Detecting type and severity of speech impairment using deep-learning and machine learning algorithms

MSc Research Project
MSc in Data Analytics

Ankit Chatterjee
Student ID: x21169993

School of Computing
National College of Ireland

Supervisor: Hicham Rifai

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ankit Chatterjee
Student ID:	x21169993
Programme:	MSc Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Hicham Rifai
Submission Due Date:	14/08/2023
Project Title:	Detecting type and severity of speech impairment using deep-learning and machine learning algorithms
Word Count:	7045
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Ankit Chatterjee	
Date:	14th August 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Detecting type and severity of speech impairment using deep-learning and machine learning algorithms

Ankit Chatterjee
21169993

Abstract

Speech forms an important aspect of human life as it is one of the main forms of communication. Speech impairment refers to a condition that affects a person's verbal communication. There are several types of speech impediments like stuttering/stammering, aphasia, dysarthria, mutism, etc. Previous studies have used distinct deep-learning methods like neural networks to explore the area of speech recognition and emotion classification using audio data. However, the aspect of classification of the different types of speech disorders still needs to be analyzed in depth. This paper not only attempts to classify speech disorders through audio files but also attempts to detect the severity of those disorders.

***Keywords*— Long Short-Term Method, Recursive Neural Network, Empirical Mode Decomposition, Wavelet Transform, MFCC, Savitzky Golay Smoothing Technique**

1 Introduction

Communication forms a cornerstone in all human relationships. When an individual struggles to express any information verbally, it causes hindrance in personal as well as professional life. Speech impairment or disorder refers to a person's inability to articulate sounds which in turn causes a hindrance in verbal communication. Speech impediments can be of different types and can vary from dysarthria, mutism to more serious conditions like stuttering or stammering. In addition, speech impairments can range widely from speech disorders to voice disorders. Speech disorders are usually functional with or without any known cause, on the other hand, voice disorders are caused by physical issues that limit speech. As per the National Institute on Deafness and Other Communication Disorders(2010), globally nearly 18.5 million people suffer from speech, voice, or language-related disorders. There are several categories of speech disorders like fluency disorders, articulation disorders, resonance disorders and phonological disorders. Detection of speech impediments relies on the expertise of speech therapists in general. But with the advancement in technology, various deep learning algorithms were used to detect the type of speech impairments. Some of these deep learning automated techniques showed promising results in detecting and analyzing different speech disorders. These algorithms require datasets of voice recordings to learn different speech patterns and features which may indicate speech disfluencies and abnormal speech. These techniques can be used in the early detection of these disorders which can result in timely intervention and proper treatment. Having said that, it is crucial to note that even with the recent advancements, deep learning can only complement the process of speech recognition and speech disorder detection. The assessment of such impediments still requires clinical judgments and the expertise of trained professionals.

This paper provides a comprehensive study on the application of deep learning techniques to detect speech impediments using raw audio data. This study focuses on feature extraction and feature selection to detect abnormalities in speech. Feature extraction techniques such as Empirical Mode Decomposition, Wavelet Transform, and Mel-frequency cepstral coefficients are employed to detect speech patterns from the raw audio data. As part of this research study, we implement a methodology of Recurrent Neural Networks based on Long short-term memory (RNN-LSTM) architecture. This paper will demonstrate how the most relevant and effective set of extracted features like spectrograms, mean and variance will be used to pre-process the data before being fed into the LSTM-RNN model.

Research Question. The above research problem motivates the following research question:

How well the Long Short-Term Memory neural networks and K-Means Clustering can detect the type and severity of speech impairment respectively using feature extraction techniques like empirical mode decomposition and wavelet transform?

This paper implements the LSTM model for speech or voice disorder detection and uses K-Means clustering algorithm to detect the severity of the speech disorders. This research paper covers the following sections:

Literature Review

This section involves a critical analysis of the previous and related works in the field of speech recognition and speech disorder detection using deep learning techniques.

Research Methododology

This section would comprise the implementation and methodological techniques of the research to be conducted. The section will demonstrate all the steps to answer the research question and achieve the objective of the study.

Evaluation and Results

This section consists of the evaluation metrics of the performance of the LSTM-RNN model and the discussion regarding the same. We will be evaluating the model's performance in two different experiments. The first model will consist of two feature extraction techniques- Wavelet Transform and Empirical Mode Decomposition along with Mel frequency cepstral coefficients and in the second experiment, we measure the performance of the model by a single feature extraction technique i.e., wavelet transform.

2 Related Work

The aim of this research study is to detect the type and severity of speech impairments in participants. In the past, there have been a substantial amount of studies related to Automated Speech Recognition(ASR) and speech or voice disorder detection.

In this section different studies have been discussed, analyzed, and categorized below as per their respective areas of research and proposed solutions.

2.1 Alzheimer's related Speech Disorders

Detection of Alzheimer's through speech analysis has proved to be more convenient and cost-effective than methods like magnetic resonance imaging and electroencephalograms. The study Lin et al. (2022) talks about an automated screening system that can be deployed in large-scale and cost-effective screening methods. Alzheimer's Disease (AD) and other forms of dementia have become very common across the globe. The paper Lin et al. (2022) focuses on the detection of cognitive impairment and detecting whether elders suffer from mild cognitive impairment (MCI) or Alzheimer's disease based on the analysis of the audio recording of participants. The participants went through several neuropsychological tests in order to detect the above-mentioned disorders. The study uses audio waveform as the basis of its research and converts it into Mel-spectrogram. With the combination of Convolutional Neural Network(CNN) and Transformer, it achieves a high accuracy for the detection of both MCI and AD respectively. Although this study demonstrated promising results, the results are confined to a certain spectrum since it has only considered 120 participants.

