INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Check for updates

OPEN ACCESS

Received: 24.12.2020

Accepted: 30.01.2021

Published: 16.02.2021

Citation: Gupta S, Shukla RS, Shukla RK (2021) Weighted Mel frequency cepstral coefficient based feature extraction for automatic assessment of stuttered speech using Bi-directional LSTM. Indian Journal of Science and Technology 14(5): 457-472. https://doi.org/10.17485/IJST/v14i5.2276

* **Corresponding author**.

gsakshi18@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Weighted Mel frequency cepstral coefficient based feature extraction for automatic assessment of stuttered speech using Bi-directional LSTM

Sakshi Gupta[1]*, **Ravi S Shukla**[2], **Rajesh K Shukla**[3]

1 Department of Computer Science and Engineering, Invertis University, Bareilly, Uttar Pradesh, India
2 Department of Computer Science, Saudi Electronic University, Tabuk, Saudi Arabia
3 Department of Engineering and Technology, Invertis University, Bareilly, Uttar Pradesh, India

Abstract

Objective: To propose a system for automatic assessment of stuttered speech to help the Speech Language Pathologists during their treatment of a person who stutters. **Methods**: A novel technique is proposed for automatic assessment of stuttered speech, composed of feature extraction based on Weighted Mel Frequency Cepstral Coefficient and classification using Bi-directional Long-Short Term Memory neural network. It mainly focuses on detecting prolongation and syllable, word, and phrase repetition in stuttered events. **Findings**: This study has discussed and performed a comparative analysis of WMFCC feature extraction method with different extensions of widely used MFCC, namely, Delta, and Delta-Delta cepstrum. The comparison of speech parameterization techniques is carried out based on the effect of different frame lengths, percentage of window overlapping, and pre-emphasis filter alpha value. The experimental investigation elucidated that WMFCC outperforms the other feature extraction methods and provides an average recognition accuracy of 96.67%. 14-dimensional WMFCC achieves a low computational overhead compared to conventional 42-dimensional MFCC, including Delta and Delta-delta cepstrum. **Application**: The integration of Weighted MFCC based speech feature extraction and deep learning Bi-LSTM based classification techniques proposed in this study are more efficient for introducing an optimal model to automatically classify the stuttered events such as prolongation and repetition.

Keywords: Stuttering; MFCC; Delta MFCC; WMFCC; BiLSTM

1 Introduction

For communication between human beings, speech is the most habitually and widely used verbal means to precise feelings, ideas, and thoughts. Not all human beings are blessed with normal means of speech. The power of speech in the sharing of information during interaction depends on fluency[1]. If continuity between semantic units, rhythm, speed, and energy contributed to flow is natural, speech is fluent. Dysfluency is characterized as any form of fluency disruption. The complex form of dysfluency is stuttering. In stuttering, due to pauses and blocks, there is a disruption in continuity and rhythm, the rate is much slower, and efforts are greater than normal.

There may be three kinds of disorders in people who stutter (PWS): repetition of syllable, word or phrase, prolongation, and silent blocks at starting a vocalization or expression or within the middle of a word. Stuttering influences individuals of all ages, cultures, and races, irrespective of their intelligence and financial status[2]. Many research pieces have stated that stuttering affects approximately 1% of the world population and is more common in males than females[3]. Therefore, this area is mainly a knowledge base field of analysis for distinctive domains like speech pathology, physiology, psychology, acoustics, and signal analysis.

Speech-Language Pathologists (SLP) diagnose the person who stutters and assess the fluency to determine the stutterer's response during the treatment phase. SLPs were previously used to determine the severity of stuttering manually through their experience. They counted and divided the frequency of stuttered events with total spoken words. Such sorts of stuttering assessments are arbitrary, incoherent, lengthy, and error-prone. Therefore, SLPs have paid considerable attention to objective assessment methods to identify stuttered events over the past few decades[4].

1.1 Literature Survey

The survery shows a detailed comparative analysis of various feature extraction and classification techniques based on the dataset used, type of disfluency, and accuracy[5–22]. The previous work published illustrates the significance of feature extraction and classification techniques in identifying stuttered events.

Table 1. Comprehensive analysis of various research activities on stuttering detection, describing the features used, classifier employed, number of subjects, type of classification and experimental results

Year	Feature Used	Classifier Used	Dataset Used	Type of Classification	Result
2009[5]	MFCC	SVM	12 training and 3 testing samples of 15 adults who stutter	Syllable Repetition	94.35%
2010[6]	LPC	HMM	5, 10, 15, 20 samples per command and 40-50 observation symbols of HMM	-	5 samples-93.75%, 10-98.75%, 15- 100% and 20-97.5%
2010[7]	MFCC	KNN and LDA	10 samples of 8 males and 2 females (11 to 20 years) from UCLASS	Repetition and Prolongation	90%
2010[8]	LPCC	KNN and LDA	10 samples of 8 males and 2 females (11 to 20 years) from UCLASS	Repetition and Prolongation	88.05%
2011[9]	12, 13, 26 and 39 Dimensional MFCC	DTW	8 training and 2 testing samples	Repetition	12 D- 80.69%, 13 D-68.4%, 26 D- 84.01%, 39 D- 84.58%,
2012[10]	MFCC and LPCC	KNN and LDA	UCLASS database	Repetition and Prolongation	MFCC- 92.55%, LPCC- 94.51%
2012[11]	Spectral Entropy using Bark, Mel and Erb Scale	SVM	UCLASS database	Repetition and Prolongation	Average accuracy-96%. Beat result of 96.84% in Erb scale
2013[12]	MFCC, PLP, and LPC	KNN, LDA, and SVM	UCLASS database	Repetition and Prolongation	Best average classification accuracy is given by SVM using the WLPCC, PLP, and MFCC features- 95%
2013[13]	SOM	Hierarchal ANN, MLP	153 recordings of 19 PWS	Blocks, syllable repetition and syllable initial prolongation	Blocks- 96% Syllable Repetition- 84% and Prolongation-99%
2014[14]	MFCC	SVM	UCLASS database	Repetition and Prolongation	95.6%

Continued on next page

Table 1 continued

2015[15]	MFCC	KNN	80 speech samples for training and 20 for testing	Repetition with 0db to 10db babble noise	60-95% depending on the sound used
2016[16]	MFCC, Formant, Pitch, ZCR, and Energy	ANN	78 recordings of 4 PWS (25-40 years)	Repetition and Prolongation	88.29%
2016[17]	MFCC, Formant and Shimmer	DTW	50 repetition events	Repetition	94%
2016[18]	MACV	Thresholding	5 Stuttering person speech samples from UCLASS database	Repetition and Prolongation	73.29%
2016[19]	MFCC and PLP	Cross-correlation, Euclidean distance using Morphological Image Processing	UCLASS database	Prolongation, word repetition, and phrase repetition	Prolongation- 99.84%, Word repetition- 98.07% and Phrase repetition- 99.87%
2017[20]	MFCC	I-Vector	1380 segments of 18 PWS from UCLASS. 80% used for training and 20% for testing	Repetition, Prolongation, and Repetition-Prolongation	Normal- 52.43%, Repetition- 69.56%, Prolongation- 40%, Rep-Pro- 50%
2020[21]	MFCC	Gated Recurrent CNN	UCLASS database	Prolongation and Repetition	Prolongation- 95% Repetition- 92%
2020[22]	MFCC	LSTM	UCLASS database	Prolongation, Blocks, and Repetition	4% and 6% higher than ANN and SVM

This paper focuses on the implementation and performance analysis of the feature extraction technique used in the proposed methodology. A wide variety of speech parameterization techniques are available for the recognition process, such as Perceptual Linear Prediction (PLP), Linear Prediction Coding (LPC), Linear Predictive Cepstral Coefficient (LPCC), and Mel Frequency Cepstral Coefficient (MFCC). In[12] and[19], the authors extracted PLP features to analyze stuttered speech samples. The PLP feature vectors show the dependency while maintaining overall spectral balance on formant amplitudes and sensitive to noise and communication channel[23]. In[24], the writers introduced a stuttered speech recognition system based on LPC features. LPC works on assuming the static nature of speech, therefore, ineffective in representing and analyzing speech accurately[25]. In[8] and[10], the authors analyzed LPCC features' performance to assess stuttered speech. LPCC delivers poor performance in high quantization noise and uses a linear scale that is not adequate for speech processing[23].

From Table 1, it can be observed that MFCC is a highly employed feature extraction technique. However, these features involve only static information of speech signals. Based on the above considerations, this paper introduces a more efficient extension of MFCC, known as Weighted MFCC (WMFCC) for feature extraction of stuttered speech samples. WMFCC includes the speech samples' dynamic information, which increases the detection accuracy of stuttered events; and reduces the computational overhead to the classification stage.

The proposed work has introduced a low dimensional and dynamic feature extraction method WMFCC, and deep-learning classification technique Bi-directional Long-Short Term Memory (Bi-LSTM) for the automatic evaluation and diagnosis of four forms of disfluency prolongation and syllable, word, and phrase repetition. The efficiency of WMFCC is determined by comparing performance of four feature extraction methods, MFCC, Delta, and Delta-Delta cepstrum, and WMFCC based on the accuracy of stuttered events classification. In[26], the authors have discussed the implementation and analysis of the classification technique employed in this study.

