IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 25, Issue 4, Ser. 2 (July. – August. 2023), PP 38-45
www.iosrjournals.org

Stuttered Speech Analysis Using Machine Learning Algorithms

Dadavali S.P * , Meharunnisa S.P**
** Departement of computer Science Government First Grade College*
*** Departement of Electronics & Instrumentation Engineering, Dayananda Sagar College of Engineering*

ABSTRACT
Stuttering is a speech disease which involves abnormalities or disorder in speech signal. It is also called as stammering. Stuttering involves disorders in speech, such as repetition of word, interjections, prolongations, broken words, revisions, incomplete phrases and silent pauses. Stammering is the research matter in many advance areas like speech psychology, signal analysis. Therefore, this speech study is advantage for advanced field for scientific work. One of the main issues still unresolved in area of speech disfluency is an aid and automatic way of detection on patient condition after initial and final speech therapy techniques and a contribution of treatment made after intercession. Generally, classification of speech language disfluency is taken up as a very difficult and complicated problem however some typical problems related to stuttering are known. Stammering is a poorly known communication with 1% predominance. Analysis of stuttered signal includes syllable per minute (SPM), Percentage disfluency (PD), number of repetitions, number of prolongation and interjection.

Keyword: *VAD Segmentation Speech Spectogram Mel Cepstrums Feature Extraction KNN Classifier*

Date of Submission: 08-08-2023 Date of Acceptance: 18-08-2023

I. INTRODUCTION

In traditional stuttering detection process, the speech signal is translated and disfluencies like counts of repetitions, interjections and prolongations are identified[1]. Then the frequency of occurrence of each disorder is counted. These detection processes are based on the ability and experience of speech consultants. The main drawbacks of such detection are wastage of time, subjective, not consistent and also poor decision when different consultants make counts on same signal.

In conventional stuttering detection process, stuttering events like repetition, interjection and prolongation are counted manually from an observed recorded speech. From these counts the number of dysfluent words and fluent words are calculated, but in this method result will not produce correctly. Therefore in order to help the speech pathologists to treat the stuttered patients effectively automatic stuttering detection process[21], it also improves the judgement of stuttered events. Conventional stuttering process will consume time to analyze and also results are not produced correctly. Therefore, in order to get the correct result automatic stuttering detection[18] process are used in stuttering analysis process.

II. FACTORS THAT LEAD TO CONTRIBUTION OF SPLITTEERING

Stuttering is detected by identifying either of four signatures [5] of speech disorder: Developmental disorders Stuttering takes place during developmental stage. In other words, it can be described as an adult stuttering. The primary reason for stuttering is that when person starts stuttering from his childhood.

Auditory processing under activities

When persons stutters brain scans found that auditory processing region is underactive. To correct this abnormal a electronic device called altered auditory feedback is used.

Over activities of speech motor control

When persons stutters brain scans found that speech muscle control [14] area is overactive. To correct this abnormal fluency shaping therapy technique is used to speak with relaxed speech muscles. Stuttering [8] person over tense his breathing, jaws, tongues, vocal folds and lips.

Stress related response

Most stuttered person speaks fluently [21] when he is in relaxation mode, but he will stutters when gone under stress. To correct this abnormality personal construct therapy is used.

Neurotransmitter and genes
High levels of neurotransmitter dopamine in left region of speech motor control areas leads to a neurological abnormality.

III. SEGMENTATION OF SPEECH SIGNAL

Speech is one of the mediums by which communication is possible for human beings. For improved speech recognition system speech or syllable segmentation is employed. Segmentation of a speech will be done into units such as words and syllables. Speech segmentation is required to make better speech recognition between machines and humans as if like between humans. For making speech better understandable and to improve interpretation speech segmentation is used. Speech signals are classified as silence, unvoiced and voiced. Silence is nothing but the representation of gaps between voiced speeches. Phonetic will not provide accurate amount of syllables. The important feature of syllable is a dynamical part of transient consonant-vowel. The feeling of a syllable edge is usually very tough and not unique. Techniques used in an automatic syllable segmentation of speech, includes signal extremes and auto regressive coefficients (AR). The segmented syllables are fed to a feature extraction process. Speech segmentation acts as sub part of a speech recognition system. Speech recognition and synthesis systems are segmented into units such as syllables or words.

