Repetition Detection using Spectral Parameters and Multi tapering features

K B Drakshayini
Research Scholar; Computer Science and Engineering, JSSSTU, Mysuru
drakshakb@gmail.com

M A Anusuya
Associate Professor, Department of Computer Science and Engineering, JSSSTU, Mysuru
anusuya_ma@sjce.ac.in

Abstract

Handling and addressing the issues in disfluent speech is a challenging task. It is very tedious to identify and remove repetition at the pre-processing step. Many speech related applications such as speech to text alignment, voice based interactive system face these hurdles while designing an automatic disfluent speech recognition system. Since speaker can utter the repeated words partially or miss some words in between makes it challenging. Spectral parameters such as Energy, Entropy, Zero Crossing Rate and centroid are used to detect repetitions. The similarity scores between phonemes and syllabus are detected and computed by employing Dynamic time warping (DTW) and polynomial curve fitting (PCF) approaches. The reconstructed speech signal features are extracted using SWEC- multi tapering window of MFCC procedure. The extracted features are modelled using SVM yielding 85% of recognition accuracy with repetition detection accuracy as 78.04% automatically.

Keywords: Spectral parameters, Spectral energy, Spectral entropy, Zero crossing rate, spectral centroid, Multi taper windowing, SVM. DTW, PCF, repetition, stuttered speech.

1. Introduction

Automatic speech recognition (ASR) technology has significantly improved over the recent decades, resulting in highly sophisticated algorithms which can recognize words with high accuracy. However, in real time applications, the performance is still limited, based on the type of utterances, training models, environmental factors, and the disorders in the fluent speech [1]. The disfluencies in the normal speech evaluation data might result in system to return inaccurate transcription. In case of stuttered speech, ASR system performance is severely degraded due to higher degree of disfluencies [2].

Stuttering is a speech disorder which disrupts the normal flow of a speech by involuntary disfluencies. The repetition events are more prominent in stuttered speech compared to the other types of disfluencies. Repetitions can occur at phoneme level, syllable level, word level, sentence level. The person suffering from stuttering has slow speaking rate, prolonged pause interval between syllables or words with stop gaps [3]. Each repetition event approximately has similar acoustic features. In this work most prominent spectral parameters such as Energy, Entropy, ZCR and centroid are applied to detect the presence of repetitions at the phoneme and syllable level.

Forced alignment techniques like Dynamic Time warping and Polynomial curve fitting techniques are applied to identify the repeated frames of phonemes. Windowing is replaced by Multi tapering features to extract MFCC coefficients of the reconstructed signal. SVM is adopted to recognize the reconstructed disfluent speech using accuracy metric.

Rest of the paper is arranged as follows. Section 2 discussed few research works reported in this area. Data set used are discussed in section 3. Repetition analysis is discussed in section 4. Methodology applied is discussed in detail in Section 5. Repetition detection and reconstruction is discussed in section 6. Feature extraction and decision making is discussed in section 7 and 8. Challenges are discussed in section 9 and followed by conclusions and Future enhancement.

2. Related Work.

As per Literature the major work exists only for detection of the disfluency i.e., repetition at the primary level with MFCC features but not with spectral parameters. Also, disfluent speech recognition with multi taper is less addressed. The research gap is depicted in Table 1.

Table 1: Related work on Repetition detection

Author	Year	Types of stuttering	Database	Features	Classifiers	Accuracy (%)
Howell [4]	1995	Repetition and Repetition	12 children who stutter (UCLASS)	Autocorrelation function and Envelope parameters	ANN	80
Howell P [5-6]	1997	Repetition and Prolongation	12 speakers (UCLASS)	Duration energy Peaks	ANN	78
Wiśniewsk [7]	2007	Repetition	38 speakers	MFCCs	HMMs	80
Wiśniewski [8]	2008	Repetition and stop gaps	38 samples for repetition of fricatives + 30 samples for stop blockade + 30 free-of-silence sample	MFCCs	HMMs	70
Tian-Swee [9]	2007	Not specific (stop gap, repetition, repetition all)	15 normal speakers + 10 disorder speakers	MFCCs	HMMs	96
Świetlicka [10]	2009	Repetition	8 stuttering speakers + 4 normal speakers (yields 59 "fluent speech samples + 59 non-fluent speech samples)	Spectral measure (FFT 512)	Kohonen, MLP, radial basis function (RBF)	88.1–94.9
Ravikumar [11]	2009	Repetitions	10 speakers	MFCCs	Perceptron	83
Lim Sin Chee [12]	2009	Repetition and prolongation	10 speakers [UCLASS]	MFCCs	kNN, LDA	90.91
Lim Sin Chee [13]	2009	Repetition and prolongation	10 speakers [UCLASS]	LPCC	kNN, LDA	89.77
IreneuszCodello[14]	2011	Repetition	10 speakers	CWT	Kohonen network	90
Wiśniewski M.[15]	2011	Repetition	425 utterances	MFCC	HTK	83
M. Hariharan[16]	2012	Repetition and repetition	10 speakers [UCLASS]	LPC-based cepstral parameters	kNN, LDA	94 and above
Izabela Swietlicka [17]	2013	Blocks, repetition, Repetition	19 speakers		Hierarchical ANN	96
Marek WISNIEWSKI [18]	2015	Not specific	36 phonemes	13 MFCC	HTK	83
P. Mahesha [19]	2016	Syllable repetition Word Repetition	20 speakers [UCLASS]	LPC, LPCC, MFCC	SVM	75 92 88
S. Girirajan [20]	2020	Repetition, Repetition	UCLASS	MFCC	LSTM	_
Tedd Kourkounakis [21]	2019	sound repetition, word repetition, phrase repetition, revision, interjection, and prolongation	UCLASS	MFCC	Bi-LSTM	_
Tedd Kourkounakis [22]	2020	Repetition, revision, interjections, repetition	UCLASS	_	Fluent NET	_
Sparsh Garg [23]	2020	Not specific	UCLASS	MFCC	DNN and Bi-LSTM	82 81
Shakeel A Sheikh [24]	2021	Stop gaps, Repetition	100 speakers [UCLASS]	MFCC	TDNN	_