The paper is structured according to the following. Section 2 elaborates on the framework for the system proposed. Experimental results and a comparative analysis of the feature extraction techniques are performed in Section 3. Section 4 provides a conclusion.

2 Methodology

The proposed method for disfluency detection (Figure 1) is split into five phases: speech signal pre-processing, segmentation, and labeling of the disfluent speech signal, splitting the labeled samples into training, validation, and test sets, feature extraction, and classification. The study has conducted a comparative analysis of extensions of MFCC feature extraction techniques, namely Delta MFCC, Delta-delta MFCC, and Weighted MFCC. The University College London Archive of Stuttered Speech (UCLASS) database is utilized for analysis[27]. The Bi-LSTM classifier evaluates the efficacy of the feature extraction techniques in the classification of prolongation and repetition dysfluencies.

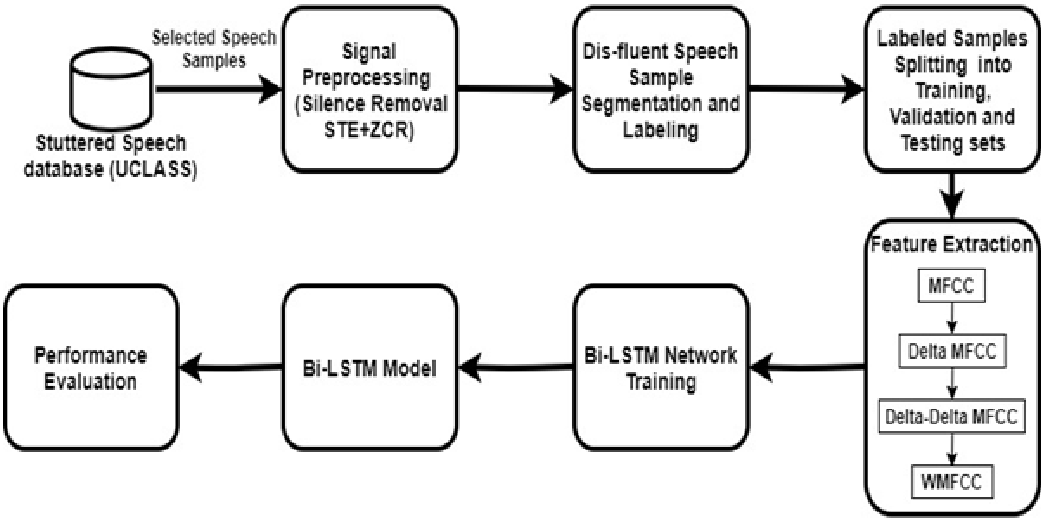

Selected Speech Samples
Stuttered Speech database (UCLASS)
Signal Preprocessing (Silence Removal STE+ZCR)
Dis-fluent Speech Sample Segmentation and Labeling
Labeled Samples Splitting into Training, Validation and Testing sets
Feature Extraction
MFCC
Delta MFCC
Delta-Delta MFCC
WMFCC
Bi-LSTM Network Training
Bi-LSTM Model
Performance Evaluation

Fig 1. Proposed system

2.1 Speech Signal Pre-processing

A signal is pre-processed by removing the silence regions[28]. There is no excitation in the vocal tract during the silence region, hence no speech production. Thus, pre-processing reduces the amount of processing and enhances the system's overall efficiency and accuracy. In this study, the integration of two widely known techniques, Short Time Energy (STE) and Zeros Crossing Rate (ZCR) (Figure 2), are applied[29]. It is a quick and straightforward approach and provides a better outcome of voiced/unvoiced/silent speech classification. (Figure 3).

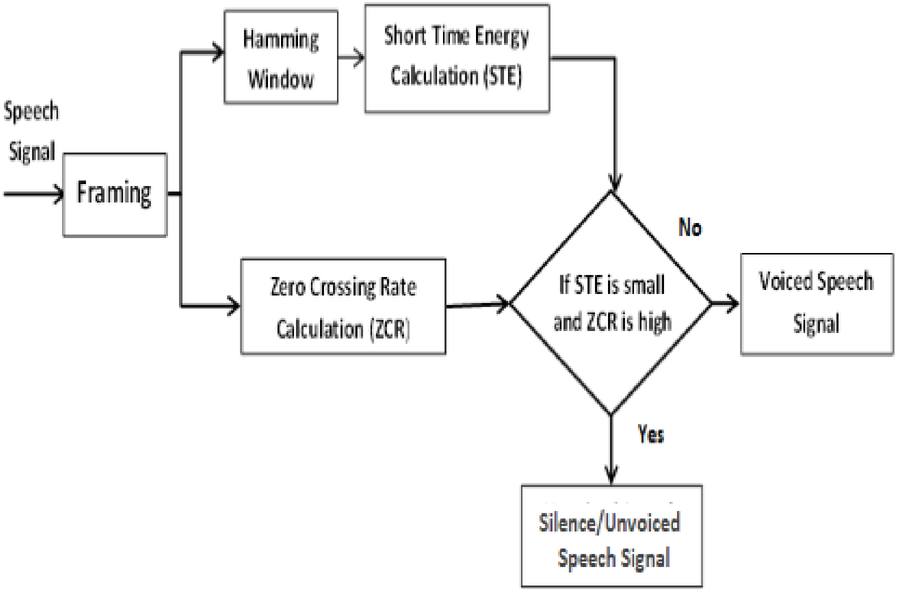

Speech Signal
Framing
Hamming Window
Short Time Energy Calculation (STE)
Zero Crossing Rate Calculation (ZCR)
If STE is small and ZCR is high
No
Voiced Speech Signal
Yes
Silence/Unvoiced Speech Signal

Fig 2. Speechpre-processing by silence removal

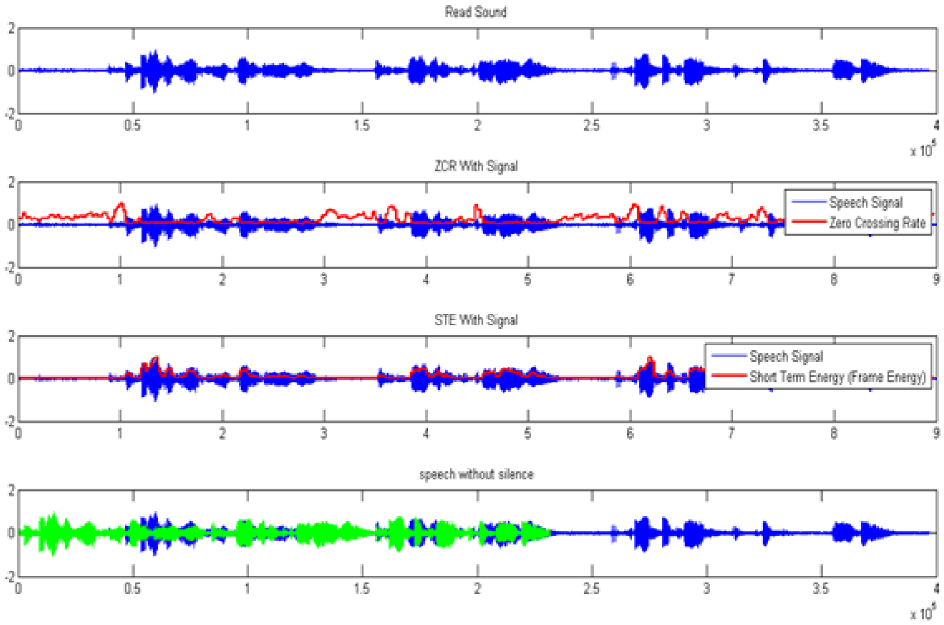


Fig 3. Silence removal using STE-ZCR method

2.2 Disfluent speech sample segmentation and labeling

The disfluent speech signals are obtained from the University College London Archive of Stuttered Speech (UCLASS)[(27)]. The dataset used in this study refers to 20 samples of stuttered speech of UCLASS Version 1 for experimentation. It comprises two female speakers and 18 male speakers aged 7years 8 months to 17 years 9 months. The purpose of the selection of speech signals is to cover a broad range of stuttering rates and ages. The samples available with transcriptions are only included in the dataset.

This paper investigates only four forms of disfluencies, prolongation, syllable, word, and phrase repetition. They are easily detectable in monosyllabic words. After pre-processing the selected speech samples, disfluent speech samples were identified and segmented manually by listening to the pre-processed signals. The segmented samples were labeled as five classes: Fluent, Prolongation, Syllable Repetition, Word Repetition, and Phrase Repetition (Figure 4).

