improve patient outcomes and quality of life.

Bibliography

[1] Léon, P. (2011). Chapitre 5. La production des sons de la parole. Phonétisme et prononciations du français. (p. 73 -89). Armand Colin. https://doi.org/10.3917/arco.leon.2011.01.0073.

[2] Segui, J., Ferrand, L. (2000). Chapitre IV. Le système phonatoire et l'articulation. Leçons de parole. (p. 91 -101). Odile Jacob. https://shs.cairn.info/lecons-de-parole--9782738107619-page-91?lang=fr.

[3] Zhang, Z. (2016). Mechanics of human voice production and control. J Acoust Soc Am. Oct;140(4):2614.

[4] Chen, J. C. (2016). Elements of human voice. World scientific publishing co.

[5] Noordzij, J. P., Ossoff, R.H. (2006). Anatomy and physiology of the larynx. Otolaryngol Clin North Am. 39(1):1-10.

[6] Daly, I., Hajaiej, Z., and Gharsallah, A. (2018). Physiology of Speech/Voice Production. Journal of Pharmaceutical Research International. 23, no. 3:1–7. http://dx.doi.org/10.9734/jpri/2018/42403.

[7] Karagama, Y. G., McGlashan, J. A. (2018). Structural Disorders of the Vocal Cords. Scott-Brown's Otorhinolaryngology and Head and Neck Surgery.

[8] Carding, P., Bos-Clark, M., Fu, S., Gillivan-Murphy, P., Jones, S. M., Walton, C. (2017). Evaluating the efficacy of voice therapy for functional, organic and neurological voice disorders. Clin Otolaryngol. 42(2):201-217.

[9] Duffy, J. R. (2019). Motor speech disorders: Substrates, differential diagnosis, and management. Elsevier Health Sciences.

[10] Jayaraman, D.K., Das, J. M. (2023). Dysarthria. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan–. PMID: 37279355.

[11] Hancock, D.B., Martin, E. R., Mayhew, G. M., Stajich, J. M., Jewett, R., Stacy, M. A., Scott, B. L., Vance, J. M., Scott, W. K. (2008). Pesticide exposure and risk of Parkinson's disease: a family-based case-control study. BMC Neurol. 28;8:6.

[12] Nandipati, S., Litvan, I. (2016). Environmental Exposures and Parkinson's Disease. Int J Environ Res Public Health. 13(9):881. doi: 10.3390/ijerph13090881.

[13] Panek, Daria, et al. (2015). Acoustic analysis assessment in speech pathology detection. International Journal of Applied Mathematics and Computer Science. vol. 25, no. 3, Sciendo, pp. 631-643.

[14] Wu, J. D., Lin, B. F. (2009). Speaker identification based on the frame linear predictive coding spectrum technique. Expert Systems with Applications. Volume 36, Issue 4, Pages 8056-8063.

[15] Panek, D., Skalski, A., Gajda, J. R. (2015). Tadeusiewicz. Acoustic analysis assessment in speech pathology detection. International Journal of Applied Mathematics and Computer Science. pp 631-643.

[16] Alsulaiman, M. (2014). Voice Pathology Assessment Systems for Dysphonic Patients: Detection, Classification, and Speech Recognition. IETE Journal of Research, 60(2), pp 156–167.

[17] 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, pp 463-480.

[18] Mohammed, H. M. A., Omergolu, A. N., Oral, E. A. (2023). MMHFNet: Multimodal and multi- layer hybrid fusion network for voice pathology detection. Expert Systems and Applications. 223(119790).

[19] Ankışhan, H., İnam, S. C. (2021). Voice pathology detection by using the deep network architecture. Applied Soft Computing. 106 (107310).

[20] 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.

[21] 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, pp 25-33.

[22] 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.

[23] Peng, X., Xu, H., Liu, J. et al. (2023). Voice disorder classification using convolutional neural network based on deep transfer learning. Sci Rep 13, 7264.