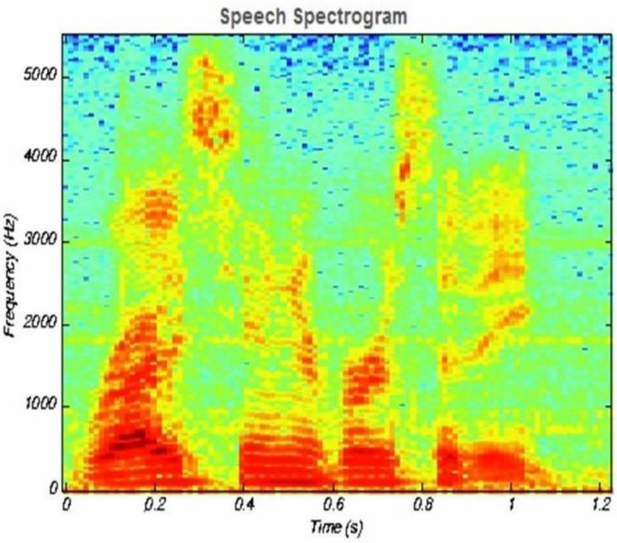


Figure 1. Schematic diagram of proposed system

Methods of Segmentation

Segmentation of speech is a process of breaking a continuous speech into primitive units with having finest edges. Recognition of speech is a crucial step; it also plays an important aspect in other certain applications. Segmentation of speech can be used for recognition of speech. Speech segmentation are of two types,

- Manual segmentation
- Automatic segmentation

In case of manual segmentation, speech segmentation will be done manually by testing a speech signal waveform with a spectrogram. In this process drawbacks are results cannot produce back, time consuming, endless process and contains more errors. Figure 1. Shows the Spectogram of a speech signal

In case of **automatic** segmentation speech segmentation is considered as better when compared to manual segmentation. In this case speech can efficiently segmented to its primitive units such as syllables and words. The various techniques involving in automatic syllable segmentation are

- Fourier transform method
- Short term energy-based method
- Minimum phase group delay method
- Wavelet method
- Word chopper technique

Voice Activity Detection (VAD) Segmentation

Voice activity detection (VAD) based segmentation detects the voiced speech from a given speech samples. The basic principle of VAD segmentation involves in extracting the features from given input samples and to compare it with the thresholds generated from noise activity. Voice activity detection (VAD=one) denotes if calculated value exceeds above thresholds. For non speech or silence portion of speech VAD is terminated as zero. The basic flow of VAD syllable segmentation is shown in Figure 2.

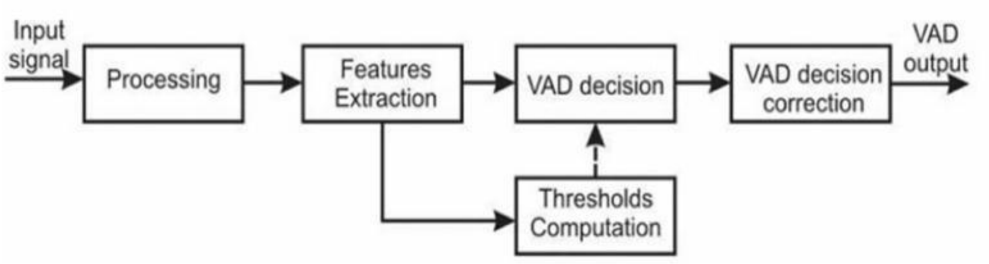


Figure 2. Basic Block Diagram for VAD Segmentation

In case of VAD segmentation it extracts a acoustic features that appropriate indicates a probability of observed speech signals. Based on the calculated acoustic features it determines whether the target speech signals present in a observed speech signals by using a computed threshold value. VAD output performance is based on frame and frame pattern where input signal frame length is given by 5-40ms duration. The reliability and accuracy of VAD segmentation is based on decision thresholds. The threshold values helps to detect the track of voice activity and thus results in efficient voice activity detection result. The threshold and energy of observed speech signal will be compared which depends on the noise in case of energy based VAD segmentation.