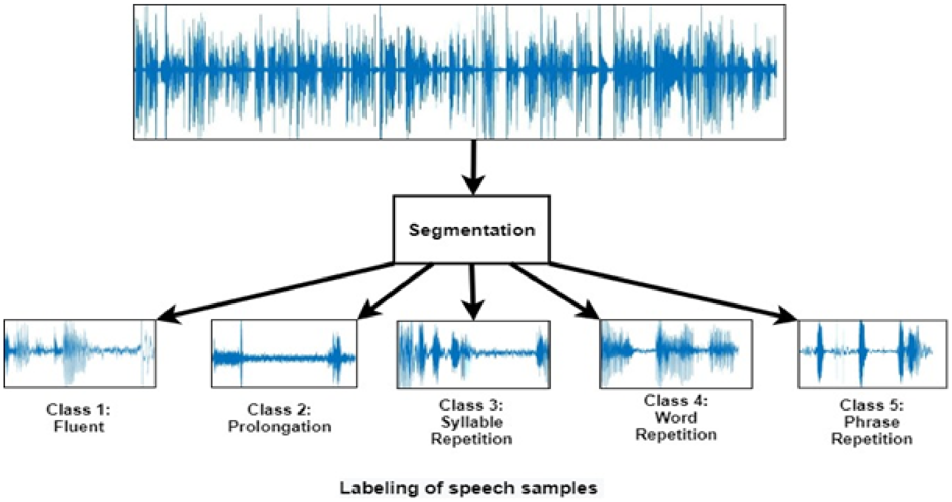

Segmentation
Class 1: Fluent
Class 2: Prolongation
Class 3: Syllable Repetition
Class 4: Word Repetition
Class 5: Phrase Repetition
Labeling of speech samples

Fig 4. Disfluent speech sample segmentation and labelling

2.3 Labeled samples splitting

The segmented and labelled disfluent speech samples were divided into three sets for training, validation, and testing. The training set is a subset of annotated stuttered speech samples for training the classification model. The validation set is used to optimize the performance of the model by reconfiguring the different hyperparameter values. It is smaller than the training set. The test set determines the absolute accuracy of the model and helps in analyzing the performance of proposed model. In this study, the datastore of disfluent speech samples is split into training, validation, and test set in the ratio of 60%, 20%, and 20%, respectively.

2.4 Speech feature extraction

The extraction of speech features is a sort of dimension reduction technique applied to minimize the enormous data to be processed by an algorithm. The critical objective of feature extraction is to upbraid the speech signal into the various acoustically recognizable elements and get the feature vectors with a minor amendment to keep the processing efficient. The proposed work has applied frequency-domain based MFCC and its type for assessing speech disfluencies (Figure 11).

2.4.1 Mel Frequency Cepstrum Coefficients (MFCC)

MFCC (30) is among the most prominent techniques for extracting features for speech recognition. It is based on the frequency domain using the Mel scale, evolved from the human ear scale. These coefficients are stable and accurate to speaker-dependent variations and recording conditions. MFCCs are commonly derived using the following steps described below (Figure 5)[30].

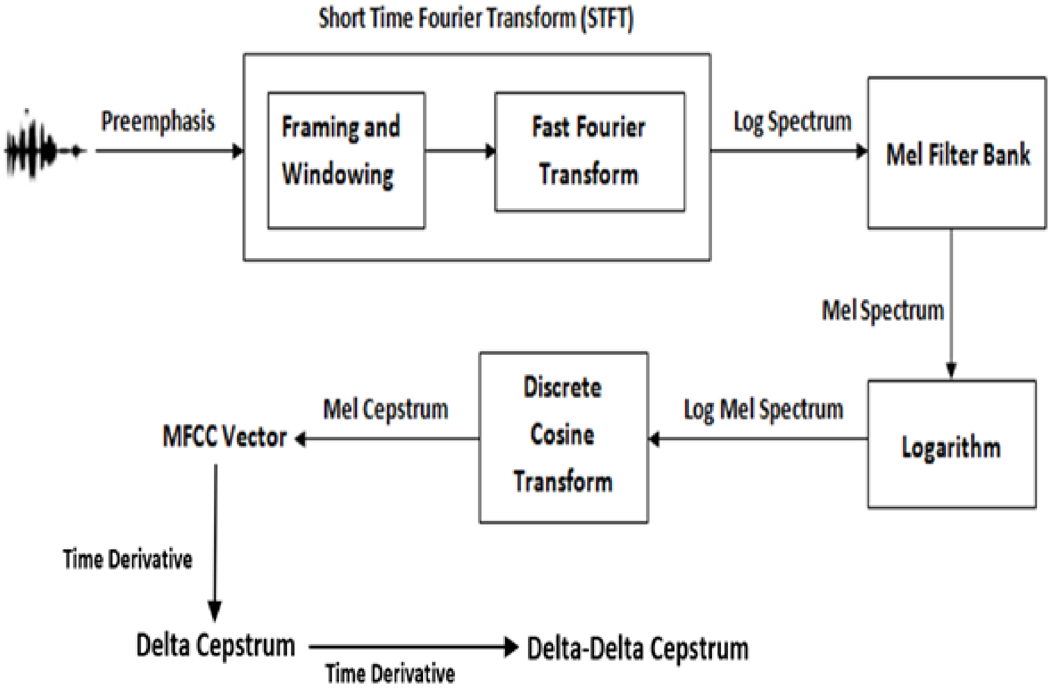


Fig 5. Block diagram of MFCCand its derivatives

(i) Pre-emphasis

The first stage pre-emphasizes the signal spectrums by raising the high frequencies (Figure 6). A low order digital system is employed to flatten the signal spectrally, making it less sensitive to find accurate results later in signal processing[28]. Generally, a first-order FIR filter is represented as Eq. (1).

$$H(z) = 1 - \alpha z^{-1} \ , 0.9 < \alpha < 1 \tag{1}$$

The standard value of α is between the range 0.91-0.99.

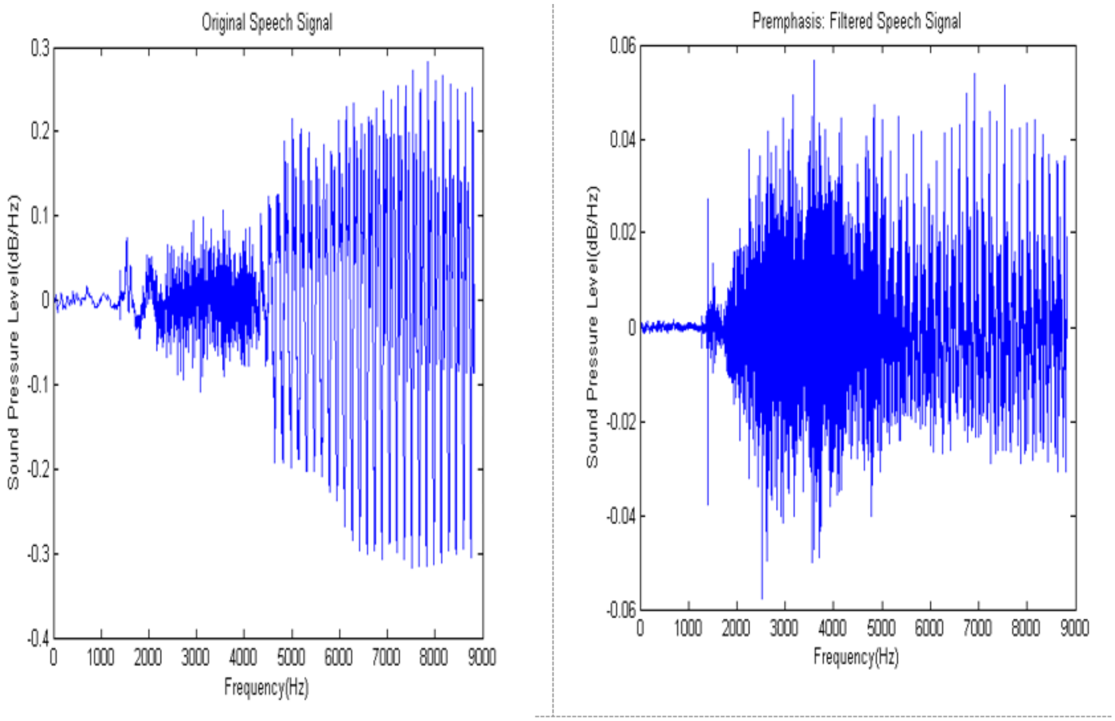


Fig 6. Pre-emphasis of the input signal

(ii) Short-Time Fourier Transform (STFT)

STFT gives the signal's information in both time as well as frequency domain. STFT consists of three steps: Framing, Windowing, and Spectral Estimation, as shown in Figure 8 [30].

Framing: The speech signal is split into small duration blocks, called frames, to perform their spectral analysis. The frame length is defined as the number of milliseconds in each frame, while frame overlapping is the number of overlapping milliseconds between two successive frames [28] (Figure 7).