The limitation of using a limited-size database has been taken care of by another study Nikolova et al. (2022) that uses a model based on linguistics obtained from verbal interviews of participants. The research study chooses close to 200 participants and uses distinct syntactic and lexical linguistic biomarkers in order to discriminate the control group patients from the group with probably Alzheimer's Disease. The study finds that the participants with probable AD have increased use of lexical components and low usage of syntactical components in their speech. Although this study uses components extracted from the speech in order to attain its objective it overlooks several other factors and medical conditions.

Both the above studies Lin et al. (2022) and Nikolova et al. (2022) have contributed to the area of speech recognition using deep learning techniques but the classification was limited to two categories so they did classify the severity of the cognitive impairments or Alzheimer's Disease(more than two classes).

2.2 Detection of Parkinson's Disease through Speech

Parkinson's Disease is known to be a neurodegenerative disorder that affects the motor neurons that cause stiffness in coordinating, walking, talking and tremors. Parkinson's Disease is very common and is often seen in elders. Machine learning has made recent advancements in the field of early detection of Parkinson's Disease(PD) to ensure prompt treatment. One of the recent studies, Mamun et al. (2022) makes use of machine-learning algorithms to detect Parkinson's Disease through vocal features. This paper uses distinct vocal features such as Pitch Period Entropy(PPE), Harmonics parameters, shimmers and Noise/Harmonic Ratio(NHR) to detect the disease. The authors implement ten different machine learning models to achieve their objective. These models include Logistic Regression, XGBoost, Support Vector Machine, Random Forest, LightGBM, Naive Baiyes, Bagging and AdaBoost and K-Nearest Neighbour.

The study Rao et al. (2023) uses an ANN Multi-layered perceptron to differentiate between patients with probable Parkinson's Disease and the control group. Both the studies [1] and [2] use machine-learning and deep-learning techniques that help in the early detection of Parkinson's Disease. Both these studies use accuracy as a decisive factor in the evaluation of their results. This can lead to partially correct results as when it comes to studies involving medical science, other parameters like specificity and sensitivity need to be considered.

2.3 Voice disorder detection

Several studies highlight voice disorders by analyzing and comparing various machine-learning models used for voice pathology detection. One such study (Mohammed et al.; 2020) uses Convolutional Neural Network (CNN) for voice pathology detection. The paper creates a comparative analysis of the speech recording samples of the vocalization of the vowel 'a'. The study uses a large dataset of around 2000 participants, out of which 687 are part of the control group of healthy individuals and 1356 participants suffer from different disorders. As part of the implementation process, the voice samples are transformed into framed signals. The authors use the Fourier Transform method to convert the framed signals to frequency-domain signals. A spectrogram is created by adding up the frequency-domain signals which in turn is analyzed for the pathology classification.

The study (Suparatpinyo and Soonthornphisaj; 2023) uses residual networks (ResNet) to diagnose of mental health issues such as depression. The study attempts to highlight the aspect of mental illness through speech features. The detection of mental illness from the audio is performed through feature extraction techniques like wavelet transform in which the generated spectrograms are studied for analysis. A study Markaki and Stylianou (2011) conducted a few years ago used the modulation features to detect and discriminate voice pathology. It explores the modulation spectrum for speech analysis. The study employs cross-validation tests on a database of sustained vowel recordings from control groups and pathological voices. The study uses the Support Vector Machine algorithm to fulfill its objective and achieves a classification accuracy of close to 95%.

Although the above-mentioned papers use advanced feature extraction techniques like spectrograms, the scope is restricted only to a single feature extraction technique i.e. Spectrogram analysis. This gap is bridged by the proposed thesis as it uses empirical mode decomposition and wavelet transform both as a feature extraction technique for disorder classification.

2.4 Detection of Aphasia and Dementia

Aphasia is known to be a condition where a person has difficulty speaking or understanding other people speaking. According to medical science, this usually happens when the part of the brain controlling speech is damaged. Aphasia is generally linked with stroke according to medical sciences. Often therapies are suggested as a way of treatment for this condition. One such study covering the severity levels of aphasia uses raw audio files and then converts them into mp3 format (Herath et al.; 2022). The files are then split into 20 chunks for feature extraction using MFCC. This paper uses MFCC as the feature extraction technique since it has a higher classification accuracy than other feature extraction methods. Different supervised machine-learning models like K-Nearest Neighbor, Decision Tree, Deep Neural Network and Random Forest are used for analysis and comparison, Out of these models, the Deep Neural Network model seems to have the highest values for accuracy, precision and F1 score. However the paper covers eight severity levels of aphasia but still restricts the samples only from low to moderate severity levels. Furthermore, this paper only focuses on a single speech impediment for its research.

The paper (Milana; 2023) uses an attention mechanism on speech data for dementia classification. The study analyzes the mel-spectrogram representation of the audio samples. As part of this paper, the attention mechanism is built using matrix multiplication making use of their context and hidden layers. The results of this study are measured in terms of accuracy. In addition to this, there have been a few studies conducted with speech transcripts in order to detect early dementia in patients. Nambiar et al. (2022) illustrates a comparative model of deep learning and natural language processing techniques like GloVe, Doc2Vec word embeddings and BiLSTM, LSTM, GRU, BERT, RoBERTa and ALBERT. The highest accuracy was generated as 0.812 using the BERT+BilSTM model. The above-mentioned studies (Herath et al.; 2022), Nambiar et al. (2022) and (Milana; 2023) use different techniques to detect speech impairments like aphasia and dementia respectively but they attempt to analyze only two classes. The gap is bridged by the author of this study where speech impediments like stuttering and dysarthria are covered for the detection of their type and severity.