G. Diwakar [25] worked on repetition detection using MFCC features and DTW resulted in 79%detection accuracy for Dysarthria speech. Pravin B. Ramteke [26] worked on repetition detection for stuttered speech in regional language using MFCC feature extraction and DTW similarity matching. This has achieved 94% detection rate for a small size data set.

3. Data set

The experiments are conducted on the UCLASS dataset. This repository consists monologs, readings, and conversational recordings. For our simulation 80 samples at word level are derived from the sentence recording repository. It includes 22 utterances from female speakers and 58 utterances from male speakers with age ranging from 11 years to 20 years. The samples are chosen to cover speech samples of different age and gender [24]. Through listening perception, uttered words with repetition dysfluency are identified and derived manually. Few Sample Data set considered for analysis is shown in Table 2. Totally 100 signals are collected.

SL.NO	Actual word	Pronounced word	Age (years)	Sex	Derived from (.wav files)
D1	Ball	Ba/ba/aa/ll	11	F	F_0142_11y3m_1
D2	Different	Di/di /di/iiii/ffernt	11	F	F_0142_11y7m_1
D3	Climbed	Cl/cl/cl/iiiii/mbed	8	M	M_0016_1_08y7m_1
D4	Department	Depa/pa/pa/apartment	14	F	F_0101_14y8m_2
D5	Tuesday	Tueeeeesday	16	M	M_0104_17y1m_1
D6	Just	Ju/ju/ssssssss/t	11	F	F_0142_11y7m_1
D7	Every	E/vvvv/ery	16	M	M_0016_07y11m_1
D8	understanding	Under/st/st/st/ttt/anding	14	F	F_0101_14y8m_2
D9	Money	Mo/nnnnn/ey	11	F	F_0142_11y7m_1
D10	Finding	Fin/dddddd/ing	12	M	M_1206_12y3m_1
D11	Moment	Mo/me/me/nt/	11	F	F_0142_11y3m_1
D12	Started	Starte/te/te/d/	11	F	F_0142_11y3m_1
D13	Step	St/st/eeee/p	10	M	M_1202_10y11m_1
D14	Part	Pa/pa/aaaart	7	M	M_0016_07y11m_1
D15	Called	Ca/ca/called	8	M	M_0016_08y3m_1

Table 2: Data samples of stuttered speech sample words with Repetition

4.Repetition Analysis with tempo-spectral properties

Repetition analysis is an important pre-processing step in stuttered speech recognition. It is very challenging to detect repetition of Phonemes in word in stuttered speech, since the repeated words are not exactly identical type of repetition considered for analysis are repetition exists with stop gap and prolongation. Repetition are usually longer than fluent words. As seen in the oscillogram in Figure 1, the repetitive sections have a pattern of alternating energy and silence that is designated as a fragmentary pattern. Each repetition emitted will tend to have a similar spectral structure. These are the representations of the first part of the same phonemic, syllabic, or lexical event. This is observed from the Figure 1.

Figure 1 shows the repeated words have similar Spectro-temporal properties and energy contours. uttered word text representation (bottom row), spectrographic (second row) and oscillograph (bottom row) formats. These serve to illustrate the acoustic pattern of the repeated dysfluent repeated events.