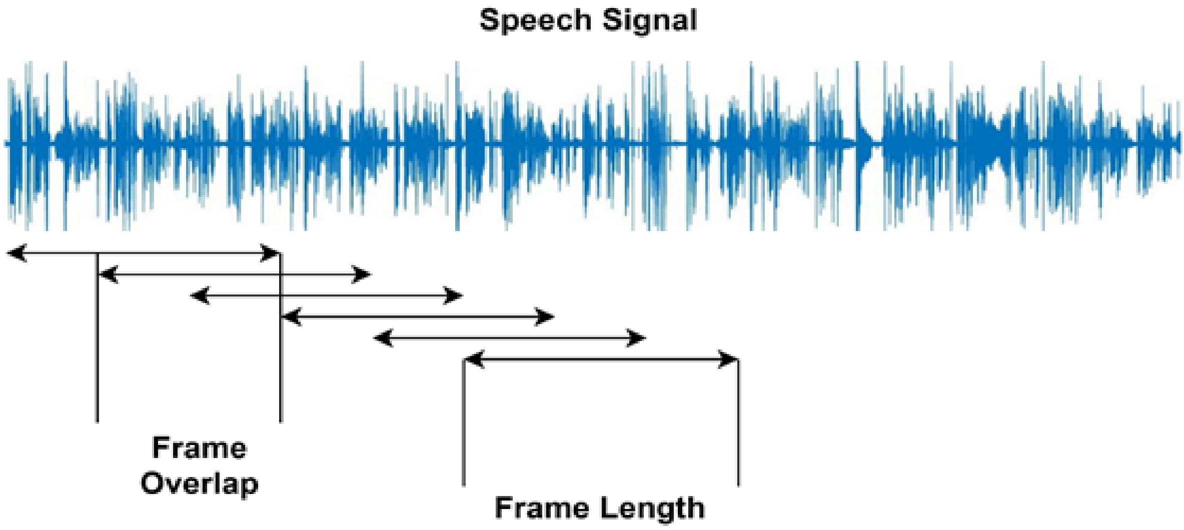


Fig 7. Framing Process

Windowing: A Hamming window function is multiplied by each frame. The window function attenuates the sample values at the beginning and end of each frame, reducing the discontinuity effect caused by framing. The Hamming window is defined as Eq. (2):

$$w(n) = 0.54 - 0.46\,cos(2\pi/N-1) \tag{2}$$

Spectral Estimation**:** Discrete Fourier Transform (DFT) extracts spectral coefficients for discrete frequency bands for a discrete-time signal. DFT is computed by an algorithm known as the Fast Fourier Transform (FFT). It only provides the magnitude of the spectral coefficients. DFT can be defined as Eq. (3):

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{j2\frac{\pi}{N}kn} \tag{3}$$

where X[k] is the spectral coefficients, x[n] is the framed signal, and 0<=n, k>=N-1.

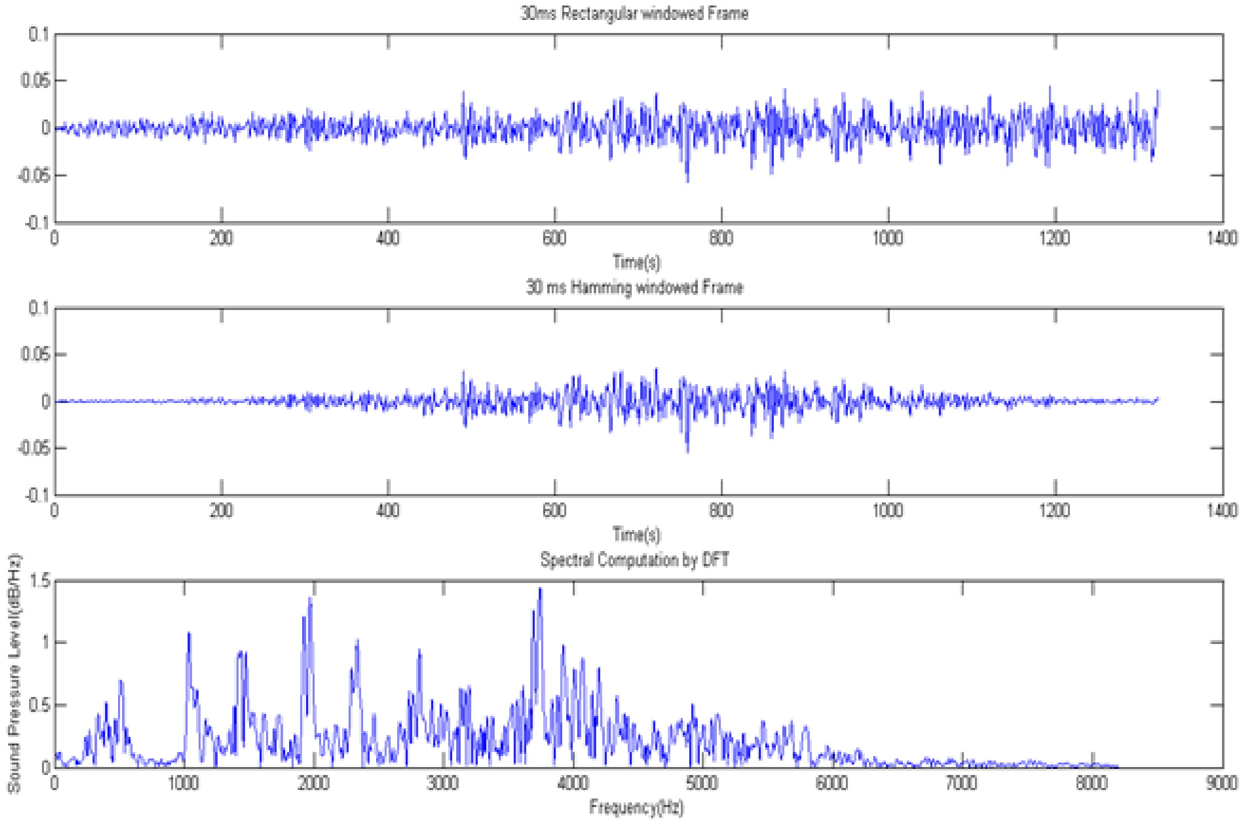


Fig 8. STFT Process

(iii) Mel Frequency Filter Bank

The frequencies output by the DFT is wrapped onto the Mel scale. It constructs a bank of 20 triangular Mel frequency filters that captures energy from each frequency band. The bank of filters (Figure 9) consists of ten filters linearly spaced below 1000 Hz, and the remaining filters spaced logarithmically above 1000Hz. Eq. (4) shows the conversion of linear scale frequency to Mel scale frequency.

$$Mel\,(f) = 2595log_{10}\left(1+\frac{f}{700}\right) \tag{4}$$

Eq. (5) represents the filter bank with M (m = 1, 2, 3….M) filters, where m is the number of triangular filters in the filter bank.

$$Hm(k) = \begin{cases} 0, \text{ for } k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, \text{ for } f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, \text{ for } f(m) \le k \le f(m+1) \\ 0, \text{ for } k > f(m+1) \end{cases} \tag{5}$$

Each triangular filter in the filter bank satisfies Eq. (6).

$$\sum_{m=0}^{M-1} Hm(k) = 1 \tag{6}$$

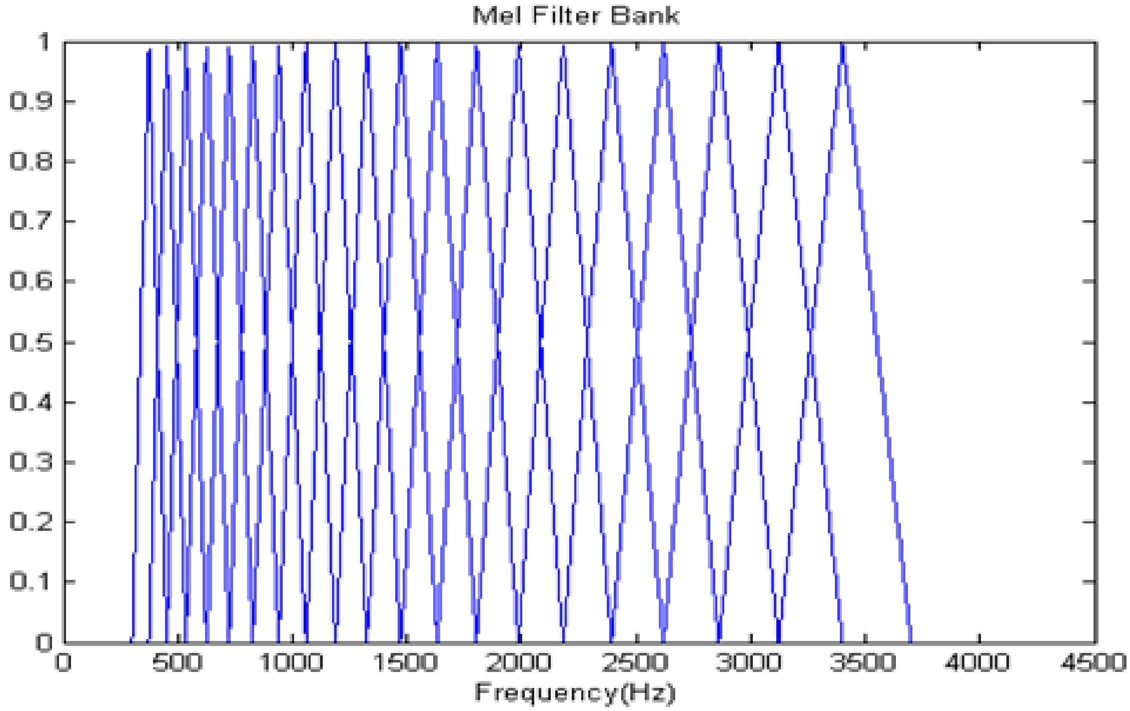


Fig 9. Mel Scale Filter Bank

(iv) Logarithm of Filter Energies

Finally, it calculates the log of all the Mel spectrum values. Eq. (7) is used for computation of the log-energy. 20 numeric values are obtained for each frame as output. They are stored in a matrix. The matrix has rows equal to the number of frames and columns equal to the number of filters in the filter bank.

$$S(m) = \log_{10}\left[\sum_{k=0}^{N-1} |X(k)|^2 Hm(k)\right], 0 \le m \le M \tag{7}$$

(v) Discrete Cosine Transform (DCT)

The filter banks computed above are all overlapping; thus, the filter bank energies are strongly correlated. Hence, the DCT of the log filter bank energies is computed. However, only 14 coefficients are kept for each frame called Mel Frequency Cepstral Coefficient. DCT can be defined as Eq. (8).

$$C(k) = \sum_{m=0}^{M-1} S(m) cos\left(\frac{\pi k\left(m + \frac{1}{2}\right)}{M}\right), 0 \le k \le K \tag{8}$$

where K is chosen as 14. This stage outputs a matrix with rows equal to the number of frames and columns equal to 14.