2.5 LSTM-based techniques for speech recognition

Automated Speech Recognition (ASR) is one of the most demanding areas in Natural Language Processing(NLP). Deep-learning techniques such as the Long Short-Term Memory (LSTM) method has been found to be more reliable than other traditional methods for automatic speech recognition problems. One such study (Oruh et al.; 2022) highlights the issues with the traditional LSTM method and proposes Recurrent Neural Network (RNN) along with LSTM to detect automatic speech recognition using raw audio data as input. In this study, RNN is incorporated as a ”forget gate” to the memory block to allow the resetting of cell states at the beginning of the sub-sequences. As part of this study, along with RNN, some CNN-based LSTM model architecture was also used on the same dataset but the LSTM-RNN model outperformed all the other models and achieved an accuracy of 99.36%˙

A similar study (Bhaskar and Thasleema; 2023) uses LSTM deep-learning technique for visual speech recognition using with the help of facial expressions. The basis of this study is that hearing-impaired people are more expressive in nature while speaking. The study uses CNN-based LSTM model architecture for the visual speech recognition system. The dataset is collected from 2 people - 1 male and 1 female and the results are evaluated in terms of classification accuracy. The paper demonstrates that classification accuracy is far better for features extracted by GoogleNet model as compared to the other two models-ResNet and AlexNet models. The results show that the recognition rate for both the speaker-dependent and speaker-independent experiments proves that the analysis of the facial expressions in the participants plays a crucial role in visual speech recognition. Both the above-mentioned studies have used LSTM techniques for speech recognition but both have limited scope in terms of feature extraction. For reliable results, it is crucial for researchers to base their results on more than one feature extraction technique. One interesting study Gupta et al. (2020) uses bidirectional Long Short-Term Memory to detect the prolongation and repetition of words in stuttered speech through weighted MFCC(Mel Frequency Cepstral Coefficients). The main intention behind the use of WMFCC(Weighted MFCC) instead of the traditional MFCC is to extract the dynamic features from the audio signals for enhanced accuracy. The study achieves its objective in five stages: preprocessing of audio signals, labeling of speech signals, splitting the data into training and test sets, validation of the test sets, feature extraction and classification using model training. Although this study uses the LSTM algorithm which is considered to be highly suitable for speech recognition, the study focuses only on a single speech impediment i.e. stuttering. On the other hand, our study focuses on different speech impediments like stuttering and dysarthria.

2.6 Emotion Classification Using Speech:

The study Andayani et al. (2022) uses a hybrid LSTM Transformer model to recognize the emotions of the speech through audio files. Emotions play a vital role in human communications. As part of this study, the features are extracted with the help of Mel Frequency Cepstral Coefficient (MFCC) and then fed into the model for training and results. The study shows the model achieves better results as compared to the traditional models. However, the combination of the traditional LSTM model and the Transformer Encoder makes the model very complex and a bit unstable.

A recent study Caschera et al. (2022) uses the Markov model implementing a multimodal approach to classify emotions from speech. The multimodal data for the study is extracted from videos. This study involves extracting features like facial expressions, gestures and text from the video of the participants. Each multimodal data is mapped with emotion using a hidden Markov model. The trained model is evaluated on the samples against seven basic emotions. The study shows promising results and a good classification rate for emotions. However, often getting access rights to the video data to use the same in the study is difficult.

A study Ayadi and Lachiri (2022) similar to the proposed research uses MFCC(Mel Frequency Cepstral Coefficients) for feature extraction and classifies emotions based on speech and song attributes. This study uses the CNN-LSTM model for training purposes. The project uses MFCC in order to use the frames instead of the waveforms of the audio data. The results obtained from the model are tested against the following emotions- 'happy' 'calm', 'fear', 'angry' and 'sad'. The results obtained from this study do not look very promising since it has employed only a single feature extraction technique(MFCC).

2.7 Music Genre Classification:

Deep learning techniques have gained huge popularity in the field of music classification. One such study Srivastava et al. (2022) illustrates the classification of audio clips into well-defined music genres. The study uses CRNNs (Convolutional Recurrent Neural Networks) for this task. This study uses MFCC as the representation of the audio data. The study achieves high accuracy for both the implemented models: CNN-LSTM and CNN-GRU.

The majority of reviewed research studies are conducted in the area of speech recognition and early detection of diseases like Alzheimer's and Parkinson's using distinct speech features but very few papers explore the area of speech severity detection. There have been certain studies that cover speech disorders in general. Anthony et al. (2022) reviews the different speech disorders and processing of disordered speech. This study uses the changes in cognitive-linguistic functions to detect any speech disorder. The study Anthony et al. (2022) provides a detailed approach for detecting impediments like stuttering, aphasia and dysarthria. The main objective of this research study is to classify speech impairments based on the two datasets of stammering and dysarthria and then label the severity of these impediments using clustering. One of the distinct attributes of this paper is that it uses two complex feature extraction methods- Empirical Mode Decomposition and Wavelet Transform.

3 Methodology

The LSTM along with the combination of Recursive Neural Network(RNN) has gained wide popularity in the field of speech recognition as compared to the traditional pattern recognition methods. Yuan et al. (2023) demonstrates how the LSTM-RNN method has high perception ability and can store the data in a decentralized manner compared to a human brain. The methodology of this project comprises all the steps from the data collection to the model building section. This section provides a systematic approach to demonstrate how the LSTM-RNN deep-learning technique is used to detect the type and severity of speech impairment through raw audio data.
As part of this project, we will be using the KDD technique. KDD stands for Knowledge Discovery in Databases. This technique involves various steps like selecting the data, cleaning and pre-processing the data and then interpreting outcomes from the observed results.
An overview of the methodology of the project is illustrated in the below figure.

