International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-12 Issue-4, April 2023

Hybrid Approach to Detect Prolonged Speech Segments

K B Drakshayini, Anusuya M A

Abstract: ***In the last 10 decades various methods have been introduced to detect prolonged speech segments automatically for stuttered speech signals. However less attention has been paid by researches in the detection of prolongation disorder at the parametric level. The aim of this study is to propose a hybrid approach to detect the prolonged speech segments by combining various spectral parameters with their recognition accuracies for the reconstructed speech signal. The paper presents prolonged segments detection by considering the parameters individually, combining various spectral parameters, validation of prolongation detection system, MFCC feature extraction process, basic model accuracies for the reconstructed signals. The proposed methods are simulated and experimented on UCLASS derived dataset. Obtained results are compared with the existing works of prolongation detection at parametric and word level. It is observed that hybrid parameters yield 92% of recognition rate for larger frame sizes of 200ms when modeled with SVM. The results are also tabulated and discussed for various metrics like sensitivity, specificity and accuracy metrics in detecting the prolonged segments. The study also focuses on the prolongation characteristics of vocalized and non-vocalized sounds at phoneme level. The detection accuracy of 71% is observed for Vocalized prolonged vowel phonemes over non-vocalized prolonged signal. Objectives: The objective of this work is to propose a hybrid algorithm to detect prolonged segments automatically for speech signal with prolongation disorder. The other objective is to evaluate the obtained spectral parameters performances by applying to various evaluation metrics and models to compute the recognition accuracy of a reconstructed signal. The objective is further extended to bring out the importance of variable frame size concept and to analyze the variations in vocalized and non-vocalized sounds. Methods: The methods adopted to detect prolonged speech segments are discussed at two levels namely at the preprocessing and modeling levels. The Preprocessing level is discussed by applying various parameters at an individual level, hybrid level by combing the Centroid, Entropy, Energy, ZCR parameters and MFCC feature extraction method. A new method has been applied using Specificity, Sensitivity and accuracy metrics to validate the prolongation detection model performance. In modeling level, the above parameters are discussed by applying evaluation metrics for the clustering and classification models like K-means, FCM and SVM. The performance of these methods is considered for evaluating and estimating the prolonged segment detection accuracy of the reconstructed speech signals of vocalized and non-vocalized sounds. All these methods are discussed in detail in the following sections. Findings: Hybridizing the spectral parameters to detect the prolonged speech segment automatically is a major finding of this work. It is also found that Specificity, sensitivity and accuracy metrics plays a major role in designing and validating the prolongation detection model. From the further experiments it is identified that the hybrid and verification metrics suits better for vocalized and non-vocalized sounds when larger frame lengths are considered. SVM has been found to perform better for all the above considerations. Novelty: As per Literature survey it is observed that individual and few parameters are applied to detect the prolongation. But works are not addressed on applying or combining more than two parameters to detect the prolonged speech segments. The novelty of this work lies in selecting and combining the spectral parameters at the preprocessing stage to detect the prolongation disorder. Spectral centroid and entropy are considered as appropriate parameters along with ZCR and Energy parameters. Hence hybridizing these parameters results in a novelty to propose an automatic prolongation detection system. Novelty is further brought by applying Specificity, sensitivity and accuracy metrics to build and evaluate the detection system for vocalized and non-vocalized prolonged sounds.***

Keywords: Prolongation, Centroid, Entropy, Specificity, sensitivity, Autocorrelation, frame length, threshold.

I. INTRODUCTION

Speech is the most common mode for communication to express ideas, feelings and thoughts [1]. There exists 2 to 5 % of population suffering from speech dysfluency. Stuttering is a break of normal speech such as repetition, prolongation and interjections of phonemes, sounds, phrases or word. It is required to identify the dysfluent part of speech automatically to enhance the process of speech signal analysis [2]. Few authors have worked by considering individual parameters for different datasets. In this Paper a hybrid method using Energy, Centroid, ZCR and Entropy is discussed to detect the prolonged dysfluent speech automatically. It also focuses on the performances of vocalized (for example: rrrrunning, aaaaple) and non-vocalized (for example: sssseven, ffffourteen) prolonged sound evaluated through various validation metrics. Recognition rate using various models are tabulated with their recognition performance. In the first level of preprocessing, prolonged segments are detected and removed. This signal is reconstructed to extract the MFCC coefficients and these features are modeled using various decision-making methods like KMeans, FCM and SVM are used to compute the recognition performance. These simulation results are analyzed and presented for various evaluation metrics at the parametric and the model levels. Section 1 and section 2 discusses about related literature and the methodology adopted.

Manuscript received on 29 March 2023 | Revised Manuscript received on 06 April 2023 | Manuscript Accepted on 15 April 2023 | Manuscript published on 30 April 2023.

*****Correspondence Author(s)
K B Drakshayini*, Research Scholar, Visvesvaraya Technological University (VTU), Belgaum (Karnataka), India. E-mail: drakshakb@gmail.com, ORCID ID: https://orcid.org/0000-0002-4860-2890
Dr. Anusuya M A, Associate Professor, Department of Computer Science and Engineering, JSS Science and Technology University, Mysore (Karnataka), India. Email: anusuya_ma@sjce.ac.in

Challenges and dataset used for simulation is presented in Section 3 and section 4. Section 5 discusses the results and observations. Conclusion and future enhancements are discussed in section 6.

1.1. Related literature

Table 1 depicts the related work available in the literature in brief using spectral parameters to detect the prolonged speech segments in a disfluent speech.

Table 1: Existing work on parameter-based prolongation detection process

Author	Database	Parameters	Decision making	Detection rate at parametric level	Recognition rate
Deshmukh [3]	UCLASS	Energy, pitch, duration, frequency formants	–	–	–
Katarzyna Barczewska[3]	6 recordings	First two formants (F1, F2), segment duration	–	–	68%
G. Manjula [4]	20 speakers (AIISH)	Epochs, Zero Frequency Filter	–	–	–
Sadeen Alharbi [4]	40 speakers (UCLASS)	ZCR and Short-term energy	–	70%	–
Waldemar Suszyński[5]	10 speakers	Frequency and duration	Fuzzy classifier	91%	–

II. METHODOLOGY

Figure 1 depicts the complete architecture of the proposed system. The architecture is considered at three phases. The First phase is aimed at preprocessing the signal to reduce the noise and to detect the prolonged segments automatically by using the proposed hybrid parameterization technique. In the second phase the reconstructed signal is further processed to extract first level 12 speech features using MFCC procedure. In the last phase these features are modeled using various models to compute speech recognition accuracies.

