MFCC and Texture Descriptors based Stuttering Dysfluencies Classification using Extreme Learning Machine

Roohum Jegan, R. Jayagowri
Department of Electronics and Communication Engineering
BMS College of Engineering
Bangalore, India

***Abstract*—Stuttering is a type of speech disorder which results in disrupted flow of speech in the form of unintentional repetitions and prolongation of sounds. Stuttering classification is important for speech pathology treatment and speech therapy techniques which decreases speech disfluency to some extent. In this article, a method for prolongation and repetition classification is presented based on Mel-frequency cepstral coefficients (MFCC) and texture descriptors. Initially, MFCC and filter bank energy (FBE) matrix are computed. Gray level co-occurrence matrix (GLCM) and Gray level run length matrix (GLRLM) textural features are extracted from these matrices. Laplacian score-based feature selection approach is employed to choose relevant features. Finally, extreme learning machine (ELM) is utilized to classify the speech audio event as repetition or prolongation. The algorithm is evaluated using UCLASS database and has achieved improved performance with classification accuracy of 96.36%.**

Keywords—Voice disorder; Mel-frequency cepstral coefficients; gray level co-occurrence matrix; GLRLM; Laplacian score; extreme learning machine

I. Introduction

Speech forms a major part in day-to-day communication and is used by humans to express their emotions and to exchange their ideas. Thus, speech helps in the efficient communication of ideas that determines how a person thinks and feels. Speech is a special gift to the mankind because animals and other species cannot speak [1]. Stuttering is classified as one of the speech disorders and is identified by reiteration of utterances, phonics, phrases, or terms; elongation of sounds during utterance; and interventions in speech called as blocks [2]. Even though there is no complete cure for stuttering at present, there are numerous speech pathology approaches that may aid to decrease speech disfluency to certain extent. To judge the performance of the stutterers before and after the treatment, stuttering assessment is needed. Generally, Speech Language Pathologist (SLP) manually enumerates and categorize eventuality of disfluencies such as prolongations and repetitions in a stammered speech. But, this type of evaluation is unpredictable, uncertain, intuitive, cumbersome and erroneous. Hence it would be worthy if the stuttering assessment can be carried out automatically enabling the SLP to spend more time with the stutterer in treatment session.

This article presents a new statistical feature based on MFCC and FBE matrix to enhance stuttering event classification using UCLASS database. Prolongation and repetition event are discriminated using GLCM and GLRLM features extracted from MFCC and FBE matrix and ELM classification. Laplacian score- based feature selection algorithm is employed to discard irrelevant features resulting in improvement in the classification rate. The proposed feature extraction approach improves the prolongation and repetition classification accuracy to a greater extent. Moreover, best features are selected using Laplacian score-based feature selection algorithm, thereby, minimizing the computational complexity.

The rest of the paper is organized as follows: In Section II, the past solutions for the predetermined problem via different algorithms, classification and feature extraction techniques are presented. In Section III, the proposed method for speech dysfluencies has been discussed with brief description of each method in separate subsections. In Section IV, the simulation results of the work are discussed, and Section V presents the conclusion of the work with the future scope.

II. Literature Review

This section presents different objective approaches proposed for stuttering event classification based on various features, datasets and classifiers. In [3], automatic detection of syllable repetitions is presented using correlation of 1/3 octave spectra. The correlation features are used to identify repeated syllables with similar spectral components. Acoustic and pitch related descriptors including MFCCs, formants, tonality (pitch), zero crossing rate (ZCR) and energy are employed to classify repetitions and prolongations using Artificial Neural Networks (ANN) in [4]. The accuracy obtained using the ANN based classification was 87.39%. Line spectral frequency (LSF) representation features are extracted and classified using three different classifiers: MLP, RNN and RBF resulting 98-100% detection rate in [5].

LP-Hilbert transform based MFCC (LH-MFCC) based feature extraction method is presented to classify three different dysfluencies using Gaussian Mixture Model (GMM) classifier [6]. These features efficiently capture temporal and spectral parameters of utterances resulting in 94.98% accuracy rate. To enhance classification accuracy, a decision fusion technique is introduced, based on combination of different acoustical features like ZCR, speech envelope (ENV) for classifying filled pause (FP) and elongation (ELO) in Malay language [7]. Stuttered speech repetition detection algorithm based on MFCC and dynamic time warping (DTW) with accuracy of 83-90% is proposed in [8], [9].

Various stuttering events are classified in [10] using i-vector based KNN and LDA classification resulting in 80- 85% classification accuracy. In [11], similarity matrix image is computed using MFCC, PLP and filter bank energy feature sets. Dysfluent regions are detected using threshold based morphological image processing having an average classification accuracy of 82.5%. SVM based dysfluency classification method using a GMM supervector is introduced in [12], [13] with +96.10% accuracy. Repetition and prolongation classification using MFCC, LPC and perceptual linear predictive (PLP) and k-NN and SVM classifier is proposed in [14], [15] having classification rate of 96%.

A method using SVM classifier and fusion of prosodic (pitch and energy) and cepstral (MFCC) features is presented for stuttered speech classification with 97.80% accuracy in [16]. A deep belief network architecture is developed based on MFCC and LPCC features to classify stuttering speech signal having an accuracy of 85% in [17]. Computational intelligence approach based on ANN and SVM is developed to classify dysfluencies in stuttered speech signal in [18] with 85% accuracy rate. An objective methodology for dysfluency detection using six-level wavelet packet transform decomposition and features employing entropy features is presented in [19], [20]. Performance of the algorithm is evaluated using three distinct classifiers including k-NN, LDA and SVM classifiers resulting in classification accuracy of 96.67%. MFCC and LPCC based stuttered event classification approach is proposed in [21] using k-NN and LDA classifiers with 94% classification rate. Prolongation and repetition in stuttered speech classification technique using LPCC features and k-NN/LDA classifier is presented with 89.77% in [22].