Proposed VAD Segmentation Algorithm

In proposed VAD syllable segmentation algorithm periodicity measurement of speech signal, low frequency to high frequency energy ratio and computation of full energy band were used as feature extraction for segmentation process.Pitch period of signal is main factor to determine the signal periodicity (C). In order to reduce complexity of computation, center clipping of input signal is achieved than for pitch estimated normalized function of autocorrelation r(t) is used in Equation 1

$$r(t) = \left(\sum_{n=0}^{N=m-1} x(n) \times x(n-1)\right) \div \left(\sqrt{\sum_{n=0}^{N=m-1} x^2(n+t)}\right) \qquad (1)$$

Where x(n) denotes an input frame signal n = 0,1, …….. , N. The function of autocorrelation calculated for the values of t from tmin to tmax. The constant tmax and tmin represent the upper and lower limits of respected pitch periods. The function of autocorrelation is maximized by the value of t which is equal to pitch period of a voiced frame. The periodicity of speech signal (C) is determined by maximum value of r(t).

The RMSE of complete band is calculated from a range of frequency (0hz to 4khz). The threshold computation of complete energy level Emax and Emin obtained from incoming frames. The value of Emin and Emax are stored in the memory and threshold is computed by the following equation 2

$$\text{Threshold} = (1-\lambda)\cdot \text{Emax} + (\lambda \cdot \text{Emin}) \qquad (2)$$

Where λ is a scaling factor which controls the process of estimation. VAD detection segmentation performs effectively when scaling factor lies in the range of 0.95 to 0.999. For different values of signals λ value would not remain same it should be set properly. Computation of scaling factor by the equation 3

$$\lambda = \text{Emax} - \text{Emin} / \text{Emax} \qquad (3)$$

Energy ratio (Er) is the ratio between energy of high frequency to low frequency. The high frequency part is obtained by passing the speech signal to high pass filter of 2khz. The energy ratio (Er) is determined by following equation 4

$$\text{Er} = \text{Eh} / (\text{Ef}-\text{Eh}) \qquad (4)$$

Where Eh and Ef are high band and full band energy of signal respectively. Once feature extraction process is completed, VAD initial decision (Ivad) is computed by comparing the extracted features with thresholds. Once the thresholds have been compared to detect the value of Ivad then final decision is done by observing lower portion of flowchart. For each value of Ivad, output decision Fvad is computed by the comparison of threshold. At the final output smooth hangover algorithm is used to detect the voiced portion and silence part. At the beginning of VAD, final VAD flag (Fvad) and hangover flag (Hvad) will set to zero. The output block of algorithm check whether the Ivad = 1 if it satisfied, then presence of speech will be detected. Then output decision sets hangover flag and final VAD flag to be one. It means speech has been not detected. The output decision checks for whether smoothened value (Efs) less than Emin, then hold over has been indicated. Therefore output block maintains final VAD decision to be one even though speech signal been not detected . Figure 3 shows the flowchart for VAD Segmentation.

