ISSN 2278-3091

Volume 10, No.3, May - June 2021

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse521032021.pdf

https://doi.org/10.30534/ijatcse/2021/521032021

Stop gap removal using spectral parameters for stuttered speech signal

K B Drakshayini, Dr Anusuya M A
Assistant Prof, VVIET, Mysuru, India, drakshakb@gmail.com
Associate Professor, SJCE, Mysuru, India,anusuya_ma@sjce.ac.in

ABSTRACT

Stuttering is an involuntary disturbance in the fluent flow of speech characterized by disfluencies such as stop gaps, sound or syllable repetition or prolongation. There are high proportion of stop gaps in stuttering. This work presents automatic removal of stop gaps using combination of spectral parameters such as spectral energy, centroid, Entropy and Zero crossing rate. A method for detecting and removing stop gaps based on threshold is discussed in this paper.

Keywords: Stop gaps, Spectral parameters, feature extraction, FCM.

1. INTRODUCTION

With the advancement of digital computing and signal processing, the problem of stuttered speech recognition was considered seriously. The development was accompanied with an increased awareness of the advantages of conversational systems. The range of the possible applications is extensive and includes voice-controlled applications, command control applications, automation of operator assisted services, and voice recognition assistances for the handicapped [1]. Stuttered Speech signals usually contains stopgaps(both unfilled and filled). Unfilled stop gaps are silent where in filled stopgaps contains Low frequency band of energy[14].

In speech analysis it is necessary to detect stopgaps in order to spot clear speech segments. Spectral parameters are selected for stop gaps removal process. The reasons selecting these specific features are:
1) The spectral energy of the voiced segments is larger than the energy of the silent segments.
2) The spectral centroid for the voiced segments is larger than the silent segments [2].
3)Entropy is a measure of disorder, regions of voiced speech have lower entropy compared to regions of unvoiced speech.
4)The zero-crossing count of silence is expected to be lower than for unvoiced speech, but quite comparable to that for voiced speech

The paper is systematized as follows. Section II presents existing work on stop gap removal process, section III explains datasets used for experimentation, section IV describes the usage of spectral parameters in stopgap removal and section V presents the results of the proposed method. Conclusion and future enhancement are discussed in section VI.

2. LITERATURE REVIEW

In 1997 Howell [3-4] worked on database of 12 children stutters speech using UCLASS dataset. The features extracted are duration, energy peaks, spectral features are at word level and sub word level and obtained accuracy of 78.01% using ANN classifier. In 2000 Noth [5] experiments were conducted on 16 recordings of stuttered and non-stuttered signals for the features like Duration and frequency of disfluent portions, speaking rate, using Hidden Markov models (HMMs) as classifier. In 2003 Czyzewski [6] considered 6 normal speech samples and 6 stop-gap speech samples. The Frequency 1st to 3rd formant's frequencies and its amplitude are considered as features. ANNs and rough set are used as classifiers and achieved 73.25 and 90. % recognition rates. In 2015 V Naveen Kumar [7] worked on stuttered database using MFCC feature extraction method. and K-Means clustering is used to remove the stop gaps using the features energy and zero crossing rate. RMS energy maximum energy, minimum energy-based stop gap removal was proposed by Kirill Sakhnov [8] and Muhammad Asadullah [9]. In 2017 Formants, pitch, intensity-based stop gap removal was proposed by Pierre Arbajian[10].

3. DATABASE

The speech signals are collected from ordinary environment where background noise is present. These background noise signals are pre-processed using audacity software. Speech signals are acquired using Monochannel, 8000 KHz samples with 8-bit quantisation. Totally 100 signals are considered for the training and testing purpose. Automatic Removal of stop gaps has been done for all 90 samples (each utterance is 10 times). An Adult male and female were asked to utter numbers (1 through 10) 10 times each for recording. Among 100 speech signals only 90 signals are selected for training. For testing 10 signals (from each number) are taken independently from the trained set.

4. METHODOLOGY

During stop gap removal the spectral parameters adopted are signal energy and spectral centroid.

A. *Signal Energy*

Let $x_i(n)$, n = 1…, N the audio samples of the ith frame, of length N. For individual frame i the energy is calculated according to the equation (1)

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

This simple feature can be used for detecting stop gaps in audio signals, but also for discriminating between audio classes this feature can be used.

B. *Spectral centroid*

The spectral centroid is defined as the centre of “gravity” of its spectrum, forindividual framei the centroid is calculated according to the equation (2)

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

k = 1..., N, are the Discrete Fourier Transform (DFT) coefficients of the ith short-term frame, where N is the frame length. It indicates the sequence of spectral centroid is highly variated for speech segments [11].

C. *Zero Crossing rate*

The zero-crossing count is an indicator of the frequency at which the energy is concentrated in the signal spectrum. Voiced speech is produced as a result of excitation of the vocal tract by the periodic flow of air at the glottis and usually shows a low zero crossing count. Unvoiced speech is produced due to excitation of the vocal tract by the noise-like source at a point of constriction in the interior of the vocal tract and shows a high zero crossing count. The zero-crossing count of silence is expected to be lower than for unvoiced speech, but quite comparable to that for voiced speech[15].

A definition for zero crossing rate is

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

Where

$$\text{sgn}[x(n) = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases} \quad (4)$$

and

$$w(n) = \begin{cases} \frac{1}{2N} & for\ 0 \leq n \leq N-1 \\ 0 & for, otherwise \end{cases} \quad (5)$$

N is the duration of the window used in the method.