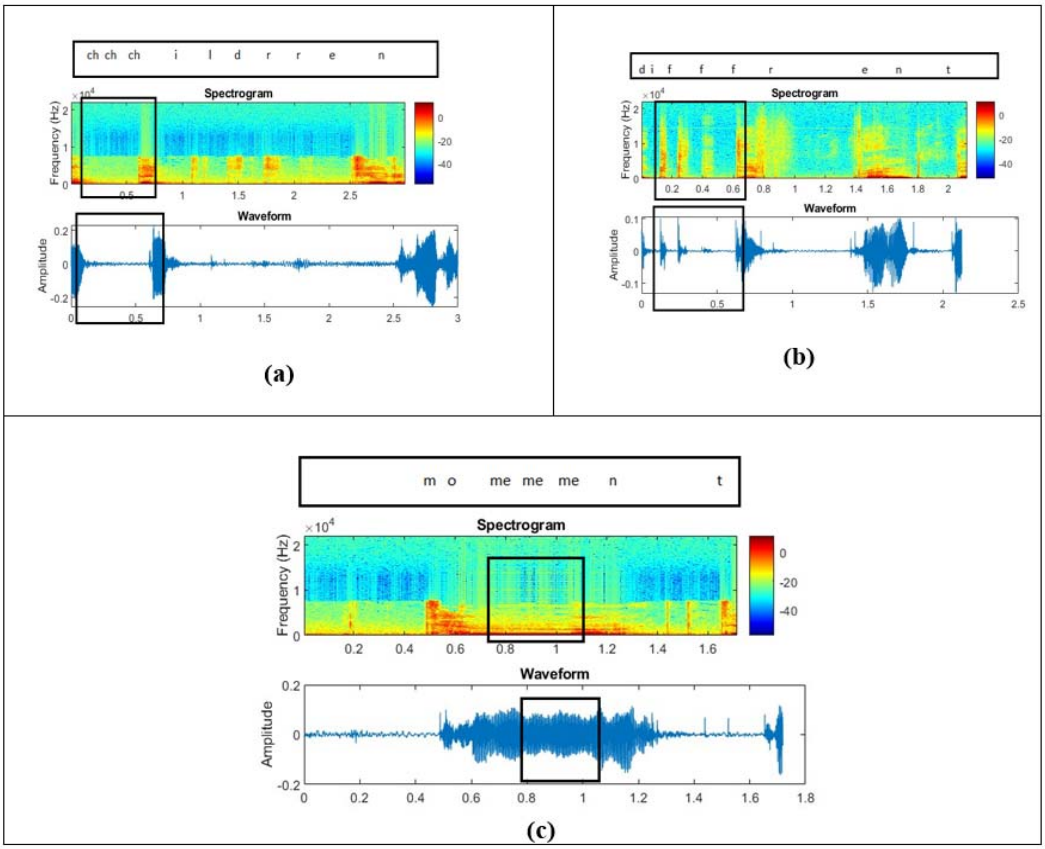

Figure 1: (a) Speech signal uttered word “children” as “ch/ch/ch/ldren” (b) Speech signal uttered word “different” as “di/fff/erent” (c)Speech signal uttered word “moment” as “mome/me/me/nt”

4. Methodology

The architecture of the proposed system is depicted in Figure 2. It consists of preprocessing, repetition detection and removal, signal reconstruction, feature extraction, and recognition phase. These are explained in this section briefly.

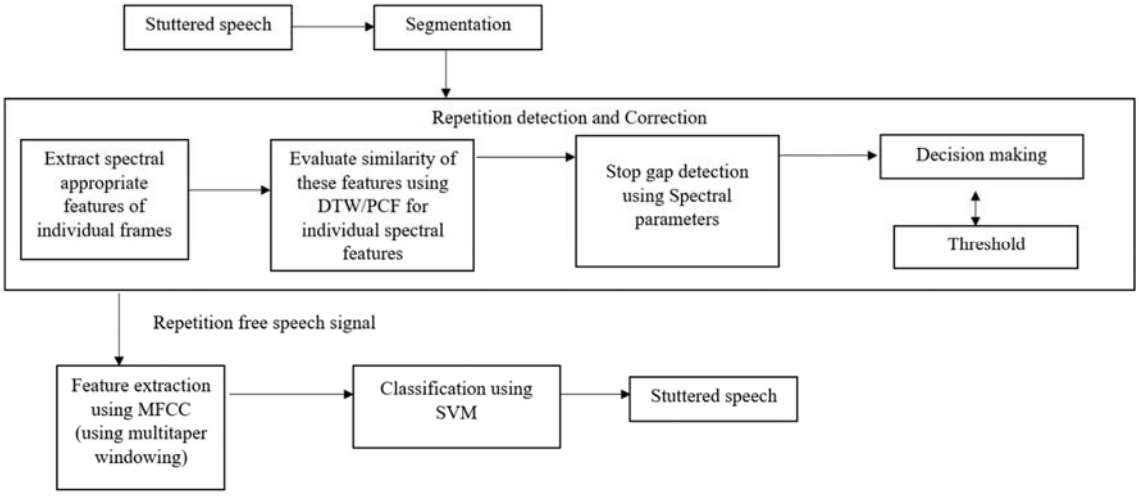

Figure 2: Repetition detection and Recognition of Stuttered speech model

The above diagram is analysed and processed in two phases namely i) Repetition detection and reconstruction of the signal ii) Speech Feature extraction and Recognition of the reconstructed signal.

5. Repetition Detection and Reconstruction of the Signal

Different phases of Repetition detection and Reconstruction is discussed in following section.

6.1 Preprocessing

The analysis of the repetition is the prominent task in handling diffluent speech processing applications. The signal is sampled for 16 KHz and noise is removed by employing first order filter. The speech segment is divided into 'X' Frames. To detect the presence of repetitions of the phonemes, each frames of a signal is subjected to extract the spectral parameters. For every disfluent speech segment frame Energy, ZCR, centroid and entropy parameters are computed [24][26]. The repeated frames will have approximately similar temporal and spectral properties as seen from Figure 1.

6.2 Similarity scores between words Computation:

To determine the similarity between two frames for repetition detection the spectral features of phonemes or syllabus are compared. The distance between these The speech signal is segmented into 200ms frames each. To identify the presence of the repeatedness of the phoneme and the syllables adjacent frames with their speech patterns are compared. To perform this comparison DTW and PCF techniques are applied. These two methods compare the threshold values of all the spectral parameters and if these values are below the threshold the frame is detected as repeated frame will be removed or else the frame is retained. The DTW and PCF is applied as follows.