In another study, [23], [24], same authors presented stuttered event detection approach using LPC, LPCC and WLPCC features with 97.78% classification accuracy. But the test segments taken were very small and it was observed that accuracy decreased for bigger test segments. MLP network architecture is presented in [25] to detect stop consonant repetitions with accuracy of 76.67%. This study presents a new approach based on MFCC and FBE matrix representation and feature extraction using GLCM and GLRLM descriptors. Convolutional Method (CNN) was used to classify different languages in [42]. The classification accuracy obtained is 97.86%.

The existing literature feature extraction approaches are primarily based on time-domain features extracted from the speech sample. The stuttering classification rate is limited between 87% and 94%. Additionally, the feature selection techniques are less explored in the existing literature hence limiting the classification accuracy. This article presents a new feature extraction algorithm based on MFCC and FBE matrix statistical features. The proposed feature extraction approach enhances the prolongation and repetition accuracy. Moreover, important features are selected using Laplacian score-based feature selection algorithm, thereby, reducing the computational complexity.

III. Proposed Method for Speech Dysfluencies Classification

This article presents prolongation and repetition classification using MFCC/FBE, two different types of texture descriptors and ELM classifier. The proposed scheme is presented in this section. MFCC and filter bank energy computation is explained next along with its importance in the stuttering classification process. GLCM and GLRLM texture features are investigated and discussed in detail. Laplacian score-based feature selection is employed in this study along with its brief introduction. Finally, extreme learning machine classifier that is employed in this work and the merits are discussed.

A. Architecture of MFCC and FBE based Dysfluencies Classification

The framework of the stuttering classification scheme is depicted in Fig. 1. It is the architecture used for stuttering classification using MFCC and texture descriptors. The sample voice is analyzed before taking it as an input from a person. At various stages, the input speech signal is manipulated and undergoes operations of Pre-processing, converting into frames, filtering and Windowing, and complementing with the uttered word. This speech algorithm has two major stages: training and testing stages and the process is shown in Fig. 2.

After pre-processing the input speech sample, MFCC and FBE matrix is obtained. GLCM and GLRLM descriptors are extracted from these two matrices. In order to reduce feature vector dimensionality, Laplacian score-based filter type feature selection is used. Finally, ELM is trained using the training database. In our experiments, 70% of the speech specimens are used for training stage and remaining 30% samples for testing.

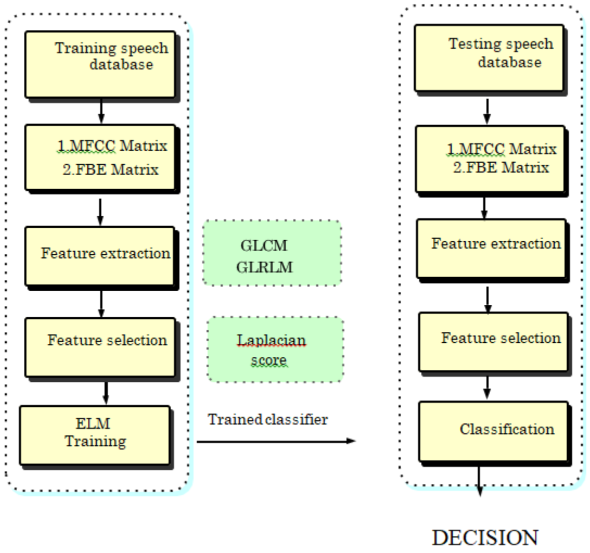

Fig. 1. Architecture of Stuttering Classification using MFCC and Texture Descriptor.

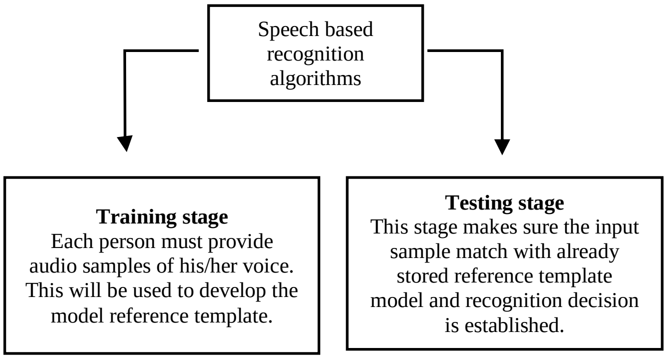
Speech based recognition algorithms
Training stage
Each person must provide audio samples of his/her voice. This will be used to develop the model reference template.
Testing stage
This stage makes sure the input sample match with already stored reference template model and recognition decision is established.

Fig. 2. Training and Testing of Voice Recognition Algorithms.

B. Mel-Frequency Cepstral Coefficients (MFCCs)

MFCC represents set of Discrete Cosine Transform decorrelated variables that are evaluated using transmutation of the energies (output from filter) that are compressed logarithmically. These are obtained from a sharply spaced triangular filter bank that precedes the Discrete Fourier Transformed audio signal. The extracted features represent parametric characterization of audio signals that plays important role to enhance the performance of the recognition approach. MFCCs is widely and commonly adopted feature extraction algorithm in variety of audio/speech/music processing algorithms [26]-[30].