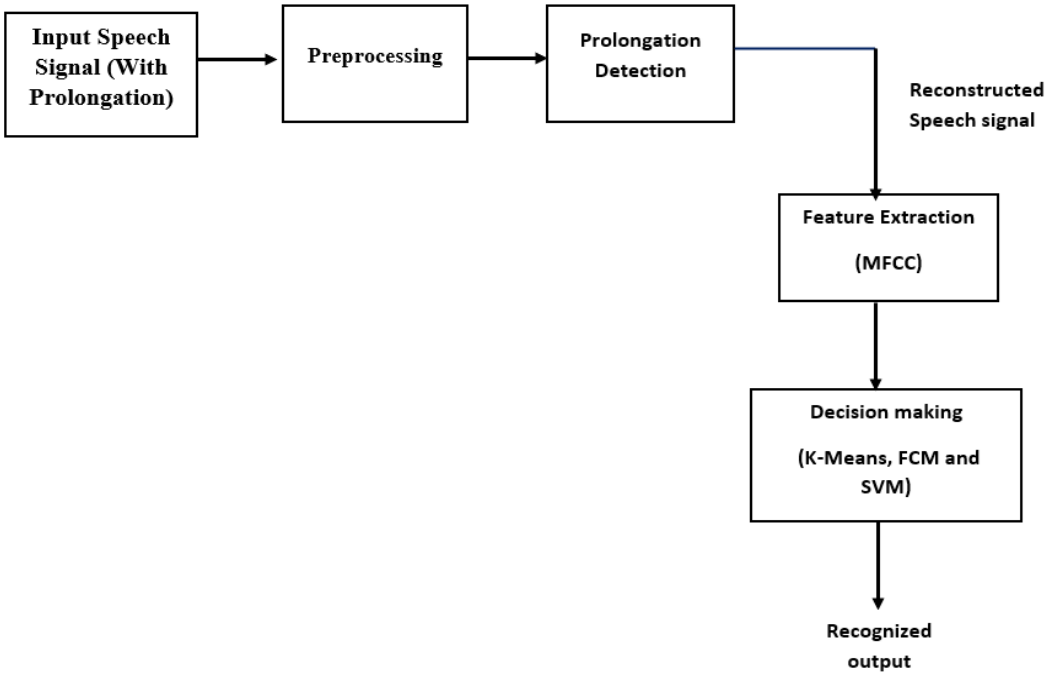

Figure 1: Stuttered Speech recognition Process

2.1. Phase I: Prolongation detection

Phase 1 consists of preprocessing prolongation detection and reconstruction of the signal.

2.1.1 Preprocessing:

The original sampling frequency of the input speech samples of UCLASS dataset is 44.1 kHz, speech samples are down sampled to 16 kHz for the simulation purpose. In the digitization process the speech signals are normalized to prevent the error estimation caused by change in speaker's volume.

2.1.2 Parameters considered for prolongation detection

To detect the prolonged speech segments the following parameters are adopted and evaluated individually. These Parameters are also combined to observe the performance of hybrid effect in the prolongation detection.

A. Short term Energy

Let $x_i(n)$, n = 1…, N the speech samples of the i[th] frame, of length N. Eqn. (1) used for computing short term energy of an individual frame. The stress on a particular phoneme has higher energy associated with it [6]. Hence short-term energy is one of the parameters to detect prolongation in stuttered speech signal.

$$E_i = \frac{1}{N}\sum_{n=1}^{N} |x_i(n)|^2 \text{----- Eqn. (1)}$$

B. Spectral centroid

The spectral centroid represents centre of gravity of speech spectrum and it is highly variated for speech segments [6]. It also indicates the centre of mass spectrum and depicts the brightness of sound. It is calculated using Eqn. (2)

$$C_i = \frac{\sum_{k=1}^{N}(k+1)x_i(k)}{\sum_{k=1}^{N}x_i(k)}.x_i(k) \text{-----Eqn. (2)}$$

k = 1..., N, are the Discrete Fourier Transform (DFT) coefficients of the i[th] frame, where N is the frame length.

C.Zero Crossing rate (ZCR)

ZCR is a measure of smoothness of a speech signal and represents frequency at which the power is focused in the signal [7]. Zero crossing rate is computed using Eqn. (3)

$$Z_n = \sum_{\infty}^{-\infty} |\,\text{sgn}[x(m)] - \text{sgn}[x(m-1)]|\, w(n-m) \text{ -----Eqn (3)}$$

Where

$$\text{sgn}[x(n)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases} \text{-----Eqn (3a)}$$

and

$$w(n) = \begin{cases} \frac{1}{2N} & for\ 0 \leq n \leq N-1 \\ 0 & for, otherwise \end{cases} \text{-----Eqn (3b)}$$

N is the window duration.

D.Spectral Entropy

Spectral Entropy calculates the regularity of power spectrum of speech signal and has peak capturing property [8]. Distinct peaks and the position of the peaks in the spectra are dependent on the phoneme under consideration. It is computed using Eqn (4).

$$Entropy = \frac{-\sum_{k=b_1}^{b_2} s_k \log(s_k)}{\log(b_2 - b_1)} \text{-----Eqn (4)}$$

where

- s_k is the spectral value at frame *k*.
- *b*1 and *b*2 are the band edges, over which spectral entropy is calculated.

Similarity measure: Autocorrelation

Autocorrelation is a measure of similarity between the successive frames. Highly correlated frames are considered as prolonged frames. To differentiate between prolonged and un-prolonged speech segments autocorrelation similarity measure is applied over adjacent frames of the signal [9]. Autocorrelation function is computed by Eqn (5).

The Autocorrelation between the samples of X= {x_1, x_2, …,x_n} and Y= {y_1, y_2, ….,y_n} is given by the correlation factor and it is given by

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - X')(y_i - Y')}{\sqrt{\sum_{i=1}^{n}(x_i - X')^2}\sqrt{\sum_{i=1}^{n}(y_i - Y')^2}} \text{-----Eqn (5)}$$

where X' and Y' are the mean values of X and Y respectively.

All the above parameters are computed for the total length of the signal on frame-by-frame basis. From our observation it is observed that these parameters play a major role in differentiating between prolonged and normal speech frames. The simulations results are discussed at two levels by considering parameters individually and by combining the parameters. The recognition accuracies for these two levels are calculated and discussed in section 5.

Observation from literature:

In the literature researchers have applied only short-term energy and ZCR parameter among the above discussed parameters to detect the prolonged segments, whereas spectral centroid and Entropy are not. But ZCR parameter is applied for dysarthric dysfluent speech but not for stuttered. In this work ZCR, Energy entropy and centroid parameters are applied together to detect the prolonged frame. To detect the prolonged segments the proposed algorithm is as follows:

Algorithm to detect and remove prolonged frames

Step 1: Read the prolonged stuttered speech signal
Step 2: Divide the signal into frames of 0.025ms and 200 ms duration.
Step 3: Compute Short term Energy, ZCR, Spectral Entropy and centroid parameters for each frame to the complete signal length.
Step 4: Autocorrelation function is applied to compute the similarity between the adjacent frames.
Step:5 Average autocorrelation values for each parameter are computed.
Step 6: Every parameter is identified and fixed with the threshold values.
Step7: If the autocorrelation value between the adjacent frames is greater than the frames are identified as prolonged speech frame and it is removed.
Step8: Retained frames in steps 7 are identified as un prolonged frames and used to reconstruct the speech signal.
Step9: steps1 to 7 are repeated by the individual spectral parameter threshold value on each frame at a
time to decide the segment is prolonged or not.