2.4.2 Delta and Delta-Delta cepstrum coefficients

The features provided by MFCC are static. The dynamic coefficients delta and delta-delta are appended with MFCC to gather dynamic information about speech signals. These features improve the recognition accuracy as they hold account of temporal variability in feature vectors. The first-order derivative of MFCC is delta coefficients, and the second-order derivative is delta-delta coefficients[30]. The delta coefficients are given as Eq. (9).

$$\triangle c_t = \frac{\sum_{K=1}^{M} (c_{t+k} - c_{t-k})}{2\sum_{K=1}^{M} k^2} \tag{9}$$

Where c and $\triangle c$ represent static and dynamin coefficients, respectively. M corresponds to the number of surrounding frames and c_t represents the MFCC feature vector. Delta-delta coefficients are computed similarly as delta coefficients. These obtained features are appended to the original features vectors, resulting in a 28-dimensional Delta MFCC and 42-dimensional Delta-delta MFCC feature vector for each frame.

2.4.3 Weighted MFCC

The overall disfluency recognition rate gets improved by employing delta and delta-delta features. However, it leads to higher computational complexity overhead due to an increase in the feature vector dimension. WMFCC utilizes the benefits of dynamic features with the reduced feature vector dimensions[31]. WMFCC is described as Eq. (10):

$$wc(n) = c(n) + p\bullet \triangle c(n) + q\bullet\ \triangle\triangle c(n), \quad q < p < 1 \tag{10}$$

where p and q are weights of Delta and delta-delta, respectively, and $wc(n)$ is a 14-dimensional WMFCC feature vector. The resultant vector is a fusion of MFCC and its derivatives, thus containing both static and dynamic information of the signal. Moreover, the feature vector is of size 14; thus, incur less computational overhead (Figure 10).

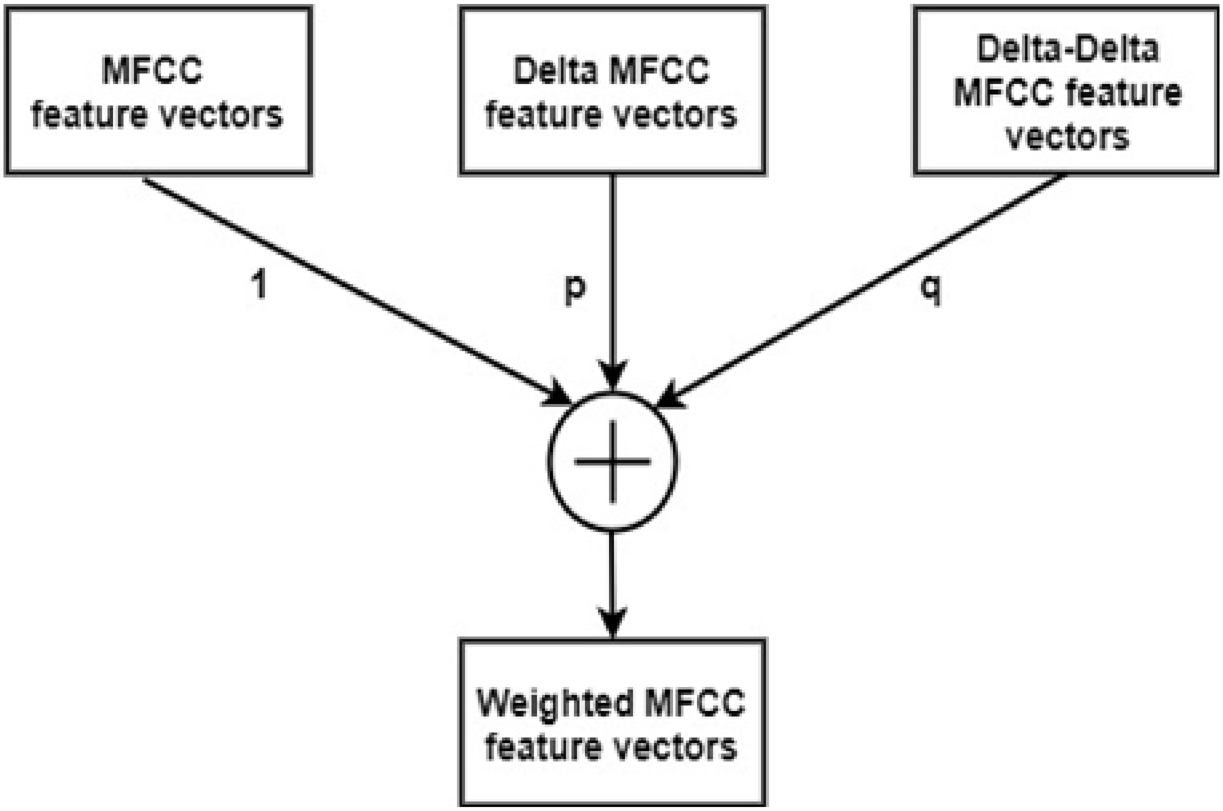

MFCC feature vectors
Delta MFCC feature vectors
Delta-Delta MFCC feature vectors
1
p
q
Weighted MFCC feature vectors

Fig 10. Weighted MFCC

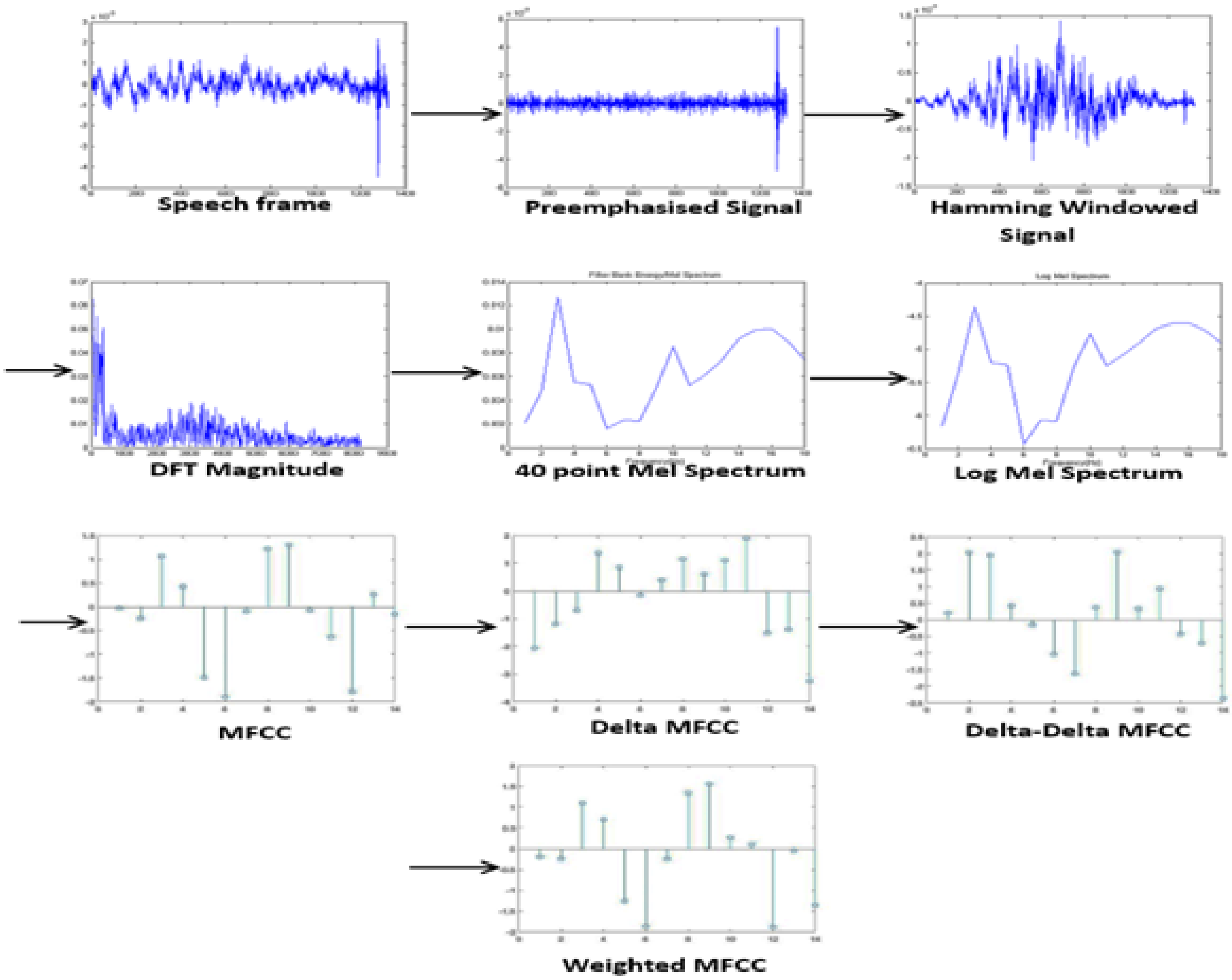


Fig 11. Feature extraction process

2.5 Stuttered speech samples classification

This study applies a deep-learning technique for the classification of stuttered speech samples known as Bi-directional LSTM. The set of features vectors extracted in the above phase are inputted to the classifier. The classifier is trained and validated with 60% and 20% of the segmented stuttered speech samples, respectively. The rest was used for testing the model. The proposed classification model has a better classification accuracy of 96.67% and performs better than other models[26].