6.3. DTW (Dynamic Time wrapping)

Dynamic time warping is an algorithm that calculates the optimal warping path between two data from sound so that the output is the path warping values and the distance between the two data frames. Warping path is the distance compared between two patterns. The smaller the warping path the two patterns are said to be the same. Two words from the same word by the same user can have different times. For example, word 'two' can be pronounced with two or twoooo or ttttttwooo. DTW solves this problem by aligning frames correctly and calculating the minimum distance between two phonemes. Different timing of speech alignment is a core problem for distance measurement in speech recognition. Small shifts result in incorrect identification [25][26]. Therefore, this algorithm is more realistic to be used in measuring the similarity of a speech pronounced pattern based on time. The sequence of speech data that varies in time is represented as in equation 1.

$$x = [x_1 x_2 \ldots x_n] \ and \ y = [\, y_1 \ y_2 \ldots y_n] \qquad (1)$$

The DTW algorithm aligns two vector sequences repeatedly until the optimal match between the two sequences is found. A linear mapping of the axis to align the two signal frames are considered. The distance between these two frames is computed using equation (2) and the minimum distance is computed using equation (3). The same procedure is adopted for all the frames and those frames which has equal or minimum distance is considered as repeated frames. This helps us to detect the repeated frames in the signal. These are removed and the signal is reconstructed Hence this is one approach to identify the similarity between the frame [25-26].

$$Dist(x,y) = |x - y| \ = \ [(x_1 - y_1)^2 + (x_2 - y_2)^2 \ + \ + (x_n - y_n)]^{\frac{1}{2}} \qquad (2)$$

$$D(i,j) = |t(i) - r(j)| + min \begin{bmatrix} D(i+1,j) \\ D(i+1,j+1 \\ D(i,j+1 \end{bmatrix} \qquad (3)$$

Apart from the above method we have also tried the PCF to compute the similarity scores between the two adjacent repeated frames. This is experiment to verify which similarity method would be better to increase the recognition accuracy of the stuttered repeated speech signal. Here distance matrix between adjacent frames using dynamic programming and then optimal path between each frame is found.

Here first it defines distance using matrix between two frames based on the spectral parameters ZCR, Entropy, Energy and Centroid. Distance matrix is computed using Euclidian distance. During the DTW computation, a distance matrix is filled using dynamic programming, where each cell in the matrix represents the

cumulative distance between elements of the two sequences up to that position. The goal is to find the path with the minimum cumulative distance, which represents the optimal alignment between the sequences.

The traceback process starts from the bottom-right corner of the distance matrix, where the final cumulative distance is found. Then, by backtracking along the path of minimum cumulative distances, we find the alignment between the elements of the two sequences. The alignment path connects the top-left and bottom-right corners of the matrix, indicating which elements are optimally matched.

By analyzing the alignment path, we can identify regions where elements in one sequence are aligned to multiple elements in the other sequence, indicating repetitions or temporal variations.

Table 3 shows part of the distance matrix computed during repetition detection and correction for the speech signal "moment" using DTW and in the optimal alignment path found stuttered speech repeated frames [0.1, 0.1] and matched the rest of the frames with some temporal shifts.

	0.1	**0.3**	**0.3**	**0.4**	**0.6**	**0.7**	**1.2**	**1.8**
0.2	0.0.2	0.08	0.18	0.26	0.48	0.06	1.18	1.98
0.3	0.18	0.00	0.06	0.04	0.26	0.44	0.84	1.58
0.1	0.34	0.06	0.02	0.06	0.08	0.24	0.64	1.26
0.4	0.40	0.10	0.10	0.02	0.16	0.14	0.46	1.04
0.6	0.68	0.38	0.26	0.06	0.04	0.32	0.16	0.62
0.7	0.88	0.58	0.34	0.20	0.02	0.14	0.24	0.22

(a)

Stuttered	0.1	0.3	0.1	0.1	0.4	0.1

(b)

Table 3: (a) distance matrix (b)Find the optimal alignment path from the top-left (0, 0) to the bottom-right (5, 7) of the distance matrix:

As Spectral parameter values varies with different signal, the threshold value can't be fixed to some specific value. Since the same data from the same speaker uttered for different times will have variations. The warping path threshold of the signal is considered by computing the average value of DTW coefficients.

The spectral parameters are ZCR, Entropy, Energy and Centroid. Distance matrix is computed using Euclidian distance. During the DTW computation, a distance matrix is filled using dynamic programming, where each cell in the matrix represents the cumulative distance between elements of the two sequences up to that position. The goal is to find the path with the minimum cumulative distance, which represents the optimal alignment between the sequences.

6.4..Polynomial Curve Fitting (PCF)

Many researchers have considered correlation coefficient matrix in normal application. In this paper we have replaced with PCF [25]. The objective of curve fitting is to find the parameters of a mathematical model that best describes a set of data in a way that minimizes the difference between the model and the data. Eqn. (4) shows the polynomial in which the dependent variable Y is expressed in terms of independent variable X.

$$Y = F(x) = a_0 + a_1x + \ldots + a_nx^n \qquad (4)$$

The phoneme or syllable is representing in terms of T × P feature vectors. The PCF method is applied to reduce the word to (O + 1) × P.
Where, T - is a number of frames in a word,
O - order of polynomial. (3^{rd},4^{th} and 5^{th} orders are considered)

Algorithm for similarity measure using PCF works as follows
1)Consider w_i and w_j frames of dimension (O+1) ×P,
2)convert 2-dimensional matrix into 1-dimensional column vector (i.e., T*P to $(O + 1) * P$)
3)Find the Euclidean distance between the vectors of w_1 and w_2 using Eq. (5)

$$D(w_i, w_j) = \sqrt{\sum_k {a_k}^2 - {b_k}^2} \quad (5)$$

PCF similarity measure is applied between two frames considering spectral parameters for various degree of polynomial. In this work the order is varied between 3rd to 5th order polynomial. Table 4 represent the magnitude of the signals variation as the order is varied. It is observed that as the order increases the magnitude of the coefficients increases. These large coefficients help to capture the better features for matching adjacent frames.