MFCC describes short time cepstral features and uses Mel scale with linear separation below 1000 Hz and logarithmic spacing above 1 kHz. The value of Mel for any frequency f (Hz) is computed as:

$$\text{Mel}(f) = 2595 \times \log10(1 + f/700) \quad (1)$$

where M is the quantity of triangular filters, L represents the total Mel scale coefficients and Ek is energy of the filter bank (log) output. Filter bank approach characterizes the speech samples efficiently. A set of triangular band-pass filter is designed, and nonlinear Mel-frequency scale is employed considering human perceptual capabilities with specific frequency spacing. Intensity from each band is computed by multiplying Mel filter bank and magnitude spectrum of speech signal. We observed that, filter bank energy spectrum varies based on the input speech sample (prolongation and repetition). This dissimilarity is exploited in this study for stuttering event classification. The overall architecture and process of generating Mel frequency cepstral coefficients is shown in Fig. 3 [6], [7]:

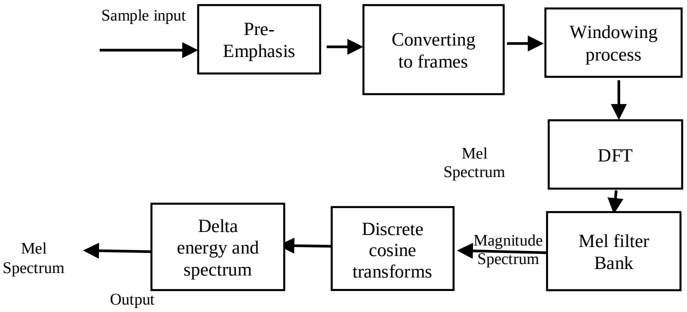

Fig. 3. MFCC Flow.

MFCC constitutes of seven major stages. Every stage has its own mathematical processes and functions described in the following steps:

Step 1: Pre–emphasis: This stage helps in escalating the signal energy at greater frequency range by allowing the signal to pass through a filter that prioritizes higher frequencies.

$$Y[x] = M[x] - aM[x - 1 \quad (2)$$

$$Y[x] = M[x] - 0.96M[x - 1] \quad (3)$$

Let us consider ‘a’ to have a value 0.96, which means it is assumed that there is 96% chance of a sample to regenerate from the previous sample.

Step 2: Converting to frames: Framing refers to segmentation of the voice samples into small framework with distance between 20 to 40 ms. The speech samples are derived from the analog to digital convertor. The input audio signal is segmented into N sample frames. Nearby frames are segregated by Y (Y < M). Y = 100 and M = 256 are the most typical values used.

Taking into account the succeeding step in feature extraction stage, it amalgamates all the immediate frequency bands. If the window is given as H(x); where x = count of samples contained in each frame, B(x) = signal output, A(x) = signal input, H(x) = window. The, the output of hamming window is:

$$B(x) = A(x) \times H(x) \quad (4)$$

$$h(x) = 0.54 - 0.46 \cos \Sigma\, 2\pi x;\ 0 \leq x \leq X - 1 \quad (5)$$

Step 4: Fast Fourier Transform: It transforms each framework of X samples that are in time domain to corresponding frequency domain. FT transforms the convolution of pulse in glottis tt[n] and impulse response I[n] of the vocal tract present in the time domain which is shown in equation stated below:

$$B(x) = \text{FFT}\,[I(t) * A(t) = I(x) * A(x)] \quad (6)$$

If A(x), I(x) and B(x) are the FT of A(t), I(t) and B(t) respectively.

Step 5: Filter Bank Conversion: The FFT range has high frequency and is broad and the audio signal is non-linear. Fig. 4 shown above describes a set of triangular filters to enumerate the weighted sum of all filter spectral samples so that the output is made to approach the Mel scale. Every filter has triangular magnitude frequency response with unit value at the center frequency and it gradually reduces linearly to zero at the Centre frequency of adjoining filters [7], [8]. Output of each filter is the filtered sum of its spectral components. Equation stated below is then used to calculate the Mel for any frequency f (HZ) as:

$$\text{Mel}(f) = 2595 \times \log10(1 + f/700) \quad (7)$$

Step 6: DCT: Log Mel spectrum is converted into time domain using Discrete Cosine Transform and this result in the formation of MFCC. MFC coefficients are also called the acoustic vectors. Hence, each input speech sample is converted into a chain of acoustic vector.

Step 7: Delta Energy and Delta Spectrum: The audio speech signal and the frameworks vary in accordance to the formant slope at its changeovers. Hence, features that relates to the variations in cepstral parameters over time have to be included. 13 delta parameters that include 12 cepstral features and one energy feature, and 39 double delta features are included. The energy E of a signal ‘a’ in a window frame from time duration t11 to time sample t12, is given by the following equation:

$$E = A^2t \qquad (8)$$

Each of the thirteen delta variables constitutes the variation happening between frames corresponding to the energy parameter. On the other hand, 39 double delta features depict the variations among frames in the corresponding delta features as,

$$r(t) = [s\,(t + 1) - s\,(t - 1)]/2 \qquad (9)$$

In short, the MFCC computation comprises of the framing stage where the input pre-processed speech sample is divided into several frames with overlap. After framing the signal, hamming window attenuates the framed signal to null at the beginning and end of the frame. The windowed signal is converted to frequency domain by applying Fast Fourier transform (FFT). The FFT spectrum is passed through a set of triangular band-pass filter to obtain the logarithm energy spectrum.