Figure 2 depicts the hybrid method by combining all the parameters to detect the prolongation detection.

Average Autocorrelation of these parameters are represented by the following terms. Short-term energy (E-ACF), Autocorrelation of ZCR (Z-ACF), Auto correlation of Entropy (En-ACF), Autocorrelation of centroid (C-ACF). These are used to compute the similarity measures in terms of spectral parameters. The average autocorrelation values of each parameter are considered as Threshold. Threshold values of the above-mentioned parameters are computed using autocorrelation function and these terms are defined as Autocorrelation threshold for Energy (TE), Autocorrelation threshold for ZCR (TZ), Autocorrelation threshold for Entropy (TEn), Autocorrelation threshold for Centroid (TC). Using the above-mentioned threshold values the decisions are made to retain or to remove the frames. Retained frames are combined to reconstruct the fluent speech signal.

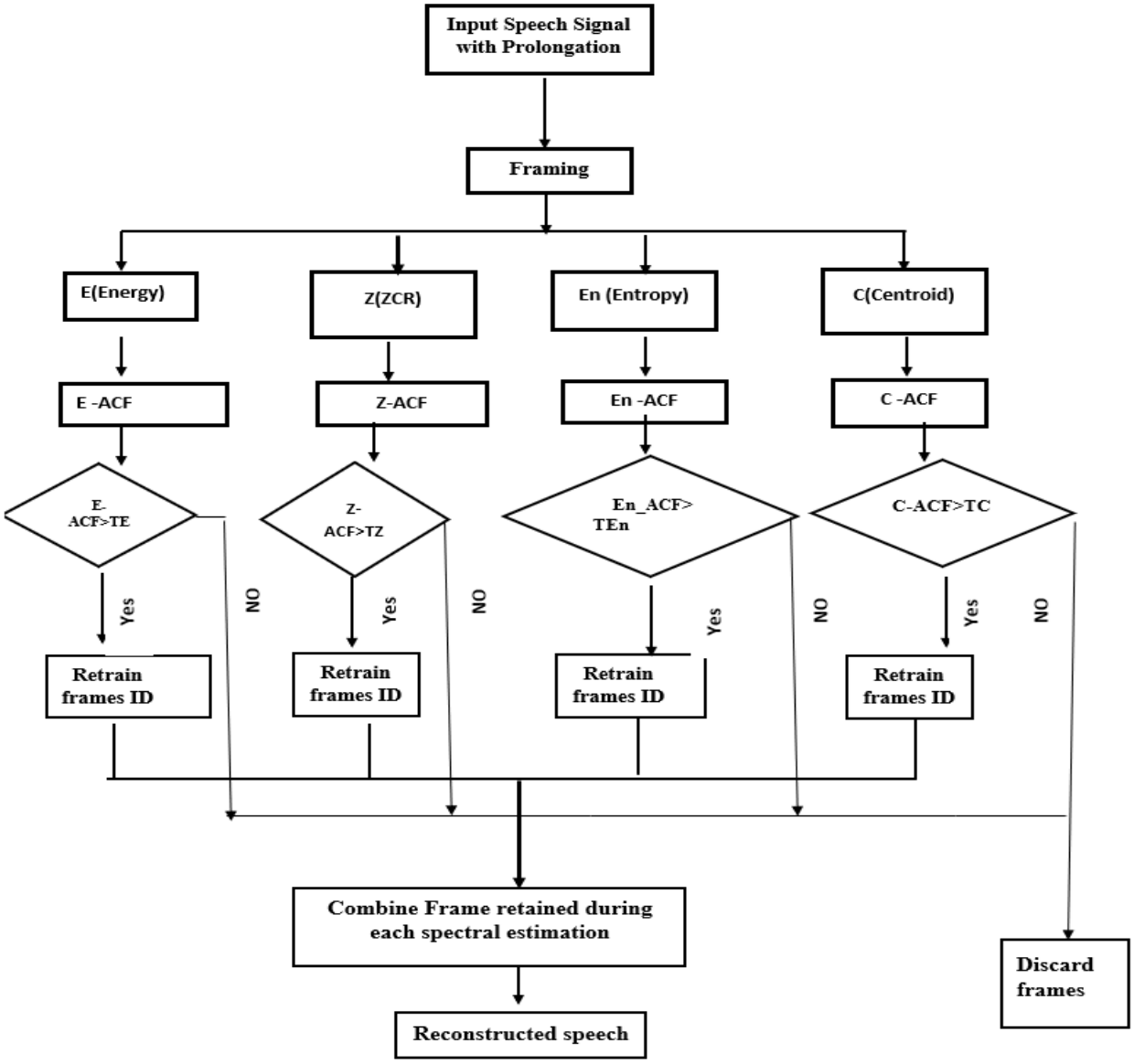

Figure 2: Prolongation Detection Process

2.2 Phase II: Feature Extraction using MFCC

The features of the reconstructed signal are computed by applying the MFCC feature extraction procedure. The first 12 coefficients of MFCC are computed for each frame. Below algorithm list the steps used to extract the speech parameters:
Algorithm for feature extraction is as follows
Step 1. Each of the speech sample is down sampled to 16 KHz to increase the quality of the signal and passed through high pass filter to remove the noise.
Step 2. The pre-emphasized signal is framed in to 200 ms and 0.025ms duration with an overlap of 10ms
Step 3. Hamming window is applied to smoothen the signal and Discrete cosine Transform is applied.
Step 4. Compute the Log filtered energies and first 12 coefficients of the signal using IDFT for all the frames.
Step 5. Average featured frames are computed.

Observations:

The advantage of considering larger frame size helps in better identification of the prolonged frames that majority checks all the parameters threshold values. It also reduces the number of frames to be processed with steady or low SNR regions. Larger frame size also helps in loosing less prolonged data.

Retrieval Number:100.1/ijeat.D41060412423
DOI: 10.35940/ijeat.D4106.0412423
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

2.3 Phase III: Modeling Phase/Decision:

The obtained features from the above process are modeled using K-Means, FCM and SVM classifiers to calculate the recognition accuracy of the reconstructed signal. The following section discusses the procedure to calculate the recognition rate for various models. This work considers two clustering methods and one classifier to compute the recognition performance of the reconstructed speech signal. The basic settings of the models for the simulation to compute the recognition accuracy is as discussed below.

2.3.1 K-means: K-means algorithm is used to generate a vector quantization codebook for data compression. It partitions 'N' observations into K clusters, in which each observation belongs to the cluster with the nearest mean. Each of clusters is defined by its central vector or centroid. Using Euclidian distance function, K-Means algorithm clusters the data in to K groups that assigns objects to their closest cluster [2].

Algorithm:

Step1.Set Observations (N=50) and clusters (K=5).
Step2. Randomly select 5 distinct centroids for each cluster.
Step3.Using Euclidian distance measure to find the distance between each observation point and the centroid.
Step4.Assigns each observation point to the nearest cluster
Step5.Calculate the mean of each cluster as new centroid
Step6.Repeat step 3 to step 5 with the new center of cluster until there is no change in the centroid.