[24] Kathard, H., Naude, E., Pillay, M & Ross, E. (2007). IMPROVING THE RELEVANCE OF SPEECH-LANGUAGE PATHOLOGY & AUDIOLOGY RESEARCH AND PRACTICE, The South African Journal o f Communication Disorders. Vol. 54.

[25] Ishikawa, K., Webster, J., & Ketring, C. (2021). Agreement between Transcription- and Rating-Based Intelligibility Measurements for Evaluation of Dysphonic Speech in Noise. Clinical Linguistics & Phonetics. 35(10), pp 983–995.

[26] Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. J Speech Lang Hear Res. 51(3):562-73.

[27] Sáenz-Lechón, N., Godino-Llorente, J. I., Osma-Ruiz, V., Blanco-Velasco, M., Cruz-Roldán, F. (2006). Automatic assessment of voice quality according to the GRBAS scale. Conf Proc IEEE Eng Med Biol Soc. 2006:2478-81.

[28] Zhaopeng, Q., Xiao, K. (2023). A Survey of Automatic Speech Recognition for Dysarthric Speech. Electronics. 12.20, 4278.

[29] Rudzicz, F. (2011). Articulatory knowledge in the recognition of dysarthric speech. IEEE Transactions on Audio Speech and Language Processing. 19(4):947 - 960.

[30] Shahamiri, S. R, Salim, S. S. (2014). A multi-views multi-learners approach towards dysarthric speech recognition using multi-nets artificial neural networks. IEEE Trans Neural Syst Rehabil Eng. 22(5):1053-63. doi: 10.1109/TNSRE.2014.2309336.

[31] Takashima, Y., Nakashika, T., Takiguchi, T., Arikii, Y. (2015). Feature extraction using pre- trained convolutive bottleneck nets for dysarthric speech recognition. In Proceedings of the 23rd European Signal Processing Conference (EUSIPCO). pp. 1411–1415.

[32] Bhat, C., Vachhani, B., Kopparapu, S. (2016). Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-Taper ral Estimation. In Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTER-SPEECH 2016). pp. 228–232.

[33] Kim, M., Kim, Y., Yoo, J., Wang, J., Kim, H. (2017). Regularized speaker adaptation of KL- HMM for dysarthric speech recognition. IEEE Trans. Neural Syst. Rehabil. 1581–1591.

[34] Yu, J.W., Xie, X.R., Liu, S.S., Hu, S.K., Lam, M.W.Y., Wu, X.X., Wong, K.H., Liu, X.Y., Meng, H. (2018). Development of the CUHK Dysarthric Speech Recognition System for the Speech Corpus. In Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018). pp. 2938–2942.

[35] Xiong, F., Barker, J., Christensen, H. (2019). Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric Speech recognition. In Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 836–5840.

[36] Wu, L. D., Zong, D.M., Sun, S.L., Zhao, J. (2021). A Sequential Contrastive Learning Framework for Robust Dysarthric Speech Recognition. In Proceedings of the ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7303–7307.

[37] Zaidi, B. F., Selouani, S.A., Boudraa, M., Yakoub, M.S. (2021). Deep neural network architectures for dysarthric speech analysis and recognition. Neural Computing and Applications. Vol 33. pp. 9089-9108.

[38] Shahamiri, S. R. (2021). Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. IEEE Trans Neural Syst Rehabil Eng. 29:852-861.

[39] Hu, S. K., Xie, X. R., Cui, M. Y., Deng, J. J., Liu, S. S., Yu, J. W., Geng, M. Z., Liu, X. Y., Meng, H. E. (2022). Neural architecture search for LF-MMI trained delay neural networks. IEEE/ACM Transactions on Audio Speech and Language Processing. Vol 30, pp 1093–1107.

[40] Almadhor, A., Irfan, R., Gao, J., Saleem, N., Rauf, H. T., Kadry, S. (2023). E2E-DASR: End-to-end deep learning-based dysarthric automatic speech recognition. Expert Systems with Applications. Vol 222, 119797.