The placement of these filters is based of Mel frequency scale, which is proportional to logarithm of linear frequency scale, reflecting human perceptual capabilities. In the last step, discrete cosine transform (DCT) is applied on logarithm energy to extract L Mel scale cepstral coefficients using the energy compaction property and is obtained as,

$$C(n) = \Sigma\, Ek \times \cos\,(n \times (k - 0.5) \times \pi/40 \qquad (10)$$

Fig. 4 and 5 shows filter bank energy spectrum for prolongation and repetition samples from UCLASS database respectively. It is clearly seen that for prolongation samples, (Fig. 4) the filter bank energy of the frames for prolongation utterances are equal. Whereas, in case of repetition (Fig. 5), filter bank frame energies are distributed more evenly as compared to prolongation frame energies. Moreover, central coefficient energy distribution is higher in Fig. 5 compared to Fig. 4. These differences are exploited for the classification in this article.

C. Gray Level Co-occurrence Matrix (GLCM)

Feature extraction technique is mainly used to make simpler the number of features that can accurately describe a large set of data. While analyzing complex data, large number of variables involves more difficulties. Huge number of variables traditionally requires more memory and computational power. Else it requires a classification algorithm which can fit the entire training sample but that result in poor generalization of new samples. Feature extraction refers to methods used for establishing fusions of the features while still describing the data without compromising on the accuracy. It is mainly used in applications that describes and retains the texture kinesthetic or visual attributes of a surface.

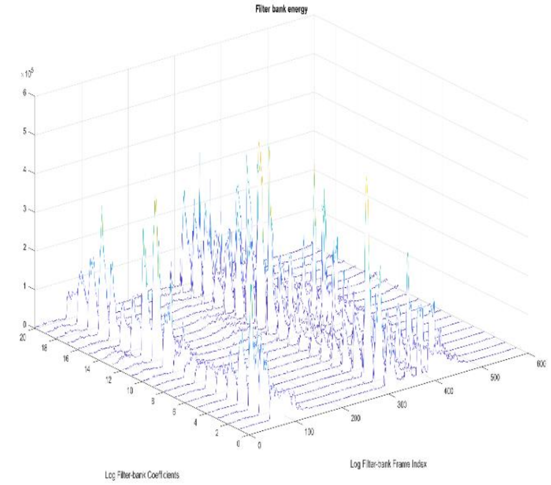

Fig. 4. Filter Bank Energy Spectrum for Prolongation Samples.

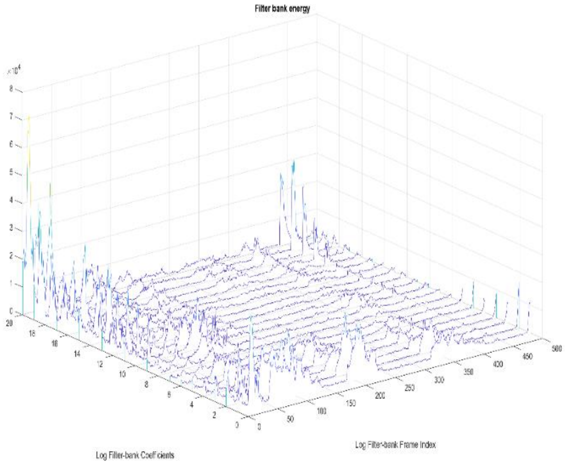

Fig. 5. Filter Bank Energies for Repetition Samples.

Texture analysis helps to find a distinctive way of illustrating the underlying/hidden characteristics of texture and personify them in an effortless and distinct form. This helps in robust, accurate classification and segmentation of samples. Although textural representation contributes a major role in image study and pattern recognition, only a few architectures implement the idea of onboard textural feature extraction method. Gray level co-occurrence matrix helps to obtain qualitative/statistical texture features. Hence GLCM is a numerical analysis method used for observing the texture that contemplates the spatial correlation between pixels [31]. GLCM is one of the most widely used texture descriptor to compute statistical features of the image based on gray level intensities and employed in different image processing applications like image segmentation, image retrieval, image classification and object recognition as discussed in [41].

The main advantage of the co-occurrence matrix computation is that, while considering the relation between two pixels at any instant of time, pixel pairs that coincide can be topologically matched in various inclinations in accordance to the distance and spatial-based angular relationships. This apparently exhibits the combination of grey levels of image matrix and their positions. Matrix relationships are defined by changing the directions with different angles and displacement vectors. In this paper, twenty features are extracted from the MFCC and FBE matrix [32].

The count of rows and number of columns in a GLCM is the count of gray levels, H, present in the image. The matrix component A (a, b δp, δq) is the frequentness at which two pixels, divided by an interval (δp, δq), transpire within a given vicinity, one with potency 'a' and the other with potency 'b'. The component A (a, b δp, δq) has got the values of second order statistical probability for corresponding variations among the gray levels 'a' and 'b' at a specific displacement length l and at a specific angle (θ).

GLC Matrices are very delicate to the dimension of the texture sample on which they are computed because of their large dimensionality. Hence, the reduction in the number of gray levels is of utmost importance. In most of the cases, a reference pixel and its quick neighbor is contemplated. Usage of larger offset is feasible if the window is large enough. The top most cell at the left will contain the frequency of occurrence of combination 0,0. That means, it will contain the information about the total number of times a neighbor pixel having 0 gray level gets placed to the right of reference pixel with 0 gray level, within the image area.