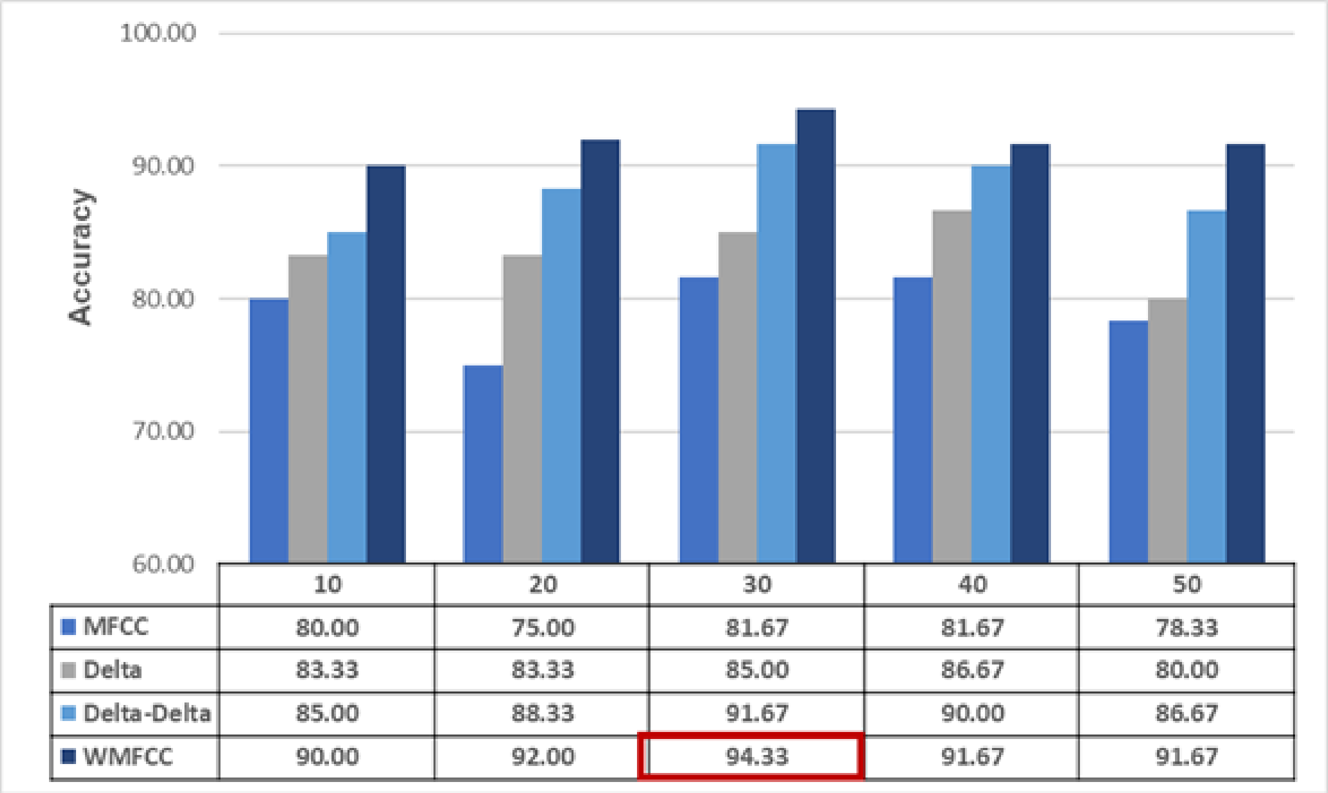
3 Experiments and Results

This section discusses the comparative analysis of proposed WMFCC feature extraction with feature extraction techniques such as, MFCC, Delta MFCC and Delta-Delta MFCC, based on the Bi-LSTM classification results and with some existing works, and also determines the optimal values of parameters required for efficient feature extraction process. The performance of feature extraction methods depends on various parameters such as frame length, frame overlapping (Figure 7) and pre-emphasis factor (Figure 6). Therefore, the classification results were discussed under situations such as different frame sizes, different pre-emphasis filter alpha values, and different percentages of frame overlapping.

The experiments were performed based on the parameter's configuration tabled in Table 2. The first observational study determines the best frame length value by setting the alpha at 0.97 and the percentage of overlapping at 50%. The frame length was varied from 10ms to 50ms for analysis, and the result is presented in Figure 12. It can be seen that 30ms frame length generated better classification accuracy of 94.33% for WMFCC for available stuttered data. The observation states that MFCC, Delta-MFCC, and WMFCC provide the highest average accuracy of 81.67%, 91.67%, and 94.33% respectively for 30ms frame length while Delta-MFCC of 86.67% for 40ms frame length. The above experiment concludes two things, WMFCC outperforms the other three feature extraction techniques with the classification accuracy of 94.33%, and frame length of 30ms gives the best recognition accuracy.

Table 2. Experiments of parameters configuration

Experiments	Frame Length	Alpha values	Frame Overlapping (%)
Frame Length	10ms to 50ms	0.97	50%
Alpha values	30ms	0.91-0.99	50%
Frame Overlapping (%)	30ms	0.98	0 to 75%

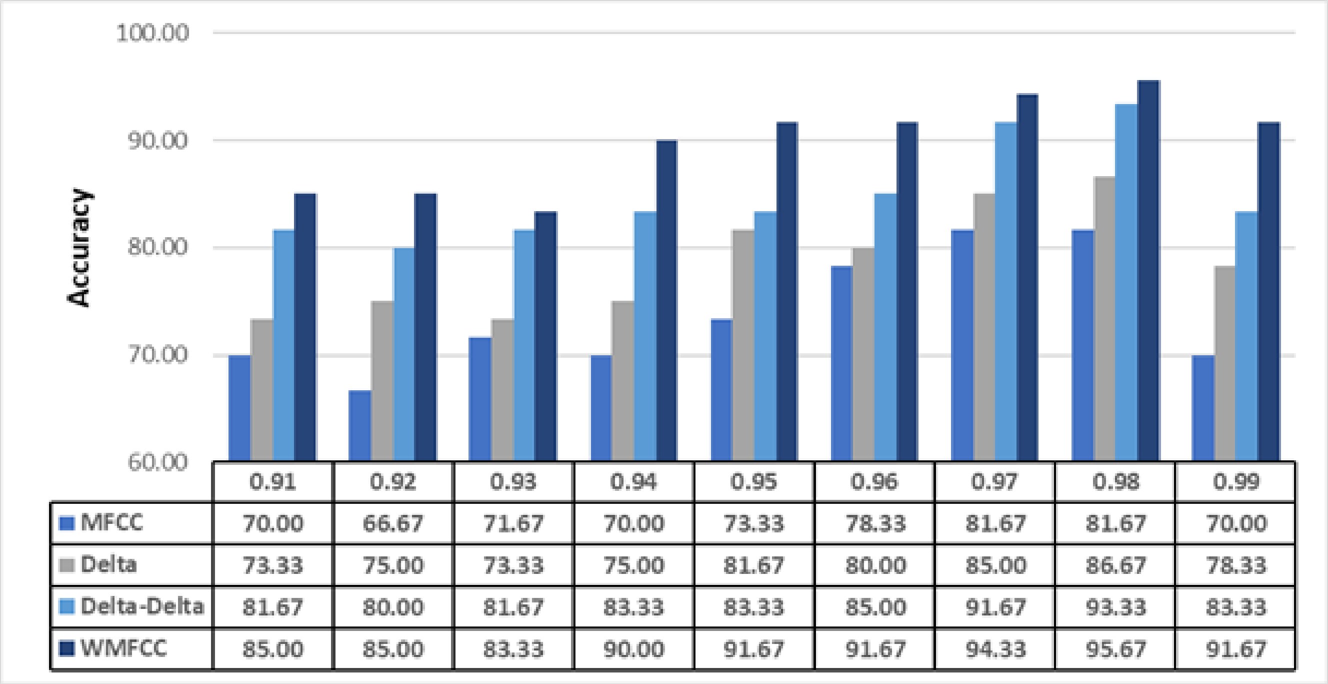


	10	20	30	40	50
■ MFCC	80.00	75.00	81.67	81.67	78.33
■ Delta	83.33	83.33	85.00	86.67	80.00
■ Delta-Delta	85.00	88.33	91.67	90.00	86.67
■ WMFCC	90.00	92.00	94.33	91.67	91.67