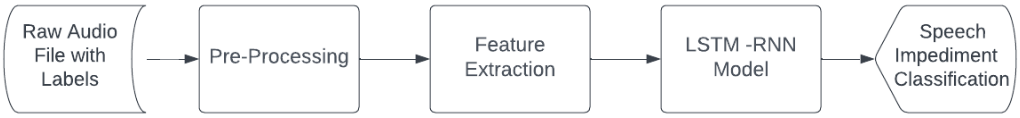

Figure 1: Overview of Methodology

The methodology for this project involves the following steps:
1.**Data Collection:** The dataset for this project consists of 3 sets of audio files. One set contains the files of the audio recordings of the participants with stuttering. The second set comprises the files of the participants with dysarthria and the third set comprises of the control group or the people without any speech impediment.

2.**Data Preprocessing:** The data preprocessing is an essential step in the detection of speech impairment using LSTM-RNN model architecture. The data is in the form of raw audio files which need to be cleaned and preprocessed so that it can be fed into the model. It is very important for the data to be preprocessed properly so that the results generated are effective and reliable. The data preprocessing step consists of the following sections:

2.1 **Loading the data:** The raw audio data files are stored in the directory and are loaded from the same. The entire dataset is divided into three categories- 'Stammer', 'Normal' and 'Dysarthria' based on their subdirectories.

2.2 **Noise Reduction:** In order to achieve effective results, the raw audio data needs to be made suitable for the model architecture. We use the 'split' method for splitting audio files on silence/muteness. The 'split' function can also be used to distinguish the mute and non-mute sections of the audio files.

2.3 **Savitzky-Golay Smoothing Technique:** As part of this project, another noise-reducing method is applied which is the Savitzky-Golay smoothing filter. These filters are used to "smooth out" audio signals with large frequencies. This method is often used in data mining projects since it has proved to be quite effective in preserving the high frequency of the signals.

The mathematical details of the Savitzky-Golay smoothing technique involve piece-by-piece fitting of a polynomial function to the audio signal. This fitting is achieved by the Least Squares Estimate :

$$y = Xb. \tag{1}$$

In the above equation (1), 'X' is the matrix, and 'y' is the vector.
The standard Least Square Estimate is given by the below equation:

$$b = (XTX) - 1XTy. \tag{2}$$

The estimated values used for the smoothing technique are:

$$Xb = X(XTX) - 1XTy = Hy. \tag{3}$$

In the above equation (3), the product $H = X(XTX) - 1XTy$ is known as Hat Matrix and its value is the same at any vector for a given polynomial.

2.4 **Normalization:** In order to ensure that all the amplitude values of the audio data are within the standard range, we use the 'l.normalize()' function.

2.5 **Waveform Padding:** The dataset used for this project consists of audio files that are of different lengths. For deep-learning models, it is very important for the entire data to be of a standard length. In order to avoid any discrepancies in the results, the 'pad_waveform' function is used either to pad or truncate the audio waveforms. The padding of the waveform means to standardize the length of the waveform if the length of the waveform is shorter than the standard length. On the other hand, truncating the waveform refers to reducing the length of the waveform if the length of the given waveform is longer than the decided standard length.
This step forms an important step in the pre-processing as it ensures that the input data for the LSTM-RNN model is consistent.

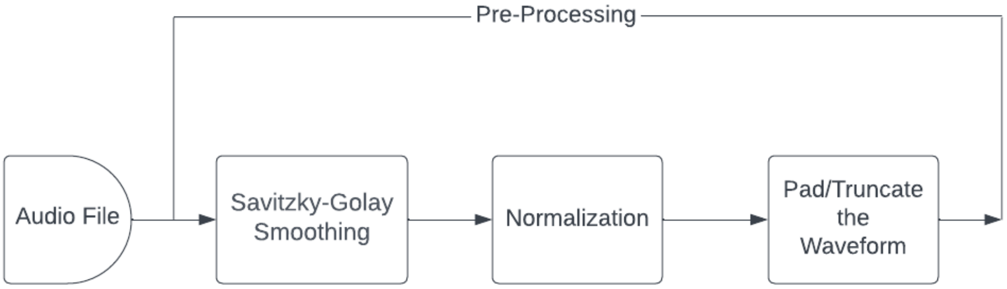

Figure 2: Pre-Processing Steps

3 **Feature Extraction:** Feature extraction is one of the most important aspects of machine learning when dealing with images or audio files. In order to get more effective computational results, audio signal processing has gained huge importance in the field of data analytics and data mining. Audio feature extraction refers to the processing or manipulation of the audio signals to remove unwanted noise and balance time-frequency ranges by converting digital and analog signals. Figure 4 represents the Mel-spectrogram for a smoothed waveform.
In this project, as part of feature extraction, we use two powerful techniques to decompose the audio signals into different frequency components. Both the processes- Empirical Mode Decomposition and Wavelet Transform are used to derive patterns and information from the signals which are then used by the LSTM-RNN model to predict speech impairments type. Figure 3 represents the overview of Feature Extraction.