[41] Wang, H., et al. (2024). Enhancing Pre-Trained ASR System Fine-Tuning for Dysarthric Speech Recognition Using Adversarial Data Augmentation. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 12311-12315.

[42] Sanez, N., Juan, I., Victor, O., Pedro, G. (2006). Methodological issues in the development of automatic systems for voice pathology detection. Biomedical Signal Processing and Control. Vol. 1, pp 120–128.

[43] Eye, M., Infirmary, E.: Voice disorders database. (1994). version. 1.03 (cd-rom). Lincoln Park, NJ: Kay Elemetrics Corporation.

[44] Loizou, P. C. (2013). Speech enhancement: Theory and practice. (2nd ed.). CRC Press.

[45] Lu, C.T., Shen, J.H., Tseng, Kun-Fu. (2011). Speech enhancement using three-step decision gain factor with optimal smoothing. International Journal of Electrical Engineering. Vol 18, pp 209-221.

[46] Borrie, S. A., Baese, B. M., Engen, K. V., Bent, T. (2017). A relationship between processing speech in noise and dysarthric speech. J Acoust Soc Am. 141(6):4660–7.

[47] Ghanbari, Y., Karami, M. (2004). Spectral subtraction in the wavelet domain for speech enhancement. International Journal of Software & Information Technology (IJSIT). Vol 1, pp 26–30.

[48] Karam, M., Khazaal, H., Aglan, H. and Cole, C. (2014). Noise Removal in Speech Processing Using Spectral Subtraction. Journal of Signal and Information Processing. Vol 5, pp 32-41.

[49] Jaiswal, R., Romero, D. (2021). Implicit Wiener Filtering for Speech Enhancement In Non-Stationary Noise. 11th International Conference on Information Science and Technology (ICIST). pp 39-47.

[50] Ayat, S., Manzuri, M. T., Dianat, R. (2004). Wavelet based speech enhancement using a new thresholding algorithm. Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing. pp. 238-241.

[51] Ravi, B. R., Deepu, S. P., Ramesh Kini, M., Sumam, D. S. (2018). Wavelet based Noise Reduction Techniques for Real Time Speech Enhancement. 5th International Conference on Signal Processing and Integrated Networks (SPIN). pp. 846-851.

[52] Gutiérrez-Muñoz, M., Coto-Jiménez, M. (2022). An Experimental Study on Speech Enhancement Based on a Combination of Wavelets and Deep Learning. Computation. 10, 102.

[53] Ding, P., He, L., Yan, X., Zhao, R., Hao, J. (2006). Optimizing the Implementation of MMSE Enhancement for Robust Speech Recognition. Proc. International Symposium on Chinese Spoken Language Processing. pp 318-327.

[54] Berouti, M., & Schwartz, R. (2022). Enhancement of speech in additive noise based on spectral subtraction. Journal of the Acoustical Society of America. 151(6), pp 4237-4248.

[55] Zhang, Q., Nicolson, A., Wang, M., Paliwal, K. K., Wang, C. (2020). DeepMMSE: A Deep Learning Approach to MMSE-Based Noise Power Spectral Density Estimation. In IEEE/ACM Transactions on Audio, Speech, and Language Processing. Vol. 28, pp. 1404-1415.

[56] 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), pp 82-87.

[57] Al-Qatab, B. A., Mustafa, M. B. (2021). Classification of Dysarthric Speech According to the Severity of Impairment: an Analysis of Acoustic Features. In IEEE Access. Vol. 9, pp. 18183- 18194.

[58] Mu, X. and Min, C. H. (2023). MFCC as Features for Speaker Classification using Machine Learning. IEEE World AI IoT Congress (AIIoT). pp. 0566-0570.