Order / Magnitude	5	8	10	15
W_0	0.19	0.82	0.31	-1.24
W_1	1.24	-1.27	1.24	0.82
W_2	-2.4	-1.24	-1.24	0.12
W_3	-1.24	0.82	17.82	2340.06
W_4	0.82	0.12	0.12	-1.24
W_5	0.12	0.06	25.06	-1.24
W_6	0.06	-1.34	-1.24	450.82
W_7		0.82	0.82	560.12
W_8		0.23	0.12	230.06
W_9		0.23	1. 06	-1.24
W_{10}			-1.24	-1.24
W_{11}			1.82	2430.82
W_{12}			0.12	0.12
W_{13}				-1.24
W_{14}				450.82
W_{15}				10.12
W_{16}				30.06

Table 4: Values of Coefficients for different polynomials of order

Here Average of PCF coefficients are considered as threshold value to decide whether frames are repeated or not. For example, frame 4 and frame 5 having PCF similarity as 0.08 which is greater than the average value 0.02 considered as threshold. Hence, these two frames are considered as repeated Frames. It is also observed that in between these repeated frames there is stop gap occurrences identified using hybrid approach [27].

6.5. Evaluation parameter to detect the repeated

Accuracy:

Accuracy (A) represents the ability of the repetition detection system to correctly identify repetitions in stuttered speech, taking into account both the true positives and false positives. Accuracy is calculated using Eqn (6)

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+TP+FN} \quad (6)$$

- True positive (TP): Repeated frames correctly identified as Repeated frames
- False positive (FP): Proper frames incorrectly identified as Repeated frames
- True negative (TN): proper frames correctly identified as Proper frames
- False negative (FN): Repeated frames incorrectly identified as Proper frames

6.7. Repetition detection and correction

Through perception and computation form DTW and PCF Techniques the signal “mo/mo/moment” is identified with 45 repeated frame and 78 non-repeated frames. TP and FP measures using DTW and PCF with Hybrid (combination of all 4 parameters). Each spectral parameter quantifies the number of frames to reduced and retained using TP, FP, TN, FN parameters individually for all the parameters. It is observed that 34 repeated frame and 11 non-repeated frames for the stuttered speech signal “moment” uttered as “/mo/mo/moent” and also observed the results of PCF and also individual parameters performance for the detection of repetition detection.

Type of similarity measure		Energy		Centroid		Entropy		ZCR		Hybrid	
		RP	NRP	RP	NRP	RP	NRP	RP	NRP	RP	NRP
DTW	45(RP)	22	23	21	24	34	11	28	17	34	11
	78(NRP)	15	63	10	68	15	63	12	66	16	62
3rd order PCF	45(RP)	16	29	15	30	21	24	12	33	23	22
	78(NRP)	10	68	13	65	12	66	21	57	12	66
4th order PCF	45(RP)	17	28	15	30	23	22	10	35	24	21
	78(NRP)	22	56	12	66	11	67	11	67	12	66
5th order PCF	45(RP)	19	26	14	31	15	30	19	26	33	12
	78(NRP)	12	66	15	63	11	67	13	65	10	68
6th order PCF	45(RP)	19	26	25	20	30	15	24	21	11	34
	78(NRP)	13	65	12	66	11	67	11	67	12	66

Table 5: TP and FP measures using DTW and PCF with Hybrid

Table 6 depicts the repetition detection accuracy. Among these DTW with hybrid approach yields a good result by obtaining 78.04 % accuracy.

Approach	**Detection accuracy %**
DTW	78.04
3rd order PCF	75.10
4th order PCF	71.42
5th order PCF	70.53
6th order PCF	69.51

Table 6: Repetition detection accuracy

Signal Reconstruction:

The signal is reconstructed by removing the repeated frames using DTW with hybrid approach [27][28]. This provides the stuttered free signal. This signal is further processed to extract the multi tapered speech features using MFCC. These features are modeled using SVM to recognize the stuttered speech recognition accuracy. The steps of Multi tapering are as discussed below.

7. Feature extraction using SWCE Multi tapered window:

The feature extraction process using MFCC is as depicted in the following Figure 3. The frames of the reconstructed signal are applied with multi tapering concept. In the windowing phase each frame is applied with various types of windows like hamming, Thomson, SWCE, multipeak windows. The table 5 depicts first level extracted features of MFCC. Among the four SWCE window is identified to extract best features. In SWCE number of tapers are varied to 5,8,10,15.