2.3.2 FCM (Fuzzy C Means)

This algorithm works by assigning a membership value to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point [11]. The performance of the algorithm depends on membership value(U), and the fuzzy parameter(m), termination criteria, norm of the matrix,

Algorithm

1. Randomly initialize the clusters (K=5)
2. Create the distance matrix from a data point to each of the cluster center using Euclidean distance
3. The Membership matrix is computed using fuzzification parameter m= (2.0)
4. Values of the membership matrix should be less than or equal to one otherwise return to step2
5. Termination parameter is set as 0.0001

2.3.3 SVM (Support Vector Machine)

SVM is a powerful machine learning tool which attempts to obtain a good separating hyper-plane between two classes in the higher dimensional space. It is a predominant technique to estimate the basic parameters of speech. Speech sample can be approximated as a linear combination of speech sample by minimizing the sum of squared difference between the actual speech samples and predicted values. A unique set of parameters or predictor coefficients can be determined. These coefficients form the basis for linear prediction of speech samples. The importance of this method lies in its stability to provide extremely accurate estimates in understanding nonlinear phase characteristics of stuttered speech[10].

Algorithm:

Step 1. Best hyperplane is decided by considering the distance between data and exiting hyperplane.
Step 2. To make decision on deciding optimal hyperplane, polynomial kernel function with degree 4 is to decide optimal hyper plane with the hold out factor 0.15.
Step3.To classify frames into different classes multiclass classification is broken down into multiple binary classification. Classifier uses m (m=5) binary SVM classifiers.
Step 4: In each binary SVM classifier, member ship is predicted to make decision of belongingness.

III. CHALLENGES

Deciding frame length is a challenging task because it has to accommodate all parametric variations of phonemes and also it should not break the disorder part of the speech segment while framing. If the prolongation duration extends, it is difficult to analyze and append smaller frames in terms of prolonged and un-prolonged phonemes. It is also an equally challenging task to classify prolonged speech sound categories that are mixed with other types of disorders. Fixing the threshold values to each type of the sounds is tedious a task, that to when combined with nasal sounds in non-vocalized words.

IV. DATASET

The experiments are conducted on the UCLASS dataset. This repository consists monologs, readings, and conversational recordings [12]. For our simulation 80 samples at word level are derived from the sentences recorded in the repository. It includes 22words of female speakers and 58 words of male speakers with age ranging from 11 years to 20 years. The samples are chosen to cover speech samples of different age and gender. By perception, vocalized and nonvocalized prolonged words are identified and derived manually. A total of 80 speech recordings of prolongation with vocalized and nonvocalized sounds are collected and few examples are depicted below. Table 2 depicts example words considered for vocalized and non-vocalized prolonged sounds used in our simulations.

Table 2: Sample Data set considered for experiment

Actual word	Pronounced Pattern	Type of Prolonged sound
Ball	/b/aaaaa/ll	Vocalized
Came	/c/aaaaa/me	
climbed	/c/laaaaa/imbed	
Every	/eeee/very	
Moment	/mmm/oment	
Favourite	/fffff/avourite	Non-vocalized
Finding	/fffff/inding	
Fish	/ffffff/ish	
Step	/ssss/ta/eee/p	
Secondary	/ssss/econdary	

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

V. RESULTS AND DISCUSSIONS

This section discusses the simulation results for prolongation detection and recognition accuracies obtained for the prolonged stutter speech. The results are discussed at four levels as listed below:

i) Prolongation detection using individual parameters
ii)Prolongation detection by combining/hybridizing spectral parameters
iii)Performance evaluation of prolongation detection using validation metrics at parametric level
iv)Recognition accuracy of the reconstructed speech using models.

5.1 Prolongation detection using individual parameters

This section discusses the results and observations individually by considering all the spectral parameters to detecting the prolonged speech segments.

Short term energy:

The Short-term energy and autocorrelation between the adjacent frames are computed using Eqn (1) and Eqn (5). To distinguish between prolonged and un-prolonged frames the average threshold value (TE) is set as 1. 43e06db.If autocorrelation of frame energy value exceeds 1.43e^{06}db it is detected as prolonged else it is un-prolonged. In the Prolonged frames partly or completely the energy of two adjacent frame will be approximately same due to the prolongation of certain phoneme. The threshold value changes as the type of the phoneme and sound changes.
For Example, for the speech signal pronounced as name where phoneme 'a' is prolonged 'n/aaaa/me'. The experiments are conducted for two frame sizes by varying the frame length. For 200ms and 0.025 ms, total 4 and 17 number of prolonged frames were identified. For example, considering 200 ms, obtained total 10 frames, out of which one set of adjacent frames (7th and 8th frames) are identified as prolonged frames having short term energies as 1.17e^{03}db and 2.26e^{03}db respectively. The difference between autocorrelation of these two is computed as 4.8e$^{06.}$ Since this value is greater than the threshold value 1.43 e^{06}it is considered as prolonged frame. Figure 3a shows short term energy of a prolonged speech signal pronounced as "n/aaaa/me ".

Zero crossing rate

The ZCR and autocorrelation between the adjacent frame are computed using Eqn (2) and Eqn (5). To distinguish between prolonged and un-prolonged frames average threshold value (TZ) is set as 1. 5e04.If autocorrelation of Zero crossing rate value exceeds 1.5e^{04}it is detected as prolonged else it is un-prolonged. In the Prolonged frames partly or completely the ZCR of two adjacent frames will be approximately same due to the prolongation of certain phoneme. The threshold value changes as the type of the phoneme and sound changes. For Example, for the speech signal pronounced as name where phoneme 'a' is prolonged 'n/aaaa/me'. The experiments are conducted for two frame sizes by varying the frame length. For 200ms and 0.025 ms, total 4 and 17 number of prolonged frames were identified. For example, considering 200 ms, obtained total 10 frames, out of which one set of adjacent frames (7th and 8th frames) are identified as prolonged frames having 97and 271 number of zero crossings respectively. The difference between the autocorrelation of these two is computed as 6.4 e$^{04.}$ Since this value is greater than the threshold value 1.5e^{04} it is considered as prolonged frame. Figure 3b shows Zero crossing rate of a prolonged speech signal pronounced as "n/aaaa/me ".