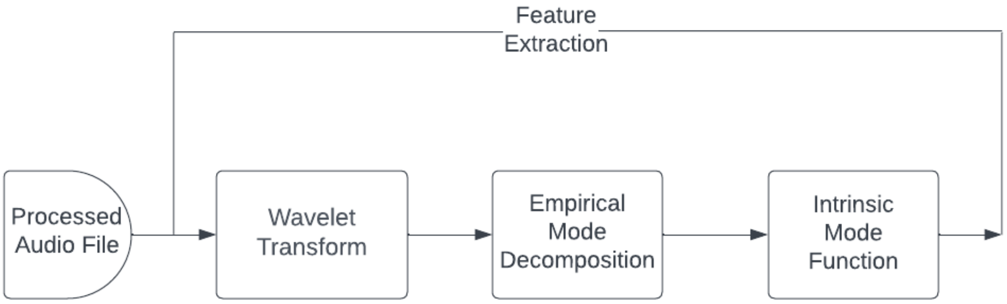

Figure 3: Feature Extraction

3.1 **Empirical Mode Decomposition (EMD):** Empirical Mode Decomposition is an algorithm that is used to decompose an audio signal. An audio signal refers to the time-series data of a specific audio file. EMD decomposes an audio signal into different Intrinsic Mode Functions (IMFs). The IMFs are based on two rules- the first being the difference between the maxima and the minima is at most '1'. Secondly, the mean of the IMF wave is always zero.

3.2 **Wavelet Transform:** Wavelet transform is one of the recent methods to extract the information from an audio signal for analysis purposes. It is widely used in the field of deep learning for audio classification purposes. The wavelet transform is used to measure the frequency of the audio signal against time. In this project, the Daubechies wavelet ('db1') is used for a level of 5 for the decomposition for analyzing the audio signals.
As part of this implementation, we use the 'preprocess_waveform' for feature extraction purposes.

4 **AudioDataset Class**: In order to access and analyze the audio data of different impediments, we use the 'AudioDataset' class which comprises the links to different paths of the data and their respective labels for the impediment classification purposes.

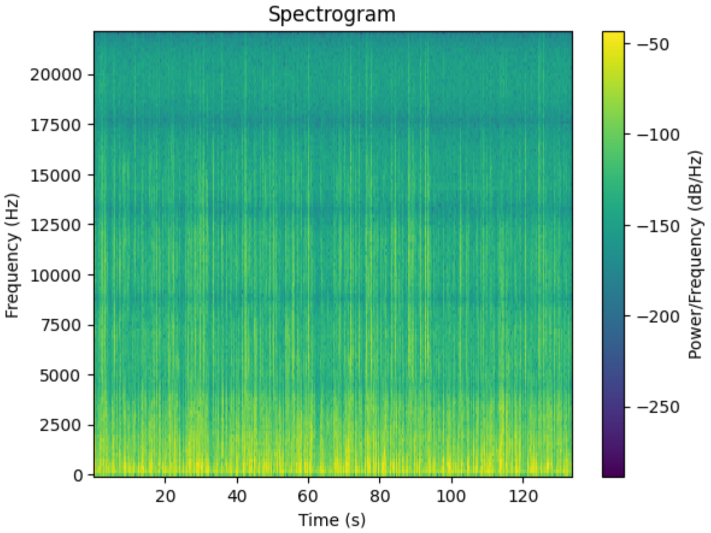

Figure 4: Spectrogram of a smoothed waveform

5. **Data Split into Test and Train Set**: In order to split the entire dataset into testing and training sections, we use the 'train_test_split' method to maintain a balance of all three classes in both sets. We use the ratio of 80:20 for training and testing respectively.

6. **Defining the LSTM-RNN Model:** For defining the LSTM-RNN model, we use a class named 'LSTMNetwork'. This class consists of three layers, the first layer consists of the temporal layer to manage the temporal dependencies. While the other two fully connected layers for the mapping of the data to the three classes. Figure 5 represents the LSTM-RNN model architecture.

7. **Model Training:** In order to provide effective results based on the evidence, we train two models. One model uses both empirical mode decomposition and wavelet transform as the feature extraction techniques whereas the second model uses only the wavelet transform. The main intention behind using these two models is to provide a comparative analysis of speech classification using both methods.

8. **Testing the Model**: In order to evaluate the performance of both models, we

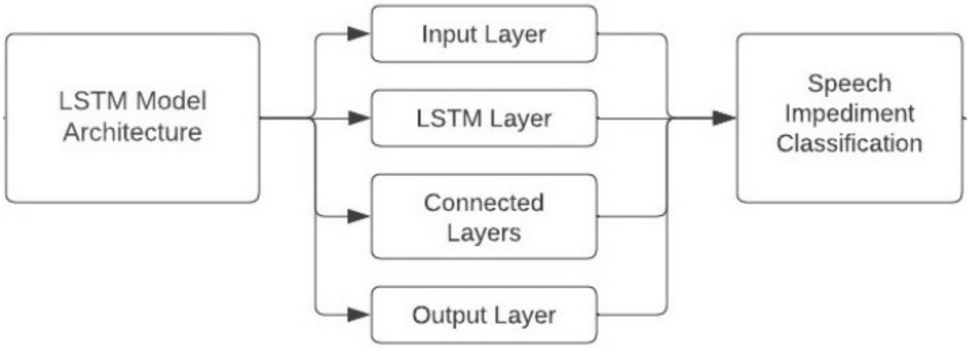

Figure 5: LSTM-RNN Model Architecture

use the 'test' function on the test set.

9. **Evaluate the Model:** For evaluating the models, we use accuracy as the main parameter. The accuracy of the model is measured by the comparison of the predicted labels with the actual labels.

The above steps involved in the methodology provide a systematic approach for the classification of the different speech impediments. These steps are in turn analyzed and used for measuring the severity of these impediments using the clustering technique. In addition to the detection of speech impediment type using the LSTM-RNN model, this project also attempts to estimate the severity levels of the speech impediments using the K-Means clustering algorithm.
The K-Means algorithm is known to be an unsupervised machine learning algorithm that attempts to partition the data into predefined subgroups or clusters where each data point belongs to only one sub-group. In this case, the subgroups are denoted by the different severities of the impediments.