Fig 12. Average classification accuracy of four feature, MFCC, delta MFCC, delta-delta MFCC, and WMFCC for different frame length

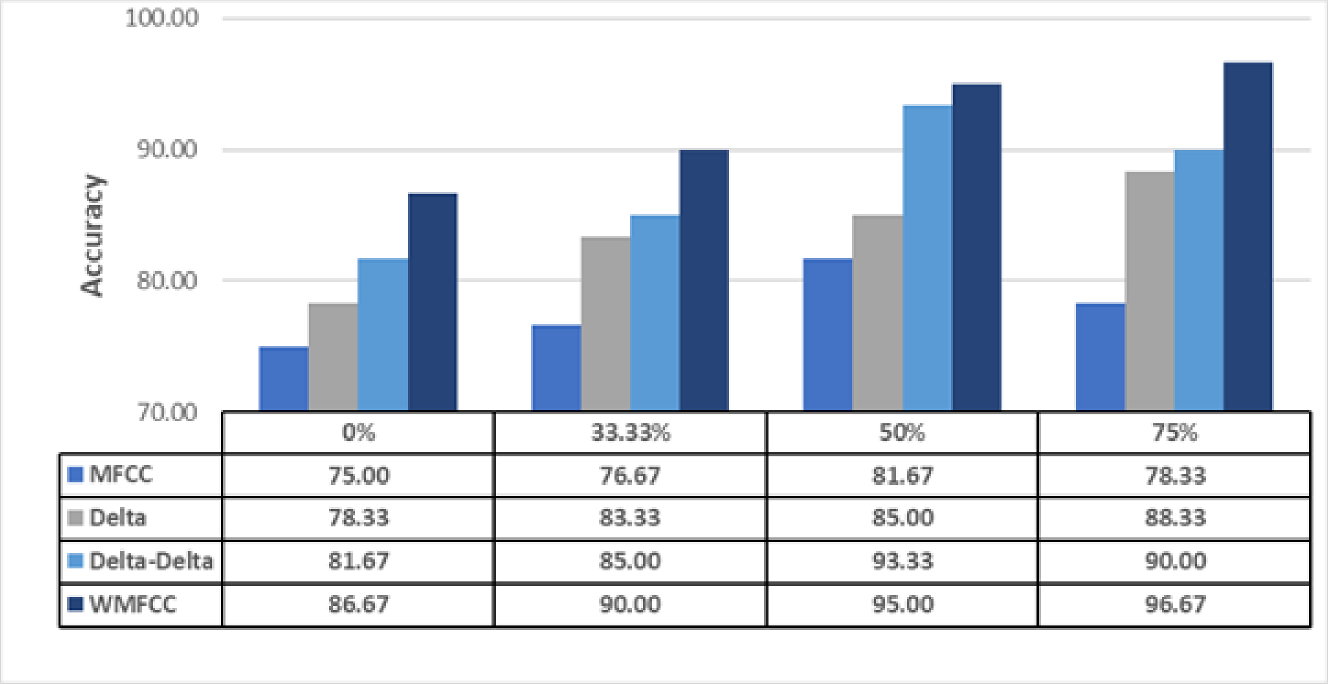
The second observational study analyses the effect of alpha for values between 0.91 and 0.99. The frame length was set as the best value determined in the first experiment and the percentage of frame overlapping as 50%. The average classification accuracy versus alpha is represented in Figure 13. The experiment showed that MFCC, Delta MFCC, Delta-delta MFCC, and WMFCC produced the highest classification accuracy of 81.67%, 86.67%, 93.3%, and 95.67% respectively for the value of alpha as 0.98. Thus, it implies that the optimal value for alpha for controlling the pre-emphasis degree is 0.98, with the WMFCC as a feature extraction technique.

The effect of the percentage of overlapping was analyzed in the third experiment by fixing frame length and alpha values to the best value found in the previous experiments. This study discussed the effect of no overlap, 33.33%, 50%, and 75%, and the results are presented in Figure 14. It can be figured out that 75% frame overlapping outputs best recognition accuracy of speech disfluencies for WMFCC with a value of 96.67% as compared to other techniques. The highest average accuracy given by MFCC and Delta-delta MFCC is 81.67% and 93.33%, respectively, for 50% frame overlapping while 88.33% and 96.67% by Delta-MFCC and WMFCC respectively for 75% frame overlapping. The results elucidated that WMFCC performed consistently better than other features, and features extracted from a higher percentage of overlapping provide optimal classification accuracy.



	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99
■ MFCC	70.00	66.67	71.67	70.00	73.33	78.33	81.67	81.67	70.00
■ Delta	73.33	75.00	73.33	75.00	81.67	80.00	85.00	86.67	78.33
■ Delta-Delta	81.67	80.00	81.67	83.33	83.33	85.00	91.67	93.33	83.33
■ WMFCC	85.00	85.00	83.33	90.00	91.67	91.67	94.33	95.67	91.67

Fig 13. Average classification accuracy of four feature, MFCC, delta MFCC, delta-delta MFCC, and WMFCC for different alpha values



	0%	33.33%	50%	75%
■ MFCC	75.00	76.67	81.67	78.33
■ Delta	78.33	83.33	85.00	88.33
■ Delta-Delta	81.67	85.00	93.33	90.00
■ WMFCC	86.67	90.00	95.00	96.67

Fig 14. Average classification accuracy of four feature, MFCC, delta MFCC, delta-delta MFCC, and WMFCC for different percentages of overlapping

The observational studies above strongly recommend that the WMFCC feature extraction method is superior to widely used MFCC feature extraction technique for automatic assessment of the stuttered speech. WMFCC combines both delta and delta-delta cepstrum with MFCC vectors according to the weights p and q as in the Eq. (10). The experiments were performed for various combinations of p and q to obtain an optimal pair. The computed results are presented in Table 3. It can be determined from the results at 1/3 and 1/6 as p and q values provide the highest recognition accuracy, respectively.

Table 3. Classification accuracy of a different combination of p and q for WMFCC

P	q	Recognition Accuracy (%)
1/2	1/3	90
1/2	1/4	92.67
1/2	1/6	85.67
1/3	1/4	94.33
1/3	1/5	94.33
1/3	1/6	96.67
1/4	1/6	86.67

From the Figures 12, 13 and 14, it can be deduced that WMFCC performs marginally better than Delta-delta MFCC and significantly outperforms Delta MFCC and MFCC for three speech parameterization parameters like frame length, pre-emphasis filter alpha, and frame overlapping for Bi-LSTM. Moreover, Delta and Delta-delta MFCC gave better accuracy than MFCC because they both are dynamic coefficients and keep account of temporal variability. By taking both Delta and acceleration coefficients and with fewer cepstral coefficients, WMFCC maximizes the classification accuracy compared to MFCC and reduces the computational overhead to the classification stage. The optimal values determined for parameters are 30ms frame with a frame overlap of 75% and the alpha as 0.98. The summarized analytical result is presented in Table 4.

Table 4. Experimentally optimal parameters after observational studies for MFCC, Delta MFCC, Delta-Delta MFCC, and WMFCC features

Features	Size of feature vector	Frame Length (ms)	Window Overlap (%)	Alpha	Recognition Accuracy (%)
MFCC	14	30	50	0.98	81.67
Delta MFCC	28	30	75	0.98	88.33
Delta-Delta MFCC	42	30	50	0.98	93.33
WMFCC	14	30	75	0.98	96.67

Table 5. Comparison with existing works

	Proposed Technique	(21)	(17)	(12)	(8)
Classification Accuracy	96.67%	93.5%	94%	95%	88.05%

Table 5 presents a comparison of classification accuracy of proposed method with the other existing feature extraction techniques, employed in papers like (21), (17), (12) and (8). The proposed WMFCC feature extraction method with the deep learning technique Bi-Directional LSTM shows an average classification accuracy of 96.67% while (21) applied Gated Recurrent CNN for classification and MFCC for feature extraction and achieved an average accuracy of 94%. (17) employed MFCC, formant, and the shimmer employed for speech parameterization and Dynamic Time Warping (DTW) for classification purposes and yielded an accuracy of 94%. (12) carried out the comparative analysis of classifiers such as k-NN, LDA, and SVM for classifying repetition and prolongation dysfluencies. The feature extraction techniques used are MFCC, PLP, and LPC. SVM achieved the highest rate of accuracy of 95%. (8) performed speech parametrization using LPCC technique and classification by two classifiers, LDA and k-NN and the average accuracy rates of recognition achieved were 88.05%. Thus, it can be concluded that proposed work provides an efficient feature extraction technique with high success rate, and is dynamic in nature, incurs less computational overhead and integrates well with the deep learning technique Bi-Directional LSTM, for the classification of stuttered events. However, a direct comparison cannot be made due to different languages, different classifiers, and different types, size, and categorical distribution of stuttered speech database, as well as ways of segmentation of database for gathering, stuttered speech samples.