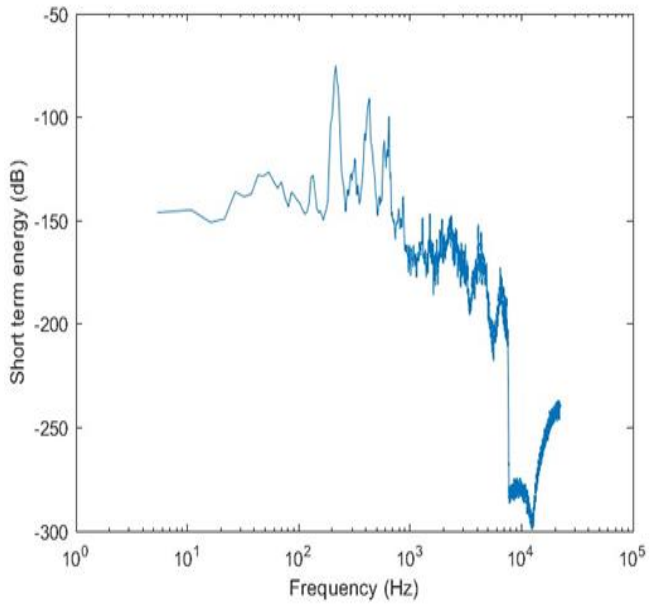

Figure 3a: Short term energy contour of a speech signal

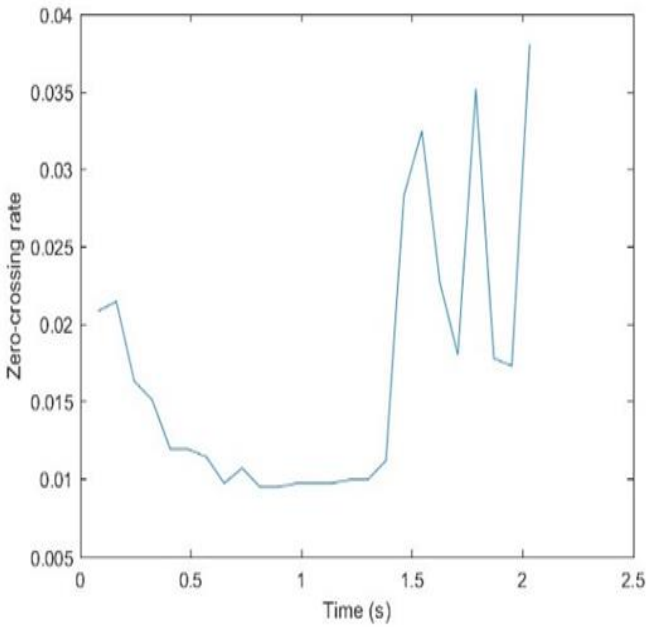

Figure 3b: Zero crossing rate of speech signal

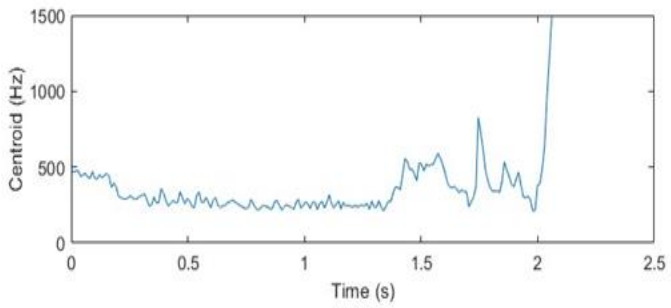

Figure 3c: Spectral Centroid of a Speech signal

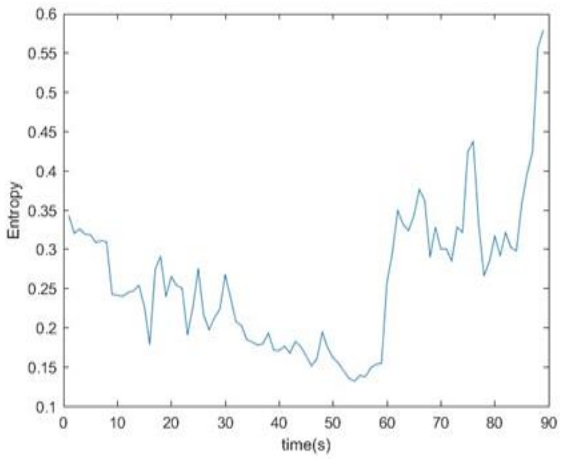

Figure 3d: Spectral Entropy of a speech signal

Spectral Centroid

The centroid and autocorrelation between the adjacent frame are computed using Eqn (3) and Eqn (5). To distinguish between prolonged and un-prolonged frames average threshold value (TC) is set as 0.0015 Hz. If autocorrelation of Spectral centroid value exceeds 0.0015 Hz it is detected as prolonged else it is un-prolonged. In the Prolonged frames partly or completely the Centroid of two adjacent frame will be approximately same due to the prolongation of certain phoneme. The threshold value changes as the type of the phoneme and sound changes.

For Example, for the speech signal pronounced as name where phoneme 'a' is prolonged 'n/aaaa/me'. The experiments are conducted for two frame sizes by varying the frame length. For 200ms and 0.025 ms, total 4 and 17 number of prolonged frames were identified. For example, considering 200 ms, obtained total 10 frames, out of which one set of adjacent frames (8^{th} and 9^{th} frames) are identified as prolonged frames having spectral centroid values 0.07 Hz and 0.05 Hz respectively. The difference between autocorrelation of these two is computed as 0.0059 Hz. Since this value is greater than the threshold value 0.0015Hz it is considered as prolonged frame. Figure 3c shows Spectral Centroid of a prolonged speech signal pronounced as "n/aaaa/me ".

Spectral entropy

The Spectral entropy (Shannon) and autocorrelation between the adjacent frame are computed using Eqn (4) and Eqn (5). To distinguish between prolonged and un-prolonged frames the average threshold value (TEn) is set as $1.1e^{05}$Hz. If autocorrelation of Spectral entropy value exceeds $1.1e^{05}$Hz it is detected as prolonged else it is un-prolonged. In the Prolonged frames partly or completely the entropy of two adjacent frame will be approximately same due to the prolongation of certain phoneme. The threshold value changes as the type of the phoneme and sound changes.

For Example, for the speech signal pronounced as name where phoneme 'a' is prolonged 'n/aaaa/me'. The experiments are conducted for two frame sizes by varying the frame length. For 200ms and 0.025 ms, total 4 and 17 number of prolonged frames were identified. For example, considering 200 ms, obtained total 10 frames, out of which one set of adjacent frames (5^{th} and 6^{th} frames) are identified as prolonged frames having spectral entropy values 0.07 Hz and 0.05 Hz respectively. The difference between autocorrelation of these two is computed as 2.4 e^{05} Hz. Since this value is greater than the threshold value 1.1 e^{05}Hz it is considered as prolonged frame. Figure 3d shows Spectral Centroid of a prolonged speech signal pronounced as "n/aaaa/me ".

Observations:

Sometimes if the signal has been shortly pronounced with the minimum of one or two frames identifies results in negative Entropy values, which is not considered. The observed result for prolongation detection with respect to individual parameters is tabulated in Table 3.

5. 2. Prolongation detection using hybrid method

From the above discussions it is realized that all the parameters are required and plays a major role in detecting the prolongation at frame level in one or the other way. Hence work motivated to observe the combined effect of these parameters, for a frame length of 200ms. Since larger frame size can accommodate all the variations of the parameters the major discussions are done for the 200ms frame duration. To present the performances difference between smaller and larger frame results are tabulated in table 3 for both the size of frames for 200 ms and 0.025ms.