D. Gray Level Run Length Matrix (GLRLM)

This matrix is another popularly used higher order texture descriptor useful in feature extraction. It is a textural representation model that helps in extracting the spatial plane features of each pixel relative to the higher order statistics [33]. A 2- dimensional feature matrix is obtained at the end of the process, exploiting the spatial variations because of the prolongation and repetitions. GLRL matrix is not just limited to 00 direction. It can also be used in other directions with θ = 450, θ = 900 and θ = 1350.

GLRL matrix gives us few textural parameters that can be extracted from it. It is observed that five texture features can be extracted from this GLRL matrix, namely: Shot Runs Emphasis (SRE), Long Runs Emphasis (LRE), Non-uniformity Gray Level (GLN), Non-uniformity Run Length (RLN), and Percentage of Run (RP). Later on, two more features called the Low Gray Level Run Emphasis (LGRE) and High Gray Level Run Emphasis (HGRE) were found to be extracted from this matrix. This parameter makes use of sequential gray level of pixels and then discriminates the texture that has equal values for Shot Runs Emphasis and Long Runs Emphasis with minor variations in the gray level distribution.

After those four more features were found to be extracted from the matrix, namely: Low Short Run Gray-Level Emphasis (SRLGE), High Short Run Gray Level Emphasis (SRHGE), Low Long Run Gray Level Emphasis (LRLGE), and High Long Run Gray Level Emphasis (LRHGE). Same intensity adjacent pixels in certain direction is termed as run length. Each element in the GLRLM characterizes the total gray level occurrences in the given direction. For a single MFCC matrix, it is possible to compute many different run- length matrices f (i, j θ) one for each chosen direction θ. Thus, given a direction, for each acceptable gray measure value, this matrix measures the total run times. GLRLM is parameterized by three different pixel features: intensity, length and direction of a run from a reference pixel. Total of 11 features per direction are extracted from the MFCC and FBE matrix [34]-[36].

E. Laplacian Score based Feature Selection

Feature selection plays important role as the preprocessing step in machine learning to select optimum features from the large input feature set. Feature selection techniques can be classified into: (a) filter and (2) wrapper techniques [37]. Filter methods are independent of the learning algorithm and faster, whereas, wrapper approaches produce higher accuracy and it needs learning algorithm. In this article, Laplacian score- based filter approach is used for feature selection [38]. As the name suggests, for every parameter, its Laplacian score is evaluated and calculated separately to reveal its locality preserving power.

Laplacian score approach is based on the reflection that two data points are probably related to the same point if they are near to each other. Generally, in all the learning tasks like classification, the local geometric structure is more important than the global structure of the given feature space. Hence, a nearest neighbor graph is designed to construct the local structure, and Laplacian score aspires those specific parameters that obey this graph model.

Laplacian score (LS) is based on the concepts of 'Laplacian Eigenmaps' and 'Locality Preserving Projection' and is used to identify importance of individual features. Locality preserving power is computed using this LS for each feature and the features are inferred to be similar if they produce very low LS. Based on the graph, structure is defined using the nearest neighbor and the geometric structure of the descriptor is evaluated. As LS is a ranking filter approach for feature selection, a threshold T is used to select number of features for classification.

F. Extreme Learning Machine (ELM) Classifier

Extreme learning machine algorithm is a contemporary state-of-the art machine learning algorithm with sole-hidden layer feed-forward neural network (SLFNs). ELM is fast; it has better generalization performance and enhances the training speed by assigning the weights randomly. ELM requires only two parameters: (1) hidden layer neural units and (2) their transfer function [39] and [43]. ELM algorithm is used in data classification and regression applications. The optimal values must be chosen for ELM training parameters to enhance the accuracy. However, while designing the classifier using ELM, the number of hidden nodes to be used for handling different problems remains a trial and error [40].

A major drawback of ELM is that the classifying borderline for the learning features of this algorithm may not be an adequate one. This is because the learning features of hidden nodes are arbitrarily allocated while they remain uninterrupted in the training stage [17]. Hence, few features might be miscategorized by the algorithm, mainly for those samples that are close to the classifying border line. Another observation made is that, in many cases, this algorithm might need additional hidden neurons compared to the already available traditional tuning-based algorithms [18]. Few researchers have proposed that the above-mentioned shortcomings of ELM can be overcome by introducing several variants of ELM, such as incremental ELM [9], pruning ELM [12], error-minimized ELM [19], dual-step ELM [20], sequential online ELM [21], evolutionary ELM [18], voting-based ELM [17], ordinal and fully complex ELM [23], and balanced (symmetric) ELM.

IV. Simulation Results

Proposed stuttering event classification algorithm is evaluated using speech samples from UCLASS database [3]. The database includes 43 speakers recording generating 107 audio samples.

In this article, 39 speech samples are selected for classification similar to the settings used in [21]. During MFCC and FBE computation, the analysis frame duration is set to 25 ms with overlap of 10 ms. The pre-emphasis coefficient is set to 0.97 with 20 filter bank channels and 12 cepstral coefficient extractions. Lower frequency limit is set to 300 Hz, whereas 3700 Hz is the high frequency limit.