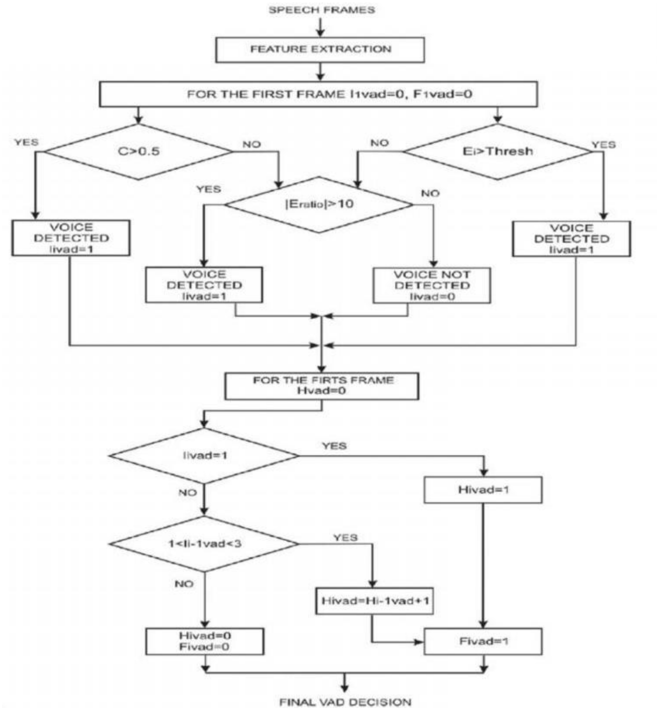


Figure 3. VAD segmentation Flowchart

Mel Cepstrums

In the final step the log spectrum which is in frequency domain converted back to time domain. Therefore, this result is named as MFCC. The cepstral of speech signal provides local spectral properties of speech signal which will useful in signal analysis. Since the calculated mel coefficients contains real numbers. It can be converted back to domain of time by using discrete cosine transform (DCT). MFCC provides speech signal spectral properties which can be extracted during the process of feature extraction as shown in Figure 4

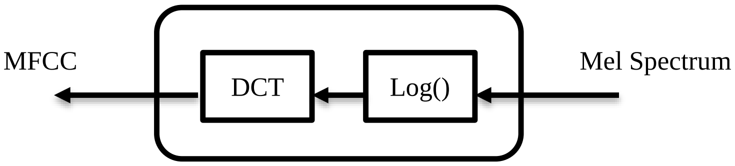


Figure 4. Mel Cepstrums Coefficients

This speech signal is represented as a convolution between quickly varying source (glottal pulse) and slowly varying filter (vocal tract impulse response) and also speech signal spectrum consists of the low frequency (spectral envelope) and the high frequency (spectral details).

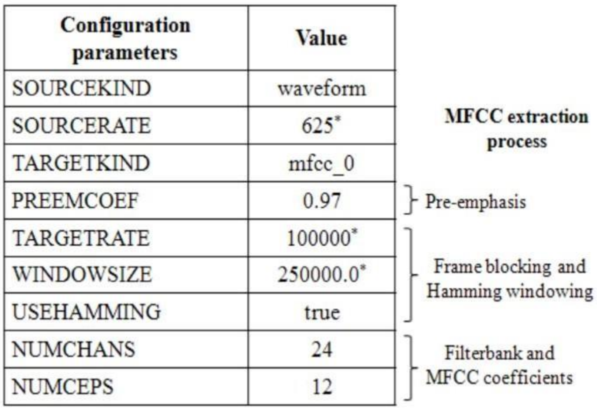
The logarithm leads to effect of change from multiplication to addition. The same technique is used to separate the spectral details and spectral envelope from a magnitude spectrum[13][17]. After this we have to apply DCT on the magnitude spectrum. From Sk values of each filter given, parameter of cepstrum in a Mel scale can be calculated by following equation 6.

$$MFCC_n = \sum_{k=1}^{k} (logSk) cos\left[n(k-0.5)\left(\frac{\pi}{K}\right)\right] \quad (6)$$

Where N is the required MFCC parameter number, S_k is the power spectrum coefficient, k is the number of filters n=1,2,…N.

This speech signal is represented as a convolution between quickly varying source glottal pulse and slowly varying filter (vocal tract impulse response) and also speech signal spectrum consists of the low frequency (spectral envelope) and the high frequency spectral details. Table 1 shows the technical parameters of MFCC process.

Table 1. Technical parameters of MFCC process (* express units in 100n)



Configuration parameters	Value	
SOURCEKIND	waveform	MFCC extraction process
SOURCERATE	625*	
TARGETKIND	mfcc_0	
PREEMCOEF	0.97	Pre-emphasis
TARGETRATE	100000*	Frame blocking and Hamming windowing
WINDOWSIZE	250000.0*	
USEHAMMING	true	
NUMCHANS	24	Filterbank and MFCC coefficients
NUMCEPS	12	

IV. CLASSIFICATION OF SPEECH SIGNAL

KNN Classifier

KNN algorithm stands for k nearest neighbors based algorithm which is a non-parametric technique used for regression and classification work. The output of k-nn classifier depends on whether it is used for regression or classification. In feature space an input consists of closest k training examples for both regression and classification work.