From the above experimental observation, it is clear that all these parameters play a major role in identifying strength of the signal hence work motivated to combine all these parameters and to propose a new hybrid prolongation detection algorithm. The above parameters are applied for variable frame length analysis to increase the prolongation detection rate. The results are tabulated prolongation detection rates for 200 ms and 0.025 ms frame duration in Table3 and analysis is shown in Figure 4.

Table 3: Prolongation detection process results (with frame duration 0.025msec and 200 msec for speech signal name.wav pronounced as 'n/aaaa/me')

Parameter	No of proper frames detected	No of prolonged frames detected	% of detection	No of frames detected as Proper frames	No of frames detected as prolonged	% of detection
Frame duration	0.025msec (Total 82 frames)			200ms (Total 10 Frames)		
Energy	65	17	20.73	6	4	40
ZCR	72	10	12.19	5	5	50
Entropy	56	26	31.70	5	5	50
Centroid	61	21	25.60	4	6	60
Proposed Hybrid method	40	42	51.21	3	7	70

From the above table it is clear that hybrid method yields better prolongation detection rate than individual parameters for larger frame size. This improvement is because, it considers the frames that are left out with individual parameter due to smaller frame size. The threshold values that lie on the frame boarder is not neglected due to the larger frame sizes during framing which can happen in smaller size. This fails to locate prolonged frames when individual parameters are used. But the above variations are overlooked in the hybrid approach.

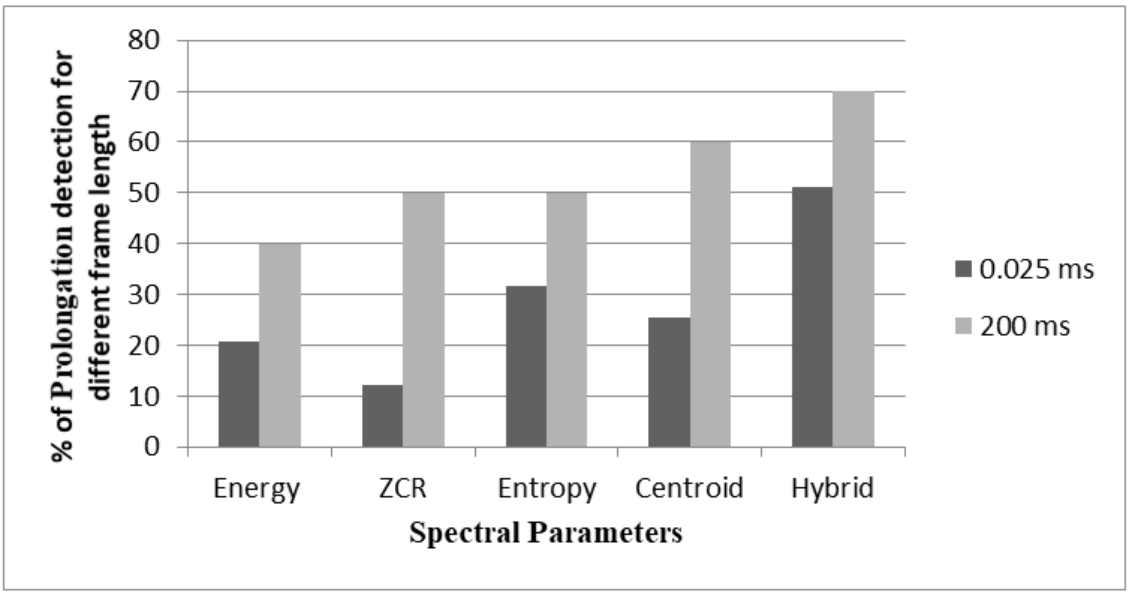

Figure 4: Analysis of proposed prolongation detection method with variable frame length for individual parameters and hybrid method

5.3. Performance Evaluation of Prolongation Detection Using Validation Metrics

Further to ensure these selected spectral features are of high prominence, the verification and validation of these spectral features performance is evaluated by considering Specificity, accuracy and sensitivity parameters as defined as follows:

This section presents the measures to evaluate the appropriate selection of the spectral parameters in detecting the prolonged frames in terms of sensitivity, specificity, and accuracy measures. As per the Table 4 the hybrid method performs well. The performance metrics are computed using Equations 6 ,7 and 8. These metrics are defined in terms of TP (true positive), TN (True negative), FP (false positive) and FN (False Negative). The validation results are tabulated in table 4 and Analysis is shown in Figure 6.

Specificity: The specificity of a prolongation detection system is to determine the proper speech frames correctly. Mathematically, this can be stated as follows.

$$\text{Specificity} = \frac{TN}{TN+FP} \text{----Eqn (6)}$$

Sensitivity: The sensitivity of a prolongation detection system is to determine the prolonged speech frames correctly. Mathematically, this can be stated as follows.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \text{----Eqn (7)}$$

Accuracy: The Accuracy of a prolongation detection system is to differentiate the prolonged speech frames and proper frames correctly. Mathematically, this can be stated as follows.

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+TP+FN} \text{----Eqn (8)}$$

Where

- True positive (TP): Prolonged frames correctly identified as Prolonged frames
- False positive (FP): Proper frames incorrectly identified as Prolonged frames

- True negative (TN): proper frames correctly identified as Proper frames
- False negative (FN): Prolonged frames incorrectly identified as Proper frames

Table 4: Comparison of Performance evaluation metrics for different spectral parameters during prolongation detection (Frame duration: 200ms, 10 frames)

Parameter	TP	TN	FP	FN	Specificity	Sensitivity	Accuracy
Energy	4	6	3	4	66	50	58
ZCR	5	5	3	3	62	62	62
Entropy	5	5	3	3	62	62	62
Centroid	6	4	3	3	57	66	62
Proposed Hybrid method	7	3	1	3	75	70	71

Observations:
From the above table the prolongation detection model has accuracy measure as a major metric to differentiate between prolonged and un-prolonged frames. As an extension, interested observations were made for vocalized and non-vocalized sounds by taking accuracy metric to detect the prolonged speech segment. Sample vocalized and non-vocalized sounds analysis for prolongation detection and correction is as shown in Figure 5(a)-5(d) and Table 5.