[59] Kishore, P. (2011). Topic: Spectrogram, Cepstrum and Mel-Frequency Analysis: Speech Technology: A Practical. https://www.cs.brandeis.edu/~cs136a/CS136a_docs/KishorePrahallad_CMU_mfcc.pdf

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

[61] 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), pp 1315–1329.

[62] Moreno, P. J., Raj, B., Stern, R. M. (1996). A vector Taylor series approach for environment- independent speech recognition. IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. Vol. 2, pp 733-736.

[63] Li, J. Y., Liu, B., Wang, R. H., Dai, L. R. (2004). A complexity reduction of ETSI advanced front-end for DSR. IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 61-64.

[64] Chung, J. S., & Lee, S. W. (2021). Speech enhancement using gammatone filter bank and deep learning-based denoising. IEEE Transactions on Audio, Speech, and Language Processing. Vol 29, pp 2715-2727.

[65] Gandia, D., Castro, S., Martínez, M., & León, A. (2023). Acoustic and Vocal Biomarkers in Children with Speech Disorders: A Review of Contemporary Approaches. Journal of Voice. 37(3), 477.e1-477.e10.

[66] Tóth, M. A., & Farkas, T. (2023). Objective and perceptual measures of jitter in voice analysis: A comparative study. Journal of Voice. 37(5), pp 719-727.

[67] 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), pp 44–53.

[68] Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal acoustic analysis jitter, shimmer and hnr parameters. Procedia Technology. 9(5), pp 1112–1122.

[69] Teixeira, J. P., Gonçalves, A. (2016). Algorithm for jitter and shimmer measurement in pathologic voices. Procedia Computer Science 100. 271 – 279

[70] Gómez, S., & Escudero, D. (2023). Shimmer and jitter analysis in voice disorders: A comparative study of objective measurements and clinical outcomes. Journal of Speech, Language, and Hearing Research. 66(7), pp 2111-2120.

[71] Molau, S. (2003). Normalization in the acoustic feature space for improved speech recognition.

[72] Bengio, Y., et al. (2009). Learning Deep Architectures for AI. Foundations and Trends in Machine Learning. Vol. 2, No. 1, 1–127.

[73] Rekha R, Tharani R S. (2021). Speech Emotion Recognition using Multilayer Perceptron Classifier on Ravdess Dataset. Proceedings of the First International Conference on Combinatorial and Optimization, ICCAP.

[74] Tajdini, F., Piri, M. (2022). Deep neural network learning for speech recognition. 9 th National Congress of Electrical and Computer Engineering of Iran.

[75] Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Trans Neural Netw Learn Syst. 33(12), pp 6999-7019.

[76] Wu, H., Soraghan, J., Lowit, A., and Di Caterina, G. (2018). Convolutional Neural Networks for Pathological Voice Detection. 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 1-4.

[77] Hasib, K. Md et al. (2024). DCNN: Deep Convolutional Neural Network With XAI for Efficient Detection of Specific Language Impairment in Children. In IEEE Access. Vol. 12, pp. 101660-101678.

[78] Zhao, X., Wang, L., Zhang, Y. et al. (2024). A review of convolutional neural networks in computer vision. Artif Intell Rev. 57(4):57-99.

[79] Zhang, Y., Liu, X., & He, J. (2023). End-to-end speech recognition using BiLSTM and attention mechanism for noisy environments. IEEE Transactions on Speech and Audio Processing. 31(4), pp 711-723.

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

[81] Belabbas, S., Addou, D., and Selouani, S. A. (2023). Improved ASR System based on CNN-LSTM Acoustic Model for Mobile Voice. 2nd International Conference on Electronics, Energy and Measurement (IC2EM). pp. 1-6.

[82] Gales, M. J. F., Young, S. (2007). The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing1. (3):195-304.

[83] Karkos, P. D., McCormick, M. (2009). The etiology of vocal fold nodules in adults. Curr Opin Otolaryngol Head Neck Surg. 17(6), pp 420-423.