In this work two training data set is made one for dysfluent speech (repetition, prolongation and interjection) and another for fluent speech set. For each test data the training samples are detected with K-nearest neighbours. Further this k-nearest neighbours suitable class is labelled based on its majority voting. This class labels can be dysfluent speech or fluent speech. K-nn algorithm is based on clustering algorithms which will partitions a provided dataset into specified quantity of cluster k. Figure 5. shows the Knn classifier flow chart.

The code generation of training vector begins from initial estimate and will continue with centroid technique and nearest neighbor until termination criterion is achieved. This procedure will continue until mean squared error between cluster centroids and feature vectors falls below some threshold.

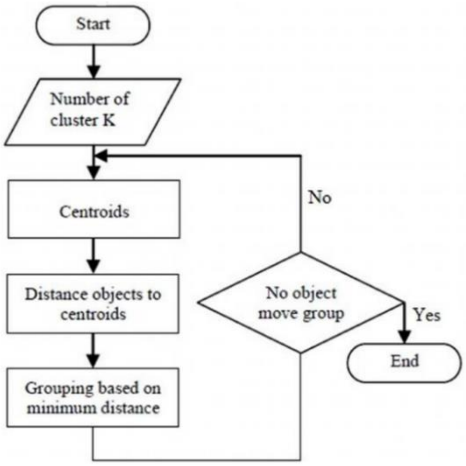

Start
Number of cluster K
Centroids
Distance objects to centroids
Grouping based on minimum distance
No object move group
No
Yes
End

Figure 5. K-nn Classifier Flow chart

The Process steps are depicted in the flow chart shown in fig 5 and squared function shown in Eq 7.
The squared function is given by,

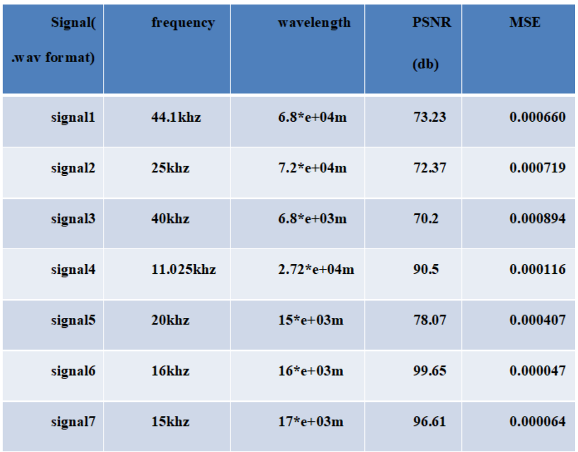
$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left|\left| x_i(j) - c_j \right|\right|^2 \qquad (7)$$

V. RESULTS AND ANALYSIS

Pre-Emphasis filter output and Signal Parameters

The first order high pass filter is designed in the preprocessing of speech signal. The objective is to enhance the total energy present in the high frequency part.

Table 2: Pre-Emphasis Signal Parameters



Signal(.wav format)	frequency	wavelength	PSNR (db)	MSE
signal1	44.1khz	6.8*e+04m	73.23	0.000660
signal2	25khz	7.2*e+04m	72.37	0.000719
signal3	40khz	6.8*e+03m	70.2	0.000894
signal4	11.025khz	2.72*e+04m	90.5	0.000116
signal5	20khz	15*e+03m	78.07	0.000407
signal6	16khz	16*e+03m	99.65	0.000047
signal7	15khz	17*e+03m	96.61	0.000064

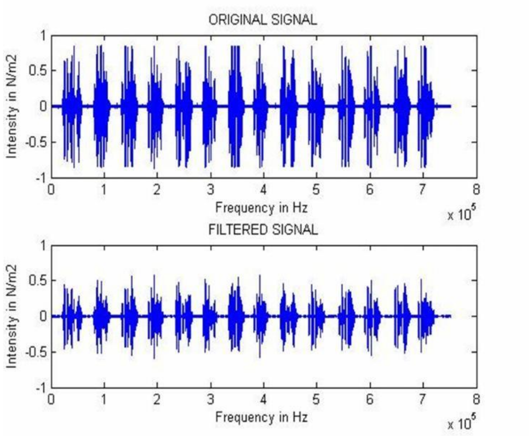


Figure 6. Pre-emphasis filter output

Figure 6 shows the Pre- emphasis filter output. From this figure, it is clear that the high frequency component that suppressed was compensated by Pre- emphasis filter. After Pre- emphasis filter applied it leads to improvement in the peak signal to noise ratio and mean square error is reduced.