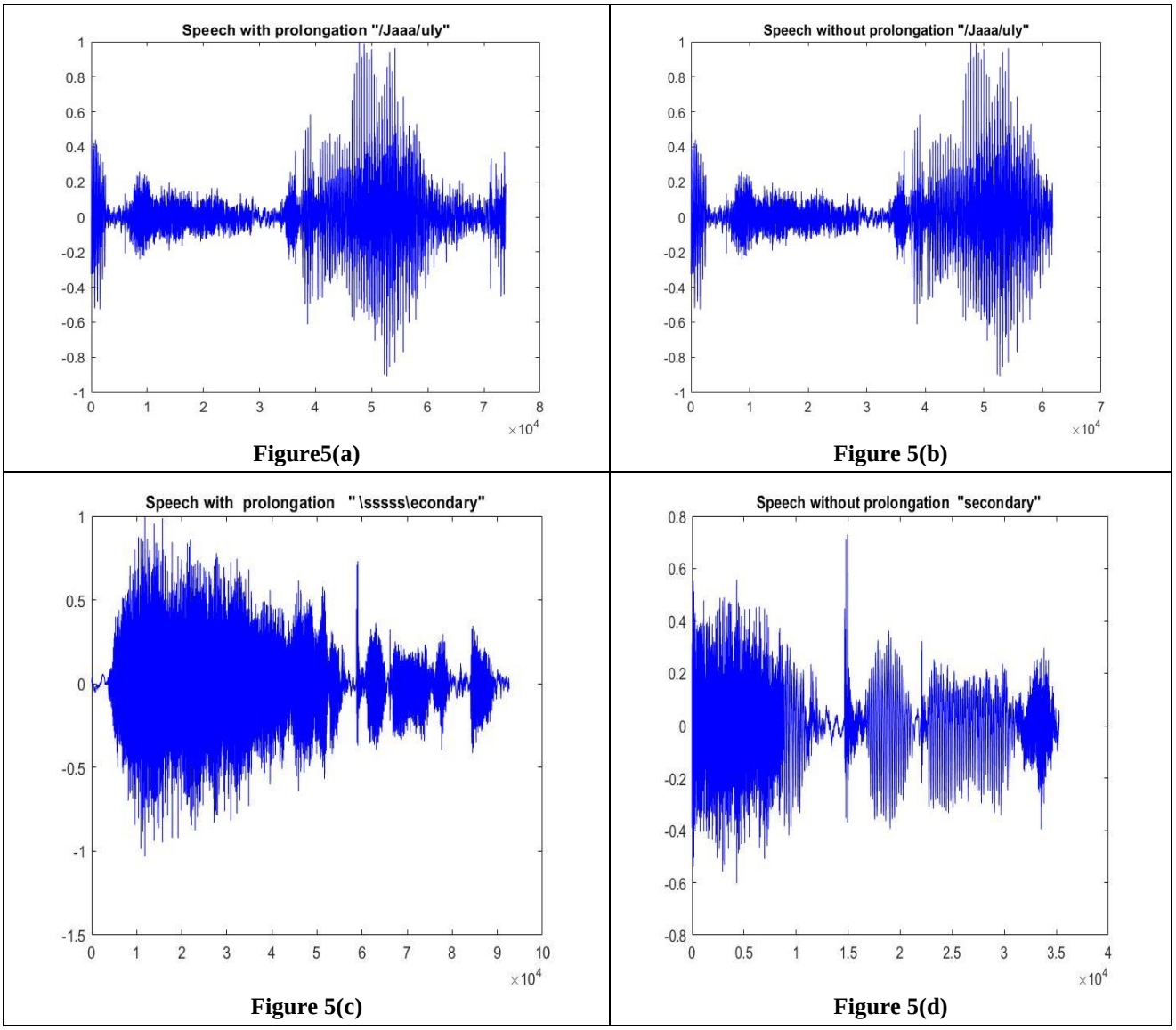

Figure5(a) **Figure 5(b)**

Figure 5(c) **Figure 5(d)**

Figure 5: Vocalized and non-vocalized prolonged sounds before and after removal of prolongation.

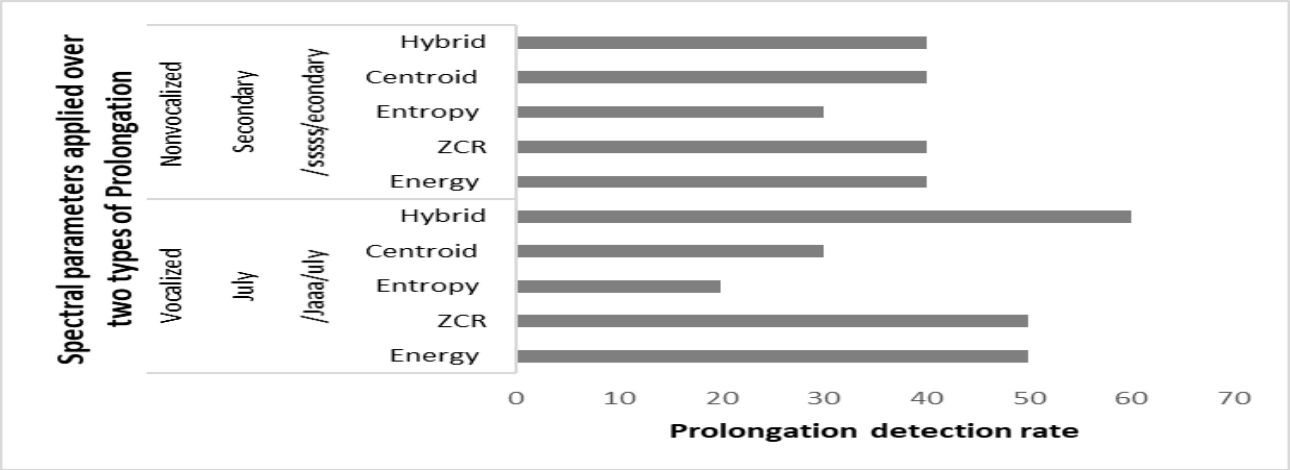

Figure 6: Analysis of proposed prolongation detection method for vocalized and nonvocalized prolongation (Frame duration 200 ms)

Observations:

In detecting vocalized prolonged sounds vowel sounds are clearly detected with major contributions of ZCR and Energy parameters. In prolongation of vocalized sound 'Jaaa/uly '(Actual word: July) 'a' is detected as prolonged phoneme. During detection process prolonged speech sample is divided in to 10 frames with frame duration of 200 msec. Out of 10 frames,6 frames are identified as prolonged frames.

In detecting non-vocalized prolonged sounds, intersection of voiceless consonants in between the phonemes could not find regularity in power spectrum due to the poor prolongation. In prolongation of non-vocalized sound 'sssss/econdary '(Actual word: secondary) 's' is detected as prolonged frame in non-vocalized prolongation audio sample (Ex: 'sssss/econdary') prolongation detection process prolonged speech sample is divided in to 10 frames with frame duration 200 msec. Out of 10 frames, 4 frames are detected as prolonged frames by combining results of all 4 spectral parameters.

5.4. Performance of Stuttered Speech Recognition System:

The signal is reconstructed by removing the prolonged frames through hybrid procedure. Features are extracted using MFCC procedure. First 12 coefficients are extracted and modeled by applying K-means, FCM and SVM classifiers. Table 5 depicts the recognition accuracies of all models and analysis is shown in Figure 7. Since the hybrid method was well suited to vocalized sounds, the simulations were performed for vocalized sounds. Among the above models SVM has obtained 92% of recognition accuracies for the reconstructed signal. From this it is clear that hybrid parameters contribute towards the detection, correction and recognition of vocalized sounds than non-vocalized sounds in a prolonged stuttered speech.