As part of this project we use 3 clusters(k=3) where each cluster denotes the type of severity-'mild', 'moderate' and 'severe'. The severity estimate is calculated on the basis of the MFCC (Mel Frequency Cepstral Coefficients) features. One of the important points to consider here is that the K-means algorithm used in this project can be referred to as an estimation of the severity levels of the speech impediments based solely on the MFCC features.

4 Design Specification

The main objective of this project is to illustrate the importance of deep learning techniques in the area of speech recognition. The project demonstrates how the LSTM-RNN model is used to classify speech impediments and determine their severity.

One of the distinctive features of this paper is the feature extraction techniques used for the preprocessing and analysis of the audio datasets. In the above project, we use a standard LSTM-based design architecture for the classification of the sequenced data. In this case, this sequenced data is in the form of raw audio files.

The feature extraction process breaks down the audio waveform into several features from each of the 2 techniques- Empirical Mode Decomposition and Wavelet Transform. The following feature is extracted from the Wavelet Transform:

Wavelet Coefficients: The wavelet transform uses wavelets to capture the details of the waveform at different frequency bands in the 'coeffs'. These coefficients are reconstructed and then in turn consumed by the Empirical Mode Decomposition process.

Empirical Mode Decomposition: The Empirical Mode Decomposition process breaks down the output waveform from the Wavelet Transform into Intrinsic Mode Functions(IMFs). The IMFs represent the oscillatory modes in the processed audio signals. As part of this project, we use only the first IMF as it uses most of the features.

Model Architecture: Post the feature extraction and the feature selection processes, the processed audio signal is fed to the LSTM-RNN model. The LSTM-RNN model is divided into the following sections:

LSTM Layer: The LSTM layer comprises the most important layer of the model architecture. The LSTM layer has the following parts:

1. **num_mfcc:** This parameter refers to the number of Mel Frequency Cepstral Coefficients (MFCCs).

2. **Number of Neurons:** This parameter shows that the model has 64 neurons.

3. **dropout:** This parameter refers to the dropout layer to prevent model overfitting.

4. **Connected Layers:** The model architecture consists of one LSTM layer and two fully connected layers. The first layer has 32 neurons and takes the output from LSTM. The second layer has 32 neurons and maps these neurons to the 3 classifying classes- stammering, normal speech and dysarthria.

5 Implementation

The code for the project provides detailed step-by-step procedures for the implementation of the code. The following are the different steps involved in the project:

1. **Import and Installations of the packages:** In order to achieve the objective of the project, we installed several libraries like 'PyEMD' and 'EMD-signal'. In addition to this, several other libraries like 'librosa', 'numpy', 'pywt' ,'matplotlib' and 'sklearn' have been imported.

2. **Mel Spectrograms:** As part of the feature extraction the spectrograms are plotted for the audio waveforms for all three classes - Stammering, Normal Speech and Dysarthria. These spectrograms denote the distribution of the frequency of the audio waveform with respect to time. Figures 6, 7, and 8 define the mel-spectrograms of the sample files for each of the labels.

3. **Model Implementation:** In order to detect the speech impairment type and severity through the LSTM-RNN model, the pre-processed data is fed into the model for training purposes. Both models are trained for a predefined number of epochs. The model using both empirical mode decomposition and wavelet transform as the feature extraction techniques is run on 50 epochs. The 'train_single_epoch' function is used by the model to perform one epoch run and then the 'train' function in turn runs this for 50 epochs.
The second model which uses only the Wavelet Transform as the feature extraction technique is run for 10 epochs. In order to ensure the correctness and robustness of the model, the application of Adam Optimizer and Cross-Entropy Loss function is introduced.
The combination of the Adam optimizer and Cross-Entropy Loss function has gained wide importance in the field of machine learning. This combination is generally used in the area of deep learning dealing with classification tasks. Since the aim of this project is to classify the speech impediment and detect its severity, the combination of the Adam optimizer and Cross-Entropy function is chosen.

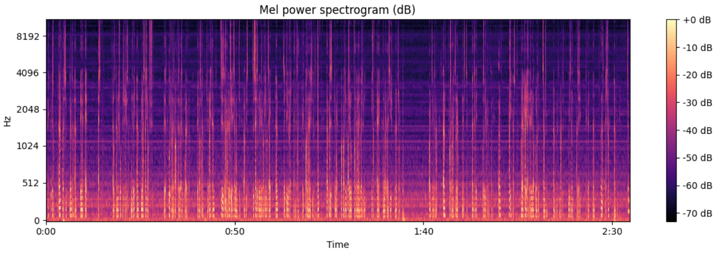

Figure 6: Spectrogram of a stuttered waveform

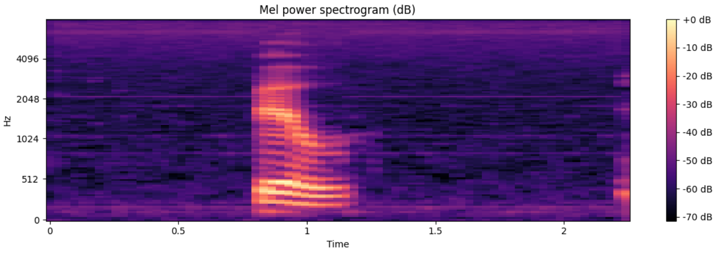

Figure 7: Spectrogram of a normal waveform

6 Evaluation

For an in-depth analysis of the model performance, we divide the evaluation section into 2 parts. First we calculate the evaluation metrics for the model detecting the type of speech impediment using LSTN-RNN and then we evaluate the severity of those speech impediments.