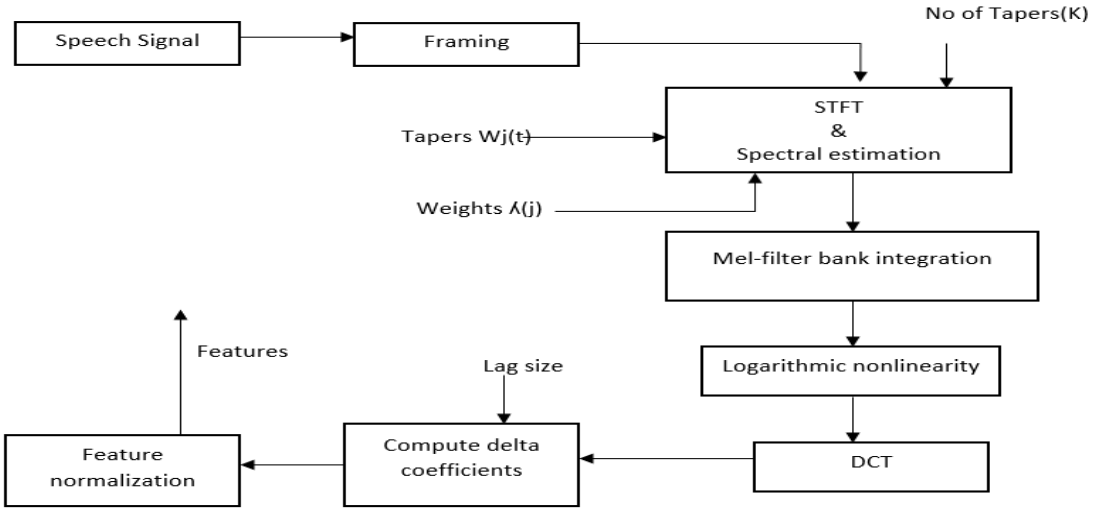

Figure 3: MFCC -Multi tapering Windowing

7.1 Multi taper windowing- MFCC

The pre-processing step includes pre-emphasizing, DC removal, signal normalization. signal normalization. In framing block the speech signal is divided small frames. Frames are again divided into small durations windows (tapers) instead of one window (Hamming). Then spectrum for each taper is estimated individually & averaged. As spectrum of each taper is uncorrelated weighted frequency domain averaging of the spectrums reduces the variance [29]. The MFCC filter bank improves Equal error rate (EER) & minimum detection cost function which indicates stable parameter setting. Then logarithmic nonlinearity is removed. Delta coefficients are estimated, then features are normalized by using feature normalization methods. It uses several window functions (tapers) to obtain a low-variance power spectrum estimate, given by

$$s(f) = \sum_{j=1}^{K} \lambda(j) \sum_{t=0}^{N-1} wj(t)\ x(t) e^{-i2\pi tf/N} \qquad (6)$$

Here, Wj (t) is the j^{th} taper (window) and λ(j) is its corresponding weight. The number of tapers, K, is an integer (typically between 4 and 15). There are a number of alternative taper sets to choose from: Hamming window, Thomson window, sinusoidal weighted cepstrum estimator (SWCE) and multi-peak.

7.1.1 Experimental set up for multi tapering windowing in MFCC

Frame Length and Hop Size: Frame length of 20 milliseconds with a hop size of 50% of the frame length (10ms for a 20ms frame) is used.
Pre-emphasis: The value for pre-emphasis is set to 0.95.
Number of Tapers: The 5,8,10,15 tapers determine the number of windows used in multi tapering.
FFT Size (n_fft): n_fft set to 400 bins
Multi tapering method: The Hamming, Thomson, SWCE or multipeak window are used for analysis.
Overlapping Frames:50 % overlap provide a smooth transition between adjacent frames and improve feature representation.

Multi tapering Approach	Multipeak	SWCE	Thom	hamming
MFCC Features(first or delta)	-2.0840372	**-2.0105558**	-1.7032807	-2.1316570
	-0.4253004	**-0.3950781**	-0.5217056	-0.4391995
	-0.2315805	**-0.2199401**	-0.2285837	-0.2975899
	0.0253836	**0.1221871**	-0.0788314	-0.0699947
	0.0803223	**0.1599228**	-0.1336564	-0.0697726
	0.1253176	**0.2315514**	-0.0499009	-0.0117938
	-0.1937938	**-0.1922349**	-0.2177202	-0.2236988
	-0.3922180	**-0.4827155**	-0.1997373	-0.2992718
	-0.2578681	**-0.3400444**	-0.2325524	-0.2245749
	0.0224257	**0.0800701**	-0.1101993	-0.0631673
	-0.1790647	**-0.1699162**	-0.1416473	-0.1817332
	-0.1569817	**-0.1665499**	-0.0699305	-0.1336261

Table 7: MFCC features for the speech signal (moment) for different windowing approaches

As per Table 7 Multi tapering windowing using SWCE resulted in extraction of better features compared to all other approaches. SWCE aims to improve the robustness of MFCC features by applying adaptive weights to each speech frame. By incorporating these adaptive weights, Table 8 depicts the tapered values of SWEC windowing for various taper levels. Taper value 15 has extracted the better features. Multi tapering windowing provides better frequency resolution compared to a single taper. It also allows for more precise localization of spectral components as depicted in Figure 4. The increased frequency resolution helps to capture the finer details of the frequency contents of the stuttered speech signal.

No of Tapers (SWCE)	15	10	8	5
MFCC tapered windowed Features	-1.76181	-1.56408	-2.13166	-2.02465
	-0.38169	-0.59848	-0.43920	-0.40277
	-0.19776	-0.20323	-0.29759	-0.17112
	-0.06470	-0.11733	-0.06999	0.10226
	-0.12254	-0.14626	-0.06977	0.20271
	-0.03025	-0.07544	-0.01179	0.22741
	-0.15763	-0.21530	-0.22370	-0.16875
	-0.16772	-0.17534	-0.29927	-0.45481
	-0.24768	-0.20324	-0.22457	-0.25808
	-0.12623	-0.11128	-0.06317	0.10432
	-0.08196	-0.13333	-0.18173	-0.46471
	-0.06486	-0.06908	-0.13363	-0.44615

Table 8: MFCC features for the speech signal (moment) using SWCE windowing approach and varying no of tapers