The GLCM features are extracted using one direction only, as during the experimentation we found that one direction is sufficient to generate satisfactory classification rate. GLCM descriptors are extracted from both MFCC and FBE matrix representation, generating 20-D feature vector for each. In addition to GLCM, GLRLM textural features are also extracted from these two matrices, thereby generating 24-D feature vector for MFCC and FBE matrix individually. Out of the total speech samples, 70% are used to train the ELM and enduring 30% are employed for testing. ELM is implemented using 300 neurons, which is set experimentally with sigmoidal transfer function. Finally, each stuttering speech sample is represented using 84-D feature vector.

Table I shows prolongation and repetition classification accuracy for individual (GLCM and GLRLM separately) and combined feature sets (GLCM+GLRLM). From Table I, it is observed that, GLCM has poor discrimination capability with classification accuracy of only 79.84%. Compared to GLCM features, GLRLM descriptors are more powerful during the classification. Finally, as expected, combined feature set (GLCM+GLRLM) resulted in highest accuracy of 92.64%. It is also evident that, feature fusion enhances the classification rate notably in the stuttering event discrimination. In order to demonstrate the effect of Laplacian score feature selection, additional experiments are performed. Table II depicts prolongation and repetition classification accuracy for individual (GLCM and GLRLM separately) and combined feature set using Laplacian score feature selection approach. Significant improvement in the classification accuracy can be observed (see Table II) by applying the feature selection technique. Combined (GLCM+GLRLM) feature set accuracy obtained was 96.36% using only 25 features. Thus, Laplacian score approach not only enhances the classification rate but decreases the number of features also (60% decrease in total number of features).

As this approach is ranking based approach, threshold T is used to select number of important features from the large input feature set. Fig. 6 shows the classification accuracy obtained using different threshold values. Highest accuracy was obtained at threshold T = 0.2, we choose final feature set with this threshold (resulting in final 25-dimensional relevant features only). As started above, GLCM descriptors can be evaluated using four different directions. Table III illustrates GLCM detection accuracy using different directions employing LS feature selection and without LS feature selection. As evident, feature selection improves the detection rate. It is also worth mentioning that, combining all four directions enhances the detection rate of the proposed technique. The present work utilizes ELM classifier with 300 hidden neurons. Tables IV and V shows the number of hidden neurons and corresponding obtained accuracy using GLCM and GLRLM features, respectively. It was observed that the highest classification rate is achieved using 300 neurons.

TABLE I. Prolongation and Repetition Classification Accuracy for Individual and Combined Feature Set

Features	Classification Accuracy (%)
GLCM	79.84
GLRLM	84.6
Combined	92.64

TABLE II. Prolongation and Repetition Classification Accuracy for Individual and Combined Feature Set using LS feature Selection

Features	Classification Accuracy (%)
GLCM	81.74
GLRLM	87.24
Combined	96.36

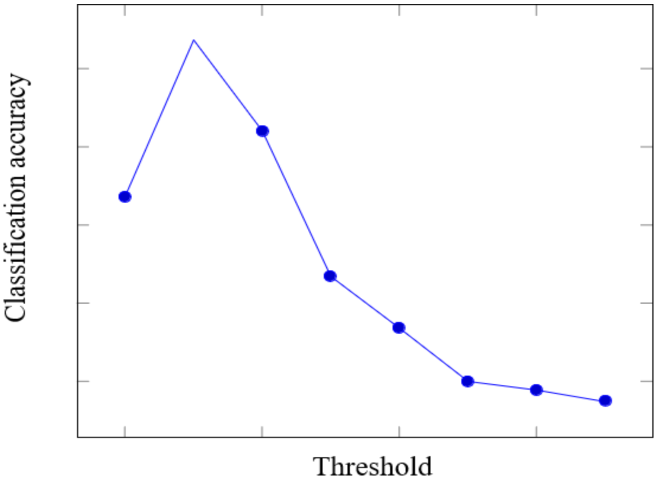

Fig. 6. Classification Accuracy vs Threshold Values using Laplacian Score-based Feature Selection for Stuttering Event Classification.

TABLE III. GLCM Detection Accuracy using Different Directions

Direction	Accuracy without FS	Direction	Accuracy with FS
00	79.84	00	81.74
00 + 450	78.12	00 +450	81.24
00+450+900	78.05	00+450+900	80.71
00+450+900+1350	77.69	00+450+900+1350	79.93

TABLE IV. NUMBER OF HIDDEN NEURONS AND ACCURACY USING GLCM FEATURES

Number of hidden neurons	Classification Accuracy (%)
100	75.29
150	76.35
200	75.85
250	78.16
300	79.84
350	78.97

TABLE V. NUMBER OF HIDDEN NEURONS AND ACCURACY USING GLRLM FEATURES

Number of hidden neurons	Classification Accuracy (%)
100	79.20
150	80.93
200	81.21
250	83.79
300	84.60
350	83.64

V. DISCUSSION

The proposed MFCC and FBE based textural feature approach is compared with existing stuttering event classification algorithms. Table VI depicts comparison of proposed method with already existing state-of- the art methods using different features and classification accuracy rates. It can be seen that the proposed technique performs better compared to all the traditional algorithms.