Type of Speech Impediment: In order to measure the performance of the model

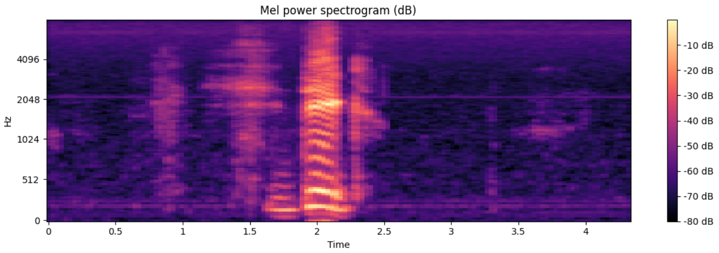

Figure 8: Spectrogram of a dysarthria waveform

in detecting the type of speech impediment using LSTM-RNN model, we use different evaluation metrics like accuracy, precision, sensitivity, specificity and F1 score.

1. **Accuracy:** As part of this model, the accuracy parameter is denoted by the measure of the classified instances out of the total instances. On the basis of our model's performance, we calculate an accuracy of 60%.

2. **Recall:** This parameter measures the count of the correctly predicted false positives by the model. For this model, we achieve a recall of 0.3333.

3. **Precision:** This parameter provides us with the fact that how many predicted false positives are positives in real. For the given model, we achieve a precision of 0.200.

4. **F1 Score:** This evaluation parameter is used when the classification data is involved. This is referred to as the harmonic mean of precision and recall. For the above-described model, we get an F1 score of 0.2500.
In addition to these parameters we also calculate the specificity that comes out to be 0.66.

Speech Impediment Severity:

For the representation of the clustering algorithm, in order to detect the severity type, we will be using a dimensionality-reducing technique to represent the high-dimension data in a lower-dimension space. In the field of machine learning, there are two widely popular techniques for dimension reduction and they are -Principal Component Analysis(PCA) and t-Distributed Stochastic Neighbour Embedding(t-SNE). In this case, we use the t-SNE technique which is more suitable for audio data. The t-SNE technique maps the MFCC features to the two-dimensional space.

Figure 9 denotes the clustering algorithm for the estimation of the severity of the impediment, each color represents a cluster and these colors may vary depending on the rendering.

Figure 9 depicts three different classes, where each class represents 'mild, 'moderate' and 'severe' severity levels. The following is the distribution of the classes in terms of severity levels-

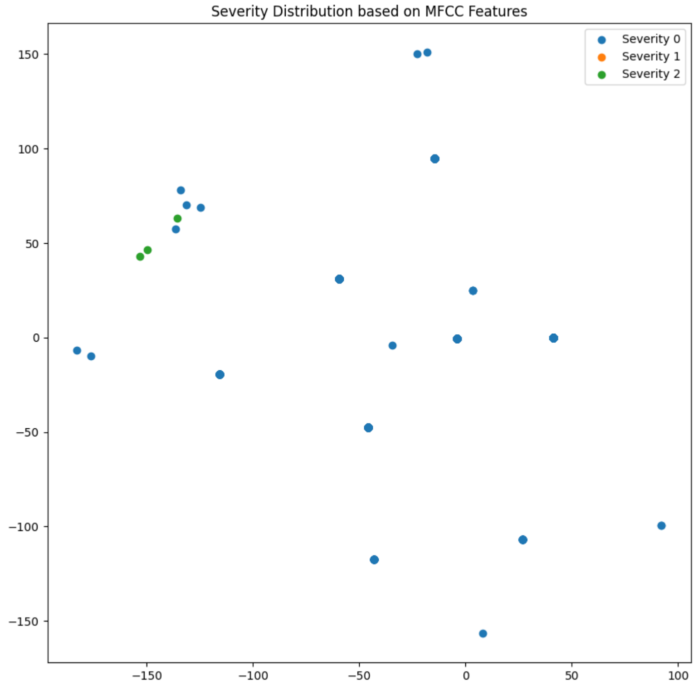

Figure 9: Cluster Representation for Severity Type of Speech Impediment

Class 0 represents 'Mild Severity'
Class 1 represents 'Moderate Severity'
Class 2 represents 'Severe Severity'

6.1 Experiment / Case Study 1

6.1.1 Model with EMD and Wavelet Transform:

The LSTM-RNN model is first trained using both advanced feature extraction techniques - Empirical Mode Decomposition and Wavelet Transform. Both these methods help us to decompose the audio signals in a suitable format for the LSTM model architecture to predict the outcomes. The model generates an accuracy of 60% for the data pre-processed using both the above-mentioned feature extraction techniques.
In this process, first the audio signals are interpreted using the time-frequency information (Wavelet Transform) and the EMD is used to break down the signals into Intrinsic Mode Functions.

6.2 Experiment / Case Study 2

6.2.1 Model with Wavelet Transform:

In this experiment, only the Wavelet Transform technique as part of feature extraction is used. In this process, the audio signal is decomposed the audio signals into coefficients that contain information about the signals at different frequencies. The reconstructed waveform created from these coefficients is used as an input for the neural network model. This method achieves an accuracy of 17.5% which is marginally lower than the first experiment.

6.3 Discussion

As part of this project, we use the LSTM-RNN method for the detection of speech impairment using audio files of patients suffering from disorders like stuttering, dysarthria and normal speech. We use two methods for a detailed analysis of speech impediment classification. The first method uses wavelet transform and EMD and the second method uses only the former technique. The results show that the experiment involving two feature extraction techniques has higher values of evaluation metrics. The first method has an accuracy of 60% and the second method has an accuracy of 17.5%.