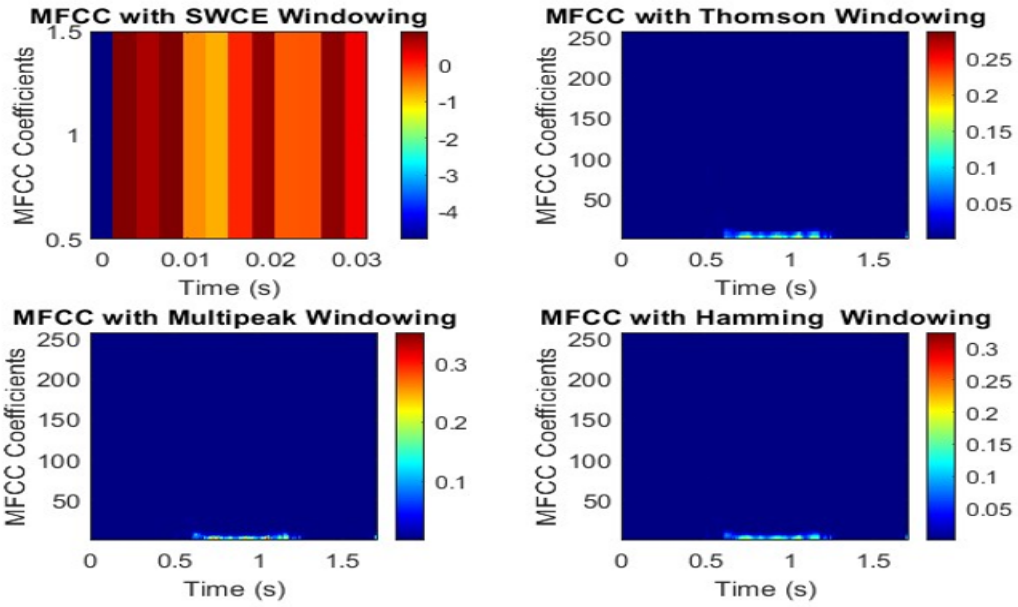

Figure 4: MFCC with multi tapering using Different windowing Approaches

8. Decision making using SVM classifier

SVM is a powerful machine learning tool to obtain a good separating hyper-plane between two classes in the higher dimensional space [30]. It is a predominant technique to estimate the basic parameters of speech. The procedure adopted to build the SVM is as discussed below.

Simulation Setup:
Step1: SVM Model Training (One-vs-One approach):

- With the one-vs-one strategy, create a binary classifier for every possible pair of classes.
- For 100 sample, create (100 * 99) / 2 = 4950 binary classifiers.
- Each binary classifier is trained on the training set using the data from the two corresponding classes.
- Step2: Model Validation:
- Evaluate the performance of each binary SVM classifier using the testing set.
- For each audio sample in the testing set, apply each binary classifier to get the class predictions.
- Combine the predictions from all binary classifiers majority voting to obtain the final predicted class for each sample.
- Step3: Prediction:
- For a new, unseen audio sample, pre-process it to extract the feature vector.
- Apply each of the 4950 binary SVM classifiers to get the confidence or probability scores for each pair of classes.
- Combine the predictions from all binary classifiers to determine the final predicted class for the sample.

Table 9 depicts the speech recognition accuracy with the SVM tool. DTW with multi tapering has obtained the 85% accuracy of the reconstructed signal. This is due to the advantage of multi taper. This yields every frame multiple independent estimate the power at component frequency with low variates and multiple spectral estimates. with for each sample frame with low variance.

Algorithm: Complete flow of the model:
Step 1: Read stuttered speech Signal With repetition Stuttredness (8KHZ)
Step 2: Pre-emphasis is performed by filtering the speech signal with first order FIR filter
Step 3: Calculate the **t**hresholds for all the spectral parameters.
Step 4: Calculate Similarity of Spectral parameters using DTW and PCF techniques
Step 5: Detection and removal of Repeated Frames with the reconstruction of the signal
Step 6: Multi tapered MFCC is applied to extract the first order speech features with SWEC multi tapered window.

The signal is divided into frames of 20ms

- Multi Tapering Windowing (no of tapers varied from 5 to 15)
- Mel-scale filter bank analysis with 400 bins
- Apply the log to obtain smaller components of a signal
- IDCT is applied to extract cepstral features

Step 7: Classification Using SVM

Detection Method (similarity measure)	**Feature Extraction**	**Classifier**	**Recognition Rate**
DTW	MFCC	SVM	78%
	MFCC+ Multi tapering windowing		85%
3rd order PCF	MFCC		70%
	MFCC+ Multi tapering windowing		72%
4th order PCF	MFCC		65%
	MFCC+ Multi tapering windowing		67%
5th order PCF	MFCC		63%
	MFCC+ Multi tapering windowing		66%
6th order PCF	MFCC		56%
	MFCC+ Multi tapering windowing		58%

Table 9: Stuttered speech recognition accuracy

Detection of Repetition using spectral parameters with DTW similarity measures followed by multi tapering windowing MFCC, classified using SVM results in better recognition rate as shown in Table 9.

9.Challenges:

- Distinguishing between normal repetitions and stuttered repetitions is very difficult because of the utterance duration.
- It is difficult to identify the repetition of sound or syllables as it involves both.
- Since speech signals can be influenced by various factors like background noise, speaker variations, and co-articulation effects, makes it challenging to accurately identify and segment stuttered repetitions.
- It is challenging to identify the specific characteristics and timing of stuttered repetitions.
- The "one-vs-one" strategy in SVM becomes computationally expensive and memory-intensive, with a large number of classes.

10. Conclusions and Future enhancement:

A knowledge-based approach has been proposed to detect the repetition for the UCLASS dataset. The spectral features are applied to detect and remove the repeated frames with the DTW and PCF techniques automatically. This work highlights the usage of hybrid spectral features with DTW similarity score technique is 78.04% better in detecting the repeated frames from the existing works. Final reconstructed speech signal MFCC features are extracted using Multi tapering window. Finally, 85% of recognition accuracy is obtained when modeled with SVM. The advantage of the proposed method is data driven. In future the above procedure is planned to be applied for word and phrases repetition detection. It can also be extended to automatic boundary detection applications.