TABLE VI. COMPARISON OF PROPOSED METHOD WITH EXISTING METHODS USING DIFFERENT FEATURES AND CLASSIFICATION RATE

Method	Features	Classification Accuracy (%)
[4]	Acoustic and pitch	87.39
[6]	LH-MFCC	94.98
[8]	MFCC-DTW	90
[9]	MFCC-DTW	89
[14]	MFCC, LPC	95
[15]	MFCC, LPC, PLP	96
[21]	MFCC, LPCC	94
Proposed	MFCC, FBE	96.36

The prime objective of this article is to present a new statistical feature approach based on MFCC and FBE matrix to enhance stuttering event classification using UCLASS database. Prolongation and repetition event are distinguished using GLCM and GLRLM features extracted from MFCC and FBE matrix and Extreme learning machine classification. Laplacian score-based feature selection algorithm is employed to remove irrelevant features resulting in improvement in the classification accuracy rate. Experiments show that GLRLM outperforms GLCM descriptors during the classification stage. On selecting the best feature set of 25 features (T = 0.2), highest accuracy of 96.36% is obtained. And it can be observed from the results and tables that the performance of the proposed algorithm is better compared to other existing methods. Besides, this article also emphasizes the use of feature selection technique to reduce the computational complexity of the algorithm.

VI. CONCLUSION

This article presents stuttering event classification approach based on MFCC and FBE using UCLASS database. Prolongation and repetition event are discriminated using GLCM and GLRLM features extracted from MFCC and FBE matrix and ELM classification. Laplacian score-based feature selection algorithm is employed to discard irrelevant features resulting in improvement in the classification rate. Experimental results show that, GLRLM outperforms GLCM descriptors during the classification stage. After selecting best feature set of 25 features (T = 0.2), highest accuracy of 96.36% is achieved. In future works, experiments can be performed with large speech samples with different feature extraction approaches and classifiers to improve the classification rate further.

REFERENCES

[1] "Statistics on Voice, Speech, and Language NIDCD" https://www.nidcd.nih.gov/health/statistics/statistics-voice-speech-and-language, Accessed: 2020-12-10.

[2] W. Suszynski, W. Kuniszyk-Jzkowiak, E. Smoka, and M. Dzienkowski, "Speech disfluency detection with the correlative method," Annales UMCS Informatica, vol. 3, no. 1, pp. 131 – 138, 2005.

[3] P. Howell, S. Davis, and J. Bartrip, "The University College London archive of stuttered speech (UCLASS)," Journal of Speech, Language, and Hearing Research, vol. 52, no. 2, pp. 556 – 568, 2009.

[4] P. S. Savin, P. B. Ramteke, and S. G. Koolagudi, "Recognition of repetition and prolongation in stuttered speech using ANN," in Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics, A. Nagar, D. P. Mohapatra, and N. Chaki, Eds. Springer India, 2016, pp. 65–71.

[5] N. K. A. M. Rashid, S. A. Alim, N. N. W. N. Hashim, and W. Sediono, "Receiver operating characteristics measure for the recognition of stuttering dysfluencies using line spectral frequencies," International Islamic University Malaysia Engineering Journal, vol. 18, no. 1, pp. 193–200, 2017.

[6] P. Mahesha and D. S. Vinod, "LP-Hillbert transform based MFCC for effective discrimination of stuttering dysfluencies", in 2017 International Conference on Wireless Communications, Signal Processing and Net-working (WiSPNET), March 2017, pp. 2561–2565.

[7] R. Hamzah, N. Jamil, and R. Roslan, "Development of acoustical feature-based classifier using decision fusion technique for malay language dis- fluencies classification", Indonesian Journal of Electrical Engineering and Computer Science, vol. 8, no. 1, pp. 262–267, 2017.

[8] P. B. Ramteke, S. G. Koolagudi, and F. Afroz, "Repetition detection in stuttered speech," in Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics, A. Nagar, D. P. Mohapatra, and N. Chaki, Eds. New Delhi: Springer India, 2016, pp. 611–617.

[9] P. Yeh, S. Yang, C. Yang, and M. Shieh, "Automatic recognition of repetitions in stuttered speech: Using end-point detection and dynamic time warping," Procedia - Social and Behavioral Sciences, vol. 193,

[10] p. 356, 2015, 10th Oxford Dysfluency Conference, ODC 2014, 17 - 20 July, 2014, Oxford, United Kingdom.

[11] A. G. Samah, A. Sherif, S. Mahmoud, and G. Nivin, "Classification of stuttering events using I-Vector," Egyptian Journal of Language Engineering, vol. 4, no. 1, pp. 11–18, 2017.

[12] I. Esmaili, N. J. Dabanloo, and M. Vali, "Automatic classification of speech dysfluencies in continuous speech based on similarity measures

and morphological image processing tools", Biomedical Signal Processing and Control, vol. 23, pp. 104 – 114, 2016.

[13] M. P. and V. D. S., "Support vector machine-based stuttering dysfluency classification using GMM supervectors," Int. J. Grid Util. Comput., vol. 6, no. 3/4, pp. 143–149, 2015.

[14] H. M., C. L. Sin, A. O. Chia, and Y. Sazali, "Gaussian mixture model-based classification of stuttering dysfluencies," Journal of Intelligent Systems, vol. 25, no. 3, pp. 387–399, 2015.

[15] K. Singh and A. K. Awasthi, "Comparison of speech parameterization techniques for the classification of speech disfluencies," Turkish Journal of Electrical Engineering and Computer Science, vol. 21, pp. 1983 – 1994, 2014.

[16] M. P. and D. S. Vinod, "Classification of speech dysfluencies using speech parameterization techniques and multiclass SVM," in Quality, Reliability, Security and Robustness in Heterogeneous Networks, Eds. Springer Berlin Heidelberg, 2013, pp. 298–308.