Table 5: Effect of prolongation detection method in stuttered speech recognition

Parameter used for Prolongation detection		Energy	ZCR	Entropy	Centroid	Proposed Hybrid
Feature Extraction		MFCC				
Decision making	K-Means	65%	68%	65%	70%	80%
	FCM	68%	70%	68%	72%	85%
	SVM	70%	72%	70%	75%	92%

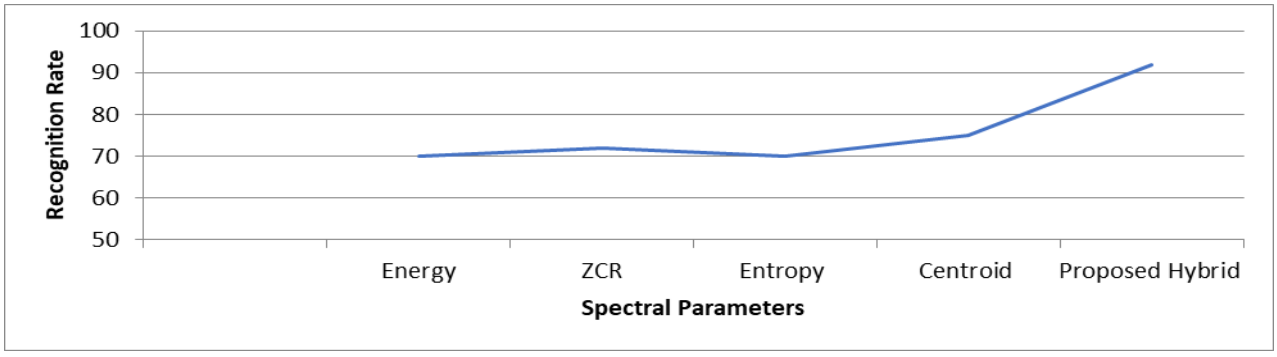

Figure 7: Analysis of effect of proposed prolongation detection method for stuttered speech recognition model

Retrieval Number:100.1/ijeat.D41060412423
DOI: 10.35940/ijeat.D4106.0412423
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

VI. CONCLUSIONS AND FUTURE ENHANCEMENT

This work proposes a hybrid approach to detect and recognize prolonged speech segments using spectral parameters. The prolongation detection is discussed for variable frame lengths of vocalized and non-vocalized prolonged sound. It is observed that 71% of prolongation detection and 92% of recognition accuracy is identified for the proposed hybrid approach over the individual parameter method. It is also observed that the hybrid approach is less suitable for non-vocalized sounds. Hence the approach can be further enhanced to increase its efficiency for non-vocalized sounds by defining various threshold measures for different phoneme sounds and by identifying few more parameters by trial-and-error methods. Further it can be tried by selecting different entropy parameters over non-vocalized sounds. The work can be further extended to phrase level to detect the prolongations at word level by increasing and considering the data set for various age and gender parameter.

DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	All authors have equal participation in this article

REFERENCES

1. Dr.MA Anusuya, SK Katti, Front end analysis of speech recognition: a review", International Journal of Speech Technology,2011, Available from: DOI:10.1007/s10772-010-9088-7
2. K B Drakshayini, Anusuya M A," Stop gap removal using spectral parameters for stuttered speech signal", International Journal of Advanced Trends in Computer Science and Engineering ,2021 Available fromhttps://doi.org/10.30534/ijatcse/2021/521032021
3. Om Dadaji Deshmukh, Suraj Satishkumar Sheth, Ashish Verma," Reconstruction of a smooth speech signal from a stuttered Speech",2013 Available from: https://patentimages.storage.googleapis.com/69/64/27/5aef3d5d69024c/US8600758.pdf
4. Katarzyna Barczewska, Magdalena Igras-Cybulska, "Detection of disfluencies in speech signal",2013 Available from: https://www.researchgate.net/publication/261913703_Detection_of_disfluencies_in_speech_signal
5. G. Manjula, M. Shiva Kumar, "Identification and Validation of Repetitions/Prolongations in Stuttering Speech using Epoch Features",2017 Available from: https://www.ripublication.com/ijaer17/ijaerv12n22_29.pdf
6. Sadeen Alharbi, Madina Hasan, Anthony J H Simons, Shelagh Brumfitt , Phil Green,"A Lightly Supervised Approach to Detect Stuttering in Children's Speech",2018, Available from: https://www.isca-speech.org/archive/pdfs/interspeech_2018/alharbi18_interspeech.pdf Available from: DOI:10.1016/j.procs.2020.04.146
7. Waldemar Suszyńskia, Wiesława Kuniszyk-Jóźkowiaka, Elżbieta Smołkaa, Mariusz Dzieńkowskib "Prolongation detection with application of fuzzy logic",2021, Available from: https://core.ac.uk/download/235272168.pdf
8. Barlian Henryranu Prasetio, Edita Rosana Widasari, Hiroki Tamura," Multiscale-based Peak Detection on Short Time Energy and Spectral Centroid Feature Extraction for Conversational Speech Segmentation",2021, ICPS Proceedings, SIET '21 Available from: https://dl.acm.org/doi/10.1145/3479645.3479675
9. Sakshi Gupta, Ravi S. Shukla, Rajesh K. Shukla, Rajesh Verma," Deep Learning Bidirectional LSTM based Detection of Prolongation and Repetition in Stuttered Speech using Weighted MFCC",2020, IJACSA, available form:10.14569/IJACSA.2020.0110941
10. Vinay N A, Bharathi S H," Dysfluency Recognition by using Spectral Entropy Features", IJEAT,2019 Available from:https://www.ijeat.org/wp-content/uploads/papers/v8i6/F7881088619.pdf
11. Manjutha M, Subashini P," Statistical Model-Based Tamil Stuttered Speech Segmentation Using Voice Activity Detection",2022, Journal of Positive School Psychology, Available form https://journalppw.com/index.php/jpsp/article/view/9958/6481
12. Salsabil Besbes, Zied Lachiri," Multi-class SVM for stressed speech recognition",2016, IEEE conference proceedings, Available from: DOI: 10.1109/ATSIP.2016.7523188
13. H.Y. Vani, Dr.M.A. Anusuya and Dr.M.L. Chayadevi "Fuzzy Clustering Algorithms - Comparative Studies for Noisy speech signals", 2019, ICTACT Journal on Soft Computing, Available from: https://ictactjournals.in/paper/IJSC_Vol_9_Iss_3_Paper_5_1920_1926.pdf
14. 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, pp. 556–569, 2009. Available from: DOI: 10.1044/1092-4388(07-0129)

AUTHORS PROFILE

K B Drakshayini is a research scholar of VTU Belgaum working under the guidance of Dr. Anusuya M A on stuttering Speech signal processing. Completed MTech in NIE, Mysore and bachelor degree in vidya Vardhanka college of Engineering. She has total of 15 years of experience in teaching. Published Papers in National/International journals and conferences in research field. Area of interest are Speech signal Processing, Data science, Machine learning

Dr. Anusuya M A is having M. Tech and PhD qualification in Computer Science and Engineering with specific research interest in the field of Speech signal processing. She has total 25 years of teaching experience and published around 60 papers in International / national journals and Conferences with special recognitions. Presently working as Associate Professor in JSS science and technological university, Mysore. Area of interest are Pattern Recognition, Speech Signal Processing, Machine learning, Machine Translation, Fuzzy based Mathematical modeling