The Table 2 describes about the peak signal to noise ratio and mean square error obtained after pre-emphasis filter is applied. From table it is clear that peak signalto noise ratio will be improved after Pre-emphasis filter is applied. Peak signal to noiseratio is the ratio between maximum possible powers to the power of corrupting noise. Better PSNR leads to a better result of the signal. From above table we will get a good PSNR for signal6, signal7, and signal4. MSE should be as low as possible for a processing of speech signal Since MSE is inversely proportional to a PSNR. For less PSNR ratio the mean square error will be high. For improved signal processing technique. PSNR ratio must be high and mean square error (MSE) should be low.

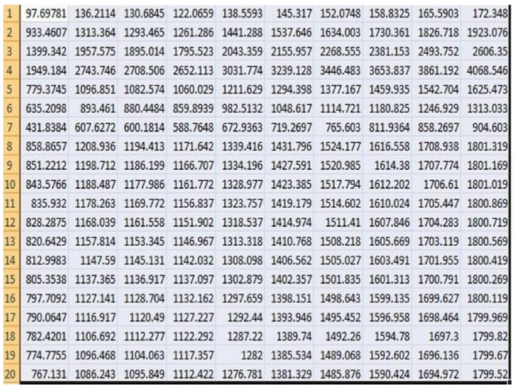
The below table 3 shows the segmented [21] test data which describes the number of words segmented using VAD segmentation. The number of words will differ for different speech signals, if time duration of speech signal is more than we can get more words.

Table 3 : Segmented test data

Test Data	Number of Words
Test Data 1	31
Test Data 6	30
Test Data 7	14
Test Data 8	72
Test Data 9	89
Test Data 10	84
Test Data 11	50
Test Data 12	44
Test Data 13	238
Test Data 14	120

MFCC Feature Output

The graph describes the Mel frequency cepstrum coefficients which are subjected to classification work. These features are stored in data matrix during training phase and used for comparison using testing phase. Figure 7 shows the MFCC feature output and Figure 8 shows Mel frequency cepstrum coefficients.



1	97.69781	136.2114	130.6845	122.0659	138.5593	145.317	152.0748	158.8325	165.5903	172.348
2	933.4607	1313.364	1293.465	1261.286	1441.288	1537.646	1634.003	1730.361	1826.718	1923.076
3	1399.342	1957.575	1895.014	1795.523	2043.359	2155.957	2268.555	2381.153	2493.752	2606.35
4	1949.184	2743.746	2708.506	2652.113	3031.774	3239.128	3446.483	3653.837	3861.192	4068.546
5	779.3745	1096.851	1082.574	1060.029	1211.629	1294.398	1377.167	1459.935	1542.704	1625.473
6	635.2098	893.461	880.4484	859.8939	982.5132	1048.617	1114.721	1180.825	1246.929	1313.033
7	431.8384	607.6272	600.1814	588.7648	672.9363	719.2697	765.603	811.9364	858.2697	904.603
8	858.8657	1208.936	1194.413	1171.642	1339.416	1431.796	1524.177	1616.558	1708.938	1801.319
9	851.2212	1198.712	1186.199	1166.707	1334.196	1427.591	1520.985	1614.38	1707.774	1801.169
10	843.5766	1188.487	1177.986	1161.772	1328.977	1423.385	1517.794	1612.202	1706.61	1801.019
11	835.932	1178.263	1169.772	1156.837	1323.757	1419.179	1514.602	1610.024	1705.447	1800.869
12	828.2875	1168.039	1161.558	1151.902	1318.537	1414.974	1511.41	1607.846	1704.283	1800.719
13	820.6429	1157.814	1153.345	1146.967	1313.318	1410.768	1508.218	1605.669	1703.119	1800.569
14	812.9983	1147.59	1145.131	1142.032	1308.098	1406.562	1505.027	1603.491	1701.955	1800.419
15	805.3538	1137.365	1136.917	1137.097	1302.879	1402.357	1501.835	1601.313	1700.791	1800.269
16	797.7092	1127.141	1128.704	1132.162	1297.659	1398.151	1498.643	1599.135	1699.627	1800.119
17	790.0647	1116.917	1120.49	1127.227	1292.44	1393.946	1495.452	1596.958	1698.464	1799.969
18	782.4201	1106.692	1112.277	1122.292	1287.22	1389.74	1492.26	1594.78	1697.3	1799.82
19	774.7755	1096.468	1104.063	1117.357	1282	1385.534	1489.068	1592.602	1696.136	1799.67
20	767.131	1086.243	1095.849	1112.422	1276.781	1381.329	1485.876	1590.424	1694.972	1799.52