In order to estimate the severity levels, we use the K-Means clustering unsupervised machine learning technique. Although this algorithm generates clusters depicting different severity levels, the method solely depends on the patterns in the MFCC features. The actual and real-time severity interpretation of the speech impediment may involve a domain expert or a speech therapist.

7 Conclusion and Future Work

In conclusion, the code provides a comprehensive approach to detecting speech impediments from audio files using an LSTM network and advanced feature extraction techniques. The results, particularly the evaluation metrics, would be key to understanding its real-world applicability. While the foundation is strong, there's always room for optimization and enhancement to cater to specific use cases or to achieve higher precision. Although this study does a good job of achieving what it aimed for, there are still quite a few future considerations that can be considered for further research.

1. **Different Deep-Learning Technique:** In this project, we have used a Recursive Neural Network to achieve our objective of determining the speech impediment type. However, there have been recent studies that have shown intriguing results for speech recognition using Convolutional Neural Networks.

2. **Involving More Labels:** As part of this project, we have used three labels for classification in the model training- stuttering, dysarthria and normal speech. For future scope, we can include other speech disorder data like mutism, aphasia, etc.

3. **Inclusion of Other Evaluation Metrics:** As mentioned earlier, incorporating metrics beyond just accuracy can provide a more holistic view of model performance. We can include metrics like sensitivity and specificity for such cases involving medical sciences.

References

Andayani, F., Theng, L. B., Tsun, M. T. and Chua, C. (2022). Hybrid lstm-transformer model for emotion recognition from speech audio files, *IEEE Access* **10**: 36018–36027.

Anthony, A. A., Patil, C. M. and Basavaiah, J. (2022). A review on speech disorders and processing of disordered speech, *Wireless Personal Communications* **126**(2): 1621–1631.

Ayadi, S. and Lachiri, Z. (2022). A combined cnn-lstm network for audio emotion recognition using speech and song attributs, *2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, IEEE, pp. 1–6.

Bhaskar, S. and Thasleema, T. (2023). Lstm model for visual speech recognition through facial expressions, *Multimedia Tools and Applications* **82**(4): 5455–5472.

Caschera, M. C., Grifoni, P. and Ferri, F. (2022). Emotion classification from speech and text in videos using a multimodal approach, *Multimodal Technologies and Interaction* **6**(4): 28.

Gupta, S., Shukla, R. S., Shukla, R. K. and Verma, R. (2020). Deep learning bidirectional lstm based detection of prolongation and repetition in stuttered speech using weighted mfcc, *International Journal of Advanced Computer Science and Applications* **11**(9).

Herath, H. M. D. P. M., Weraniyagoda, W. A. S. A., Rajapaksha, R. T. M., Wijesekara, P. A. D. S. N., Sudheera, K. L. K. and Chong, P. H. J. (2022). Automatic assessment of aphasic speech sensed by audio sensors for classification into aphasia severity levels to recommend speech therapies, *Sensors* **22**(18): 6966.

Lin, S.-Y., Chang, H.-L., Hwang, J.-J., Wai, T., Chang, Y.-L. and Fu, L.-C. (2022). Automatic audio-based screening system for alzheimer's disease detection, *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE.

Mamun, M., Mahmud, M. I., Hossain, M. I., Islam, A. M., Ahammed, M. S. and Uddin, M. M. (2022). Vocal feature guided detection of parkinson's disease using machine learning algorithms, *2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, IEEE, pp. 0566–0572.

Markaki, M. and Stylianou, Y. (2011). Voice pathology detection and discrimination based on modulation spectral features, *IEEE Transactions on audio, speech, and language processing* **19**(7): 1938–1948.

Milana, S. (2023). Dementia classification using attention mechanism on audio data, *2023 IEEE 21st World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, IEEE, pp. 000103–000108.

Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., Khanapi Abd Ghani, M., Maashi, M. S., Garcia-Zapirain, B., Oleagordia, I., Alhakami, H. and Al-Dhief, F. T. (2020). Voice pathology detection and classification using convolutional neural network model, *Applied Sciences* **10**(11): 3723.

Nambiar, A. S., Likhita, K., Pujya, K. V. S. S., Gupta, D., Vekkot, S. and Lalitha, S. (2022). Comparative study of deep classifiers for early dementia detection using speech transcripts, *2022 IEEE 19th India Council International Conference (INDICON)*, pp. 1–6.

Nikolova, M., Nanovic, Z. and Gievska, S. (2022). Approach for screening and early diagnosis of alzheimer's disease through detection of linguistic deficiencies and other biomarkers, *ICT Innovations*.

Oruh, J., Viriri, S. and Adegun, A. (2022). Long short-term memory recurrent neural network for automatic speech recognition, *IEEE Access* **10**: 30069–30079.

Rao, S. S. et al. (2023). Eadn: Enhanced auto encoder decoder network ensembled with boosting technique for feature selection of parkinson's disease detection.

Srivastava, N., Ruhil, S. and Kaushal, G. (2022). Music genre classification using convolutional recurrent neural networks, *2022 IEEE 6th Conference on Information and Communication Technology (CICT)*, IEEE, pp. 1–5.

Suparatpinyo, S. and Soonthornphisaj, N. (2023). Smart voice recognition based on deep learning for depression diagnosis, *Artificial Life and Robotics* pp. 1–11.

Yuan, Q., Dai, Y. and Li, G. (2023). Exploration of english speech translation recognition based on the lstm rnn algorithm, *Neural Computing and Applications* pp. 1–10.