Conflict of interest

All authors declare no conflicts of interest in this paper.

11.References

[1] Jinyu Li," Recent Advances in End-to-End Automatic Speech Recognition", APSIPA Transactions on Signal and Information Processing,2022.

[2] Colin Lea, Zifang Huang, "From User Perceptions to Technical Improvement: Enabling People Who Stutter to better Use Speech Recognition," Proceedings of the CHI Conference on Human Factors in Computing Systems,2023

[3] Ankit Dash, Nikhil Subramani," Speech Recognition and Correction of a Stuttered Speech", IEEE,2018

[4] P. Howell, S. Sackin, "Automatic recognition of repetitions and prolongations in stuttered speech", Proceedings of First World Congress on Fluency Disorders, pp. 372–374, 1995.

[5] P. Howell, S. Sackin, K. Glenn, "Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: I. Psychometric procedures appropriate for selection of training material for Lexical dysfluency classifiers", Journal of Speech, Language, and Hearing Research, Vol. 40, 1997.

[6] P. Howell, S. Sackin, K. Glenn, “Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: II. ANN recognition of repetitions and prolongations with supplied word Segment markers", Journal of Speech, Language, and Hearing Research, Vol. 40,1997.
[7] M. Wisniewski, W. Kuniszyk-J_o_zkowiak, E. Smolka, W. Suszynski, "Automatic detection of prolonged fricative phonemes with the hidden Markov models approach", Journal of Medical Informatics & Technologies, Vol. 11, 2007.
[8] M. Wisniewski, W. Kuniszyk-J_o_zkowiak, E. Smolka, “Automatic detection of disorders in a continuous speech with the hidden Markov models approach", Proceedings of Computer Recognition Systems 2, Vol. 45, 2008.
[9] Tian-Swee Tan, Helbin-Liboh, A. K. Ariff, Chee-Ming Ting and Sh-Hussain Salleh,” Application of Malay Speech Technology in Malay Speech Therapy Assistance Tools”, 2007
[10] I. Swietlicka, W. Kuniszyk-J_o_zkowiak, E. Smolka, “Artificial neural networks in the disabled speech analysis", Proceedings of Computer Recognition System 3, Vol. 57, 2009.
[11] K.M. Ravikumar, R. Rajagopal, H.C. Nagaraj, “An approach for objective assessment of stuttered speech using MFCC features", ICGST International Journal on Digital Signal Processing, Vol. 9,2009.
[12] L. Sin Chee, O. Chia Ai, M. Hariharan,” MFCC based recognition of repetitions and prolongations in stuttered speech using k-NN and LDA", Proceedings of IEEE Student Conference on Research and Development, 2009.
[13] L. Sin Chee, O. Chia Ai, M. Hariharan, S. Yaacob, “Automatic detection of prolongations and repetitions using LPCC", Proceedings of IEEE International Conference on Technical Postgraduates, 2009.
[14] Ireneusz Codello, W. Kuniszyk-Józkowiak, ,”Disordered sound repetition recognition in ´ continuous speech using CWT and Kohonen network”,2011
[15] Wiśniewski M. , Kuniszyk-Jóźkowiak W, ”Automatic detection and classification of phoneme repetitions using HTK toolkit”,2011
[16] M. Hariharan, L. Sin Chee, S. Yaacob, “Classification of speech dysfluencies using LPC based parameterization techniques", Journal of Medical Systems, Vol. 36, 2012
[17] Izabela Swietlicka, Wiesława Kuniszyk-Józkowiak, Elzbieta Smołka,” Hierarchical ANN system for stuttering identification”,2013
[18] Marek WISNIEWSKI, Wiesława Kuniszyk-Józkowiak,” Automatic Detection of Stuttering in a Speech”,2015
[19] P. Mahesha, D.S. Vinod,” Automatic Segmentation and Classification of Dysfluencies in Stuttering Speech”,2016
[20] S. Girirajan, R. Sangeetha, T. Preethi, A. Chinnappa,” Automatic Speech Recognition with Stuttering Speech Removal using Long Short-Term Memory (LSTM)”,2020
[21] Tedd Kourkounakis, Amirhossein Hajavi, Ali Etemad,” Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory”,2019
[22] Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad,” FluentNet: End-to-End Detection of Speech Disfluency with Deep Learning”,2020
[23] Sparsh Garg,” Transfer Learning based Disfluency Detection using Stuttered Speech”, Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India,2020
[24] Shakeel Ahmad Sheikh, Md Sahidullah, Fabrice Hirsch, Slim Ouni “StutterNet: Stuttering Detection Using Time Delay Neural Network”,2021
[25] G. Diwakar,” Repetition Detection in Dysarthric Speech”, IEEE,2017
[26] Pravin B. Ramteke, “Repetition Detection in Stuttered Speech”, Proceedings of 3rd International Conference on Advanced Computing, Networking,2016
[27] 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
[28] K B Drakshayini, Anusuya M A, “Hybrid Approach to Detect Prolonged Speech Segments”, International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249-8958 (Online), Volume-12 Issue-4, April 2023
[29] Salsabil Besbes “Multitaper MFCC Features for Acoustic Stress Recognition from Speech”, International Journal of Advanced Computer Science and Applications,2017
[30] P. Mahesha, “Support vector machine-based stuttering dysfluency classification using GMM supervectors”, International Journal of Grid and Utility Computing,2015

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 Vardhaka college of Engineering. She has total of 15 years of experience in teaching in Vidya Vikas institute of Engineering. 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 modelling.