Figure 7 MFCC Feature Output

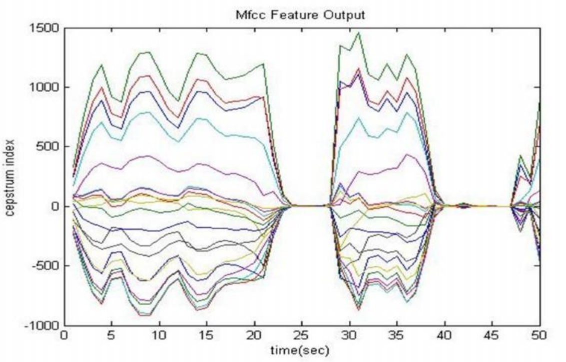


Figure 8 Mel-frequency cepstrum coefficients

The table 4 describes the performance measures of the various algorithms. From above table KNN classifier considered as the best classifier since it leads to high accuracy when compared to another classifier.

Table 4 Performance measure of proposed algorithm

Method	Accuracy (%)
Artificial Neural segmentation Network with manual	73.25
Hidden Markov segmentation Model with manual	80
KNN Classifier with VAD segmentation	85.71

VI. CONCLUSIONS

The main aim of this project is classification and analysis of stuttered speech signal using Mel frequency cepstrum coefficients based feature extraction and K-nn classifier framework. Using K-nn classifier classification of fluent and non fluent is done and obtained the accuracy of 85.71%. For training phase, hundred and fifteen speechsegmented samples are used. Analysis of signal such as stages of stuttering, , percentage disfluency, syllable per minute, and number of dysfluencies are calculated effectively.

References

[1] Junbo Zhang, Bin Dong, Yonghong Yan, "A Computer Assist Algorithm To Detect Repetitive Stuttering Automatically" Internal Conference On Asian Language Processing, Vol. 22, No. 1, PP. 249-252, 2013)

[2] Vibha Tiwari, "MFCC And Its Applications In Speaker Recognition", International Journal On Emerging Technologies, PP. 19-22, 2010.

[3] Manpreet Kaur, Amarpreet Kaur, IJET, "A Different Methods Of Segmenting A Continuos Speech Signal Into Basic Units", IJECS, PP. 3184-3183, Vol. 22, November 2013.

[4] S. Poornima, J. Satheesh Kumar, "Feature Extraction And Signal Classification Methods For Stuttering Speech Analysis", IJMCSA, Vol. No. 1, PP. 6-11, November 2013.

[5] K.M Ravikumar, Balakrishna Reddy, R. Rajagopal, "Automatic Detection Of Syllable Repetition In Read Speech For Objective Assessment Of Stuttered Disfluencies", International Journal Of Electrical, Computer, Electronic And Communication Engineering, Vol. 2, No. 10, 2008.

[6] Adel Belouchrani, Karim Abed-Meraim, Boualem Boashash, "Time Frequency And Array Processing Of Non-Stationary Signals", EURASIP Journal On Advances In Signal Processing 2012,2012:230.

[7] Alfredo Maesa, Febio Garzia, Michele Scarpiniti, Roberto Cusani, "Text Independent Automatic Speech Recognition System Using Mel-Frequency Cepstrum Coefficient And Gaussian Mixture Models", Journal Of Information Security, 2012, 3, 335-340

[8] Bachu R.G., Kopparthi S.,Adapa B., Barkana B.D. "Separation Of Voiced And Unvoiced Using Zero Crossing Rate And Energy Of The Speech Signal", Advanced Techniques In Computer Science And Software Engineering 2010,Pp:279-282.