4 Conclusion

In this study, speech parameter WMFCC, were extracted, and the Bi-directional LSTM classifier was used for automated assessment of the stuttered speech. The speech parameterization technique was compared with namely, MFCC, Delta MFCC and Delta-delta MFCC, based on the recognition accuracy of four forms of disfluencies, prolongation, syllable, word, and phrase repetition and with other existing models. Experimental results of this study display that WMFCC slightly outperforms Delta-delta MFCC and significantly outperforms Delta MFCC and MFCC in all situations of frame length, alpha values, and frame overlap percentage. The optimally configured 14-dimensional WMFCC features have the highest accuracy of 96.67%, while 14-dimensional MFCC features have 81.67% accuracy. WMFCC fusions MFCC features with dynamic coefficients, Delta and Delta-delta MFCC. Thus, WMFCC significantly increases the detection accuracy of stuttered events as compared to existing methods and reduces the computational overhead to the classification stage. The optimal values of frame length, alpha, and percentage of frame overlapping observed in the performance analysis are 30ms, 0.98, and 75%, respectively. The current study also proved that Bi-directional LSTM could be employed for the disfluency classification. In the future study, other feature extraction and classification techniques may be applied to improve speech disfluencies' recognition accuracy.

References

1) Silverman F. Stuttering and other fluency disorders. and others, editor;Waveland Press. 2004.
2) Erickson S, Block S. The social and communication impact of stuttering on adolescents and their families. *Journal of Fluency Disorders*. 2013;38(4):311–324. Available from: https://dx.doi.org/10.1016/j.jfludis.2013.09.003.
3) Guitar B. Stuttering : an integrated approach to its nature and treatment. Williams, Wilkins, editors;Williams, Wilkins. 2014.
4) Gupta S, Shukla RS, Shukla RK. Literature survey and review of techniques used for automatic assessment of Stuttered Speech. *Int J Manag Technol Eng*. 2019;9:229–240. Available from: http://ijamtes.org/VOL-9-ISSUE-10-2019.
5) Ravikumar KM, Rajagopal R, Nagaraj HC. An Approach for Objective Assessment of Stuttered Speech Using MFCC Features. 2009. Available from: http://itie.in/Ravi_Paper_itie_ICGST.pdf.
6) Thiang W. Speech Recognition Using LPC and HMM Applied for Controlling Movement of Mobile Robot. . In: and others, editor. Semin Nas Teknol Inf. 2010. Available from: http://fportfolio.petra.ac.id/user_files/97-031/Thiang-Paper-SNTI.pdf.
7) Chee LS, Ai OC, Hariharan M, Yaacob S. MFCC based recognition of repetitions and prolongations in stuttered speech using k-NN and LDA. SCOReD2009 - Proc. *IEEE Student Conf Res Dev*. 2009;p. 146–149. Available from: https://doi.org/10.1109/SCORED.2009.5443210.
8) Chee LS, Ai OC, Hariharan M, Yaacob S. Automatic detection of prolongations and repetitions using LPCC. In: and others, editor. Int Conf Tech Postgraduates. 2009. Available from: https://doi.org/10.1109/TECHPOS.2009.5412080.
9) Kumar KMR, Ganesan S. Comparison of Multidimensional MFCC Feature Vectors for Objective Assessment of Stuttered Disfluencies. *Int J Adv Netw Appl*. 2011;860:854–860. Available from: http://www.ijana.in/papers/v2i5-9.pdf.
10) Ai OC, Hariharan M, Yaacob S, Chee LS. Classification of speech dysfluencies with MFCC and LPCC features. *Expert Systems with Applications*. 2012;39(2):2157–2165. Available from: https://dx.doi.org/10.1016/j.eswa.2011.07.065.
11) Hariharan M, Vijean V, Fook CY, Yaacob S. Speech stuttering assessment using sample entropy and Least Square Support Vector Machine. In: and others, editor. Proc - 2012 IEEE 8th Int Colloq Signal Process Its Appl CSPA. 2012;p. 240–245. Available from: https://doi.org/10.1109/CSPA.2012.6194726.
12) Fook CY, Muthusamy H, Chee LS, Adom AHB, Yaacob SB. Comparison of speech parameterization techniques for the classification of speech disfluencies. *Turkish journal of electrical engineering and computer sciences*. 2013;21(1):1983–1994. Available from: https://dx.doi.org/10.3906/elk-1112-84.
13) Świetlicka I, Kuniszyk-Jóźkowiak W, zbieta Smołka E. Hierarchical ANN system for stuttering identification. *Computer Speech & Language*. 2013;27(1):228–242. Available from: https://dx.doi.org/10.1016/j.csl.2012.05.003.
14) Pálfy J. Analysis of Dysfluencies by Computational Intelligence. In: and others, editor. In-formation Sci Technol Bull ACM Slovakia;vol. 6. 2014;p. 45–58. Available from: http://acmbulletin.fiit.stuba.sk/abstracts/palfy2014.pdf.
15) Jabeen S, Ravikumar KM. Analysis of 0dB and 10dB babble noise on stuttered speech. *Proc IEEE Int Conf Soft-Computing Netw Secur ICSNS*. 2015. Available from: https://doi.org/10.1109/ICSNS.2015.7292422.
16) Savin PS, Ramteke PB, Koolagudi SG. Recognition of repetition and prolongation in stuttered speech using ANN. *Smart Innovation, Systems and Technologies*. 2016;p. 65–71. Available from: https://doi.org/10.1007/978-81-322-2538-6_8.
17) Ramteke PB, Koolagudi SG, Afroz F. Repetition detection in stuttered speech. In: and others, editor. Smart Innovation, Systems and Technologies. Springer Science and Business Media Deutschland GmbH. 2016;p. 611–617. Available from: https://doi.org/10.1007/978-81-322-2538-6_63.
18) Mahesha P, Vinod DS. Automatic segmentation and classification of dysfluencies in stuttering speech. *ACM Int Conf Proceeding Ser*. 2016. Available from: https://doi.org/10.1145/2905055.2905245.
19) Esmaili I, Dabanloo NJ, Vali M. Automatic classification of speech dysfluencies in continuous speech based on similarity measures and morphological image processing tools. *Biomedical Signal Processing and Control*. 2016;23:104–114. Available from: https://dx.doi.org/10.1016/j.bspc.2015.08.006.
20) Ghonem S, Abdou S, Esmael M, Ghamry N. Classification of Stuttering Events Using I-Vector. In: The Egyptian Journal of Language Engineering;vol. 4. Egypt J Lang Eng. Egypts Presidential Specialized Council for Education and Scientific Research. 2017;p. 11–19. Available from: https://dx.doi.org/10.21608/ejle.2017.59395.
21) Bhatia G, Saha B, Khamkar M, Chandwani A, Khot R. Stutter Diagnosis and Therapy System Based on Deep Learning. 2020. Available from: https://www.researchgate.net/publication/343005525_Stutter_Diagnosis_and_Therapy_System_Based_on_Deep_Learning.
22) Girirajan S, Sangeetha R, Preethi T, Chinnappa A. Automatic Speech Recognition with Stuttering Speech Removal using Long Short-Term Memory (LSTM). *Int J Recent Technol Eng*. 2020;8(5):1677–1681. Available from: https://doi.org/10.35940/ijrte.E6230.018520.
23) Katyal A, Kaur A, Gill J. Automatic Speech Recognition: A Review. *Int J Eng Adv Technol*. 2014;3(2). Available from: https://www.ijeat.org/wp-content/uploads/papers/v3i3/C2568023314.pdf.
24) Arjun KN, Karthik S, Kamalnath D, Chanda P, Tripathi S. Automatic Correction of Stutter in Disfluent Speech. *Procedia Computer Science*. 2020;171:1363–1370. Available from: https://dx.doi.org/10.1016/j.procs.2020.04.146.

25) Hariharan M, Chee LS, Ai OC, Yaacob S. Classification of Speech Dysfluencies Using LPC Based Parameterization Techniques. *Journal of Medical Systems*. 2012;36(3):1821–1830. Available from: https://dx.doi.org/10.1007/s10916-010-9641-6.

26) Gupta S, Shukla RS, Shukla RK, Verma R. Deep Learning Bidirectional LSTM based Detection of Prolongation and Repetition in Stuttered Speech using Weighted MFCC. *Int J Adv Comput Sci Appl*. 2020;11(9). Available from: https://doi.org/10.14569/IJACSA.2020.0110941.

27) Howell P, Davis S, Bartrip J. The University College London Archive of Stuttered Speech (UCLASS). *Journal of Speech, Language, and Hearing Research*. 2009;52:556–569. Available from: https://dx.doi.org/10.1044/1092-4388(2009/07-0129).

28) Rabiner LR, Juang BH. Fundamentals of speech recognition. and others, editor;USA. Prentice-Hall, Inc.. 1993.

29) Bachu RG, Kopparthi S, Adapa B, Barkana BD. Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy. *Adv Tech Comput Sci Softw Eng*. 2010;p. 279–282. Available from: https://doi.org/10.1007/978-90-481-3660-5-47.

30) Huang X. Spoken language processing : a guide to theory, algorithm and system development. Prentice Hall PTR. 2001.

31) Chapaneri VS. Spoken Digits Recognition using Weighted MFCC and Improved Features for Dynamic Time Warping. *Int J Comput Appl*. 2012;40(3):6–12. Available from: https://doi.org/10.5120/5022-7167.