[84] Toutounchi, S. J. S., Eydi, M., Ej Golzari, S., Ghaffari, M. R., Parvizian, N. (2014). Vocal Cord Paralysis and itsEtiologies: A Prospective Study. J Cardiovasc Thorac Res. 6(1), pp 47–50.

[85] 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. J Voice. Vol30, pp698-704.

[86] Vakil, N., van Zanten, S. V., Kahrilas, P., Dent, J., Jones, R, Global Consensus Group. (2006). The Montreal definition and classification of gastroesophageal reflux disease: a global evidence-based consensus. Am J Gastroenterol. 101(8), 1900-20.

[87] 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), pp 403-410.

[88] 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), pp 1219:1242.

[89] Kent, R.D., Kim, Y. (2009). Acoustic Analysis of Speech. The Handbook of Clinical Linguistics. pp 360-380. DOI: 10.1002/9781444301007.ch22 .

[90] Chen, L., Chen, J. (2022). Deep Neural Network for Automatic Classification of Pathological Voice Signals. Journal of Voice. 36(2), 288.e15-288.e24.

[91] Belabbas, S., Addou, D. (2022). Weighting Schemes Based Discriminative Model Combination Technique for Robust Speech Recognition. In: Hatti, M. (eds) Artificial Intelligence and Heuristics for Smart Energy Efficiency in Smart Cities. IC-AIRES 2021. pp 430–438.

[92] Pouchoulin, G., Fredouille, C., Bonastre, J. F., Ghio, A., & Giovanni, A. (2007). Frequency study for the characterization of the dysphonic voices. In Interspeech. pp 1198-1201.

[93] Hendrycks, D., Gimpel, K. (2016). Bridging nonlinearities and stochastic regularizers with gaussian error linear units. CoRR abs/1606.08415. arXiv:1606.08415.

[94] Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S. (2017). Self-normalizing neural networks. In: Proceedings of the 31st international conference on neural information processing systems. pp. 972–981.

[95] Belabbas, S., Addou, D. & Selouani, S. A. (2024). Pathological voice classification system based on CNN-BiLSTM network using speech enhancement and multi-stream approach. Int J Speech Technol. Vol 27, pp 483–502.

[96] Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T.S., Watkin, K., Frame, S. (2008). Dysarthric speech database for universal access research. Proc. Interspeech. pp 1741-1744.

[97] Hannun, A. (2017). Sequence modeling with CTC. Distill. 2(11).

[98] Mikko, K., Mathias, C., Matti, V., Ebru, A., and Murat, S. (2006). Unsupervised segmentation of words into morphemes morpho challenge 2005 application to automatic speech recognition. In Proc. Interspeech. Pages paper1512–Tue2A2O.1.

[99] Ali, A., and Renals, S. (2018). Word error rate estimation for speech recognition: e-WER. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Volume 2: Short Papers.

[100] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2020). Unpaired image-to-image translation using cycle-consistent adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 42(2), pp 321-331.

[101] Graves, A., Fernández, S., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine Learning (ICML). pp 369-376.

[102] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2022). Image-to-image translation with conditional adversarial networks. IEEE Transactions on Computer Vision and Pattern Recognition. 35(2), pp 329-337.

[103] Ankışhan, H., İnam, S. C. (2021). Voice pathology detection by using the deep network architecture. Applied Soft Computing. 106 (107310).

[104] Harar, P., Alonso-Hernandezy, J. B., Mekyska, J., Galaz, Z., Burget, R., Smekal, Z. (2019). Voice pathology detection using deep learning: a preliminary study. International Conference and Workshop on Bioinspired Intelligence (IWOBI). pp 1–4.

[105] Wu, H., Soraghan, J., Lowit, A., Di-Caterina, G. (2018). A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Networks. Proc. Interspeech. pp 446–450.

[106] Deli, F., Xuehui, Z., Dandan, C., Weiping, H. (2022). Pathological Voice Detection Based on Phase Reconstitution and Convolutional Neural Network. Journal of Voice.

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