[9] Balazs Bank, Federico Avanzini, Gianpaolo Borin, Giovanni De Polli, Federico Fontana, Devide Recchess, "Physically Informed Signal Processing Methods For Piano Sound Synthesis: A Research Review", EURASIP Journal On Applied Signal Processing 2003: 10, 941-952.

[10] David Rybach, Christian Gollan, Ralf Schluter, Hermann Ney, "Audio SegmentationFor Speech Recognition Using Segment Features", Defence Advanced Research Projects Agency(DARPA), Under Contract No.HR0011-06-C-0023.

[11] G Yen Fcok, Hariharan Muthusamy, Lim Sin Chee, Sazail, Bin Yaacob, Abdul Harnid Bin Adom, "Comparision Of Speech Parameterization Techniques ForClassification Of Speech Dysfluencies", Turkish Journal Of Electrical Engineering And Computer Science, 2013.

[12] Shuping Ran And J. Bruce Miller, "Exploring The Phonetic Structure Of Speech Signal Using Multi Layer Perceptrons " In Proceeding Of SST , 1990, Pp 22-27 .

[13] Singh Parminder, Gurpreet Singh Lehal, "Syllable Based Text-To-Speech Synthesizer For Punjabi", In 10th Punjab Science Congress, Feb 2007.

[14] Singh Prem, “Sidhantik Bhasha Vigeyan”, Patiala: Madan Publications, Pp. 391..

[15] Wagner W, ”Automatic Labelling Of Continuous Speech With A Given Phonetic Transcription Using Dynamic Programming Algorithm”, IEEE Conf. On Acoustics, Speech And Signal Processing, 1981, Pp 1156-1159

[16] Bridle J.S. And Chamberlian R.M.,”Automatic Labelling Of Speech Using Synthesis By Rule And Non Linear Time Alignment,” In Speech Communication, 1983, Pp 187-189

[17] Prasad V. K., Nagarajan T, Murthy H A ,” Automatic Segmentation Of Continuous Speech Using Minimum Phase Group Delay Functions,” In Speech Communication 42 , 2004, Pp 1883–1886.

[18] Singh Prem, “Sidhantik Bhasha Vigeyan”, Patiala: Madan Publications, Pp. 391.

[19] Vikyath Narayan K N And Meharunnisa S.P, “Detection And Analysis Of Stuttered Speech” ,International Journal Of Advanced Research In Electronics 7 Communication Engineering, Vol. 5, Issue 4, Aprol 2016

[20] Neeta Awasthy, J.P.Saini And D.S. Chauhan, “Spectral Analysis Of Speech: A New Technique,” International Journal Of Signal Processing, Vol. 2, No. 1, Pp. 19-29, 2006.

[21] Adel Belouchrani, Karim Abed-Meraim, Boualem Bosash, “Time Frequncy And Array Processsing Of Non-Stationary Signals”, EURASIP Journal 2012, 2012:230.

[22] Alfredo Maesa, Febio Garzia, Michele Scarpiniti, Roberto Cusani, “Text Independent Automatic Speech Recognition System Using Mel-Frequency Cepstrum Coefficient And Gaussian Mixture Models” Journal Of Information Security 2012, 3, 335-340.

[23] B. Guitar, Stuttering: An Integrated Approach To Its Nature And Treatment, Lippincott Williams & Wilkins, 2013.

[24] N. B. Ratner, B. Macwhinney, Fluency Bank: A New Resource For Fluency Research And Practice, Journal Of Fluency Disorders 56 (2018) 69.

[25] A. C. Etchell, O. Civier, K. J. Ballard, P. F. Sowman, A Systematic Literature Review Of Neuroimaging Research On Developmental Stuttering Between 1995 And 2016, Journal Of Fluency Disorders 55 (2018) 6–45

[26] NIDCD,Stuttering. URL :Https://Www.Nidcd.Nih.Gov/Health/Stuttering/