[17] J. L. C., P. Srikanta, and I. Nikhil, "Combining cepstral and prosodic features for classification of disfluencies in stuttered speech," in Intelligent Computing, Communication and Devices, Eds. Springer India, 2015, pp. 623–633.

[18] O. Stacey, M. Ricard, and R. Frank, "Automatic dysfluency detection in dysarthric speech using deep belief networks," in Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies. Association for Computational Linguistics, 2015, pp. 60–64.

[19] J. Palfy, "Analysis of dysfluencies by computational intelligence," Information Sciences and Technologies-Bulletin of the ACM Slovakia, vol. 6, no. 2, pp. 45–58, 2014.

[20] M. Hariharan, V. Vijean, C. Y. Fook, and S. Yaacob, "Speech stuttering assessment using sample entropy and least square support vector machine," in 2012 IEEE 8th International Colloquium on Signal Processing and its Applications, March 2012, pp. 240–245.

[21] M. Hariharan, C. Fook, R. Sindhu, A. H. Adom, and S. Yaacob, "Objective evaluation of speech dysfluencies using wavelet packet transform with sample entropy," Digital Signal Processing, vol. 23, no. 3, pp. 952– 959, 2013.

[22] L. S. Chee, O. C. Ai, M. Hariharan, and S. Yaacob, "Automatic detection of prolongations and repetitions using LPCC," in 2009 International Conference for Technical Postgraduates (TECHPOS), Dec 2009, pp. 1– 4.

[23] H. M., C. L. Sin, A. O. Chia, and Y. Sazali, "Classification of speech dysfluencies using LPC based parameterization techniques," J. Med. Syst., vol. 36, no. 3, pp. 1821–1830, 2012.

[24] L. S. Chee, O. C. Ai, M. Hariharan, and S. Yaacob, "Automatic detection of prolongations and repetitions using LPCC," in 2009 International Conference for Technical Postgraduates (TECHPOS), Dec 2009, pp. 1– 4.

[25] wietlicka Izabela, K.-J. Wiesawa, and S. Elbieta, "The application of Kohonen and multilayer perceptron networks in the speech non-fluency analysis," Archives of Acoustics, vol. 31, 01 2006.

[26] A. Benba, A. Jilbab, A. Hammouch, and S. Sandabad, "Voiceprint's analysis using MFCC and SVM for detecting patients with Parkinson's disease, "in 2015 International Conference on Electrical and Information Technologies (ICEIT), March 2015, pp. 300–304.

[27] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. OShaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors," Speech Communication, vol. 55, no. 2, pp. 237 – 251, 2013.

[28] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, "Low-variance multitaper MFCC features: A case study in robust speaker verification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 7, pp. 1990–2001, Sept 2012.

[29] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1085–1095, May 2012.

[30] M. Sahidullah and G. Saha, "A novel windowing technique for efficient computation of MFCC for speaker recognition," IEEE Signal Processing Letters, vol. 20, no. 2, pp. 149–152, Feb 2013.

[31] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 6, pp. 610–621, Nov 1973.

[32] L. Soh and C. Tsatsoulis, "Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices," IEEE Transactions on Geoscience and Remote Sensing, vol. 37, no. 2, pp. 780–795, March 1999.

[33] M. Galloway, "Texture analysis using gray level run lengths," Computer Graphics and Image Processing, vol. 4, no. 2, pp. 172–179, June 1975.

[34] X. Tang, "Texture information in run-length matrices," IEEE Transactions on Image Processing, vol. 7, no. 11, pp. 1602–1609, Nov 1998.

[35] B. V. Dasarathy and E. B. Holder, "Image characterizations based on joint gray level run length distributions," Pattern Recognition Letters, vol. 12, no. 8, pp. 497–502, August 1991.

[36] A. Chu, C. M. Sehgal, and J. F. Greenleaf, "Use of gray value distribution of run lengths for texture analysis," Pattern Recognition Letters, vol. 11, no. 6, pp. 415–419, June 1990.

[37] L. Zhu, L. Miao, and D. Zhang, "Iterative Laplacian score for feature selection," in Pattern Recognition, C.-L. Liu, C. Zhang, and L. Wang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 80–87.

[38] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in Advances in Neural Information Processing Systems 18, Y. Weiss, B. Schlkopf, and J. C. Platt, Eds. MIT Press, 2006, pp. 507–514.

[39] A. Bequ and S. Lessmann, "Extreme learning machines for credit scoring: An empirical evaluation," Expert Systems with Applications, vol. 86, pp. 42 – 53, 2017.

[40] Z. Zhou, Y. Song, Z. Zhu, and D. Yang, "Scene categorization based on compact SPM and ensemble of extreme learning machines," Optik vol. 6, no. 2, pp. 45–58, 2014.

[41] O. C. Ai, M. Hariharan, S. Yaacob, and L. S. Chee, "Classification of speech dysfluencies with MFCC and LPCC features," Expert Systems with Applications, vol. 39, no. 2, pp. 2157 – 2165, 2012.

[42] Gajanan K. Birajdar, Vijay H. Mankar "Passive Image Manipulation Detection Using Wavelet Transform and Support Vector Machine Classifier", Proceedings of International Conference on ICT for Sustainable Development pp 447-455, 2016.

[43] Gajanan K. Birajdar & Mukesh D. Patil, "Speech and music classification using spectrogram based statistical descriptors and extreme learning machine", Multimedia Tools and Applications vVolume78, pages15141–15168 (2019).