D. *Spectral Entropy*

Spectral entropy has been used successfully in voiced/unvoiced decisions for automatic speech recognition. Because entropy is a measure of disorder, regions of voiced speech have lower entropy compared to regions of unvoiced speech.Spectral entropy measures the peakness of the spectrum. Frame with Low energy will be having high entropy.

$$entropy = \frac{-\sum_{k=b_1}^{b_2} s_k \log(s_k)}{\log(b_2 - b_1)} \quad (6)$$

where

- s_k is the spectral value at frame k. The magnitude spectrum and power spectrum are both commonly used.
- $b1$ and $b2$ are the band edges, over which to calculate the spectral entropy.

E. *Simple threshold-based stop gap removal Algorithm*

As the Signal energy and spectral centroid feature sequences are calculated, a threshold-based algorithm is useful, in order to extract the speech segments. At first stage, energy and centroid thresholds are computed.The following process is used for energy and centroid thresholdcalculation, for each feature sequence:

1. Compute the histogram of the feature sequence’s values.
2. Apply a smoothing filter on the histogram.
3. Detect the histogram’s local maxima.
4. Let M1 and M2 be the positions of the first and second local maxima respectively.

The threshold value is computed using equation (3)

$$T = \frac{W.M_1 + M_2}{W+1}(7)$$

Weight(W) is a user-defined parameter. Large values of W obviously lead to threshold values closer to M1. The above process is executed for both feature sequences, leading to two thresholds: T1 and T2, based on the energy sequence and the spectral centroid sequence respectively. As long as the two thresholds have been estimated, segments are formed by successive frames for which the respective feature values (for both feature sequences) are larger than the computed thresholds.

Weights are the normalized energy of each frequency component in that frame. For detecting frequency peaks in the frame which correspond to the location of formants or pitch frequencies weights are used [12]. Large values of W obviously lead to threshold values closer to M1. Identification of local maxima and minima, is important for holding certain signal properties. True peaks are identified by Local maxima [13].

Figure 1 shows the block diagram for stop gap removal process of stuttered speech signal. Pre-processed signals are passed to threshold-based stop gap removal algorithm. MFCC and FCM techniques are used for feature extraction and clustering process.

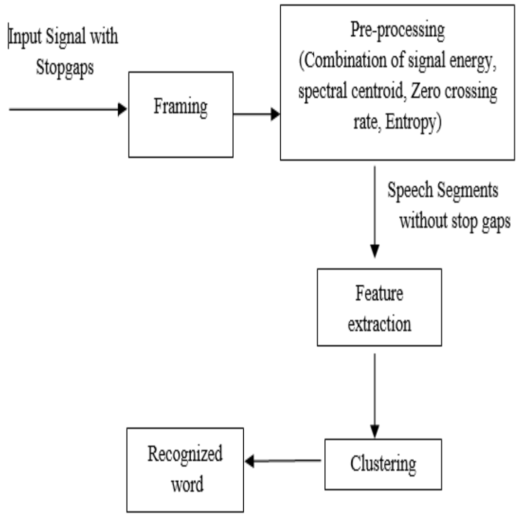
Input Signal with Stopgaps
Framing
Pre-processing (Combination of signal energy, spectral centroid, Zero crossing rate, Entropy)
Speech Segments without stop gaps
Feature extraction
Clustering
Recognized word

Figure 1: Flowchart of stop gap removal process

F. Clustering using Fuzzy C means (FCM)

The Fuzzy set theory is a generalized form of set theory. This theory is used to represent the uncertainty in the information. Since speech has uncertainty in itself, this approach is most suitable for speech signal processing application. Signal uttered from the same person for more than one time will not be exactly same. Hence we can say speech is fuzzy in nature[16]. It works as follows

Step 1: Randomly initialize the clusters centers

$$J_{KM}(X;V) = \sum_{i=1}^{C}\sum_{j=1}^{n} D_{ij}^{2} \qquad (8)$$

Step 2: Create the distance matrix from a data point to each of the cluster center using Euclidean distance using Eq(9).

$$V_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \qquad (9)$$

Step 3: The membership matrix is computed using fuzzification parameter with Eq(10).

$$u_{ij=} \frac{1}{\sum_{k=1}^{c}\left(\frac{D_{ijA}}{D_{kjA}}\right)^{\frac{2}{(m-1)}}}; 1 \le i \le c, 1 \le j \le n \qquad (10)$$

$$J_{KM(}(U, \Lambda; X) = \sum_{i=1}^{C}\sum_{t=1}^{T} u_{it}^{m}\, d_{it}^{2} \sum_{j=1}^{n} D_{ij}^{2} \qquad (11)$$

Step 4: Values of the U_{ij}matrix should be less than or equal to one ($U_{ij} \le 1$)
Step 5: Compute new centroid's
Step 6: Optimize cluster centers by generating new centroids
Step 7: Cluster assignment for the data points

Where x1- data vector, Vi - centroids of fuzzy clusters, c - number of fuzzy clusters, m - fuzzification parameter, U - assigns Fuzzy membership value to each sample indicating the membership value from one data sample to the n th cluster, € - stopping criteria, Dij - distance measure and n - number of data points.

5. EXPERIMENTAL RESULTS

Automatic stop gap removal method is applied to pre-processed signals. Speech signals are categorized into voiced speech and unvoiced speech. Energy and centroids are used to differentiate voiced and unvoiced signals. Figure 2 and 3 shows speech signals before and after removing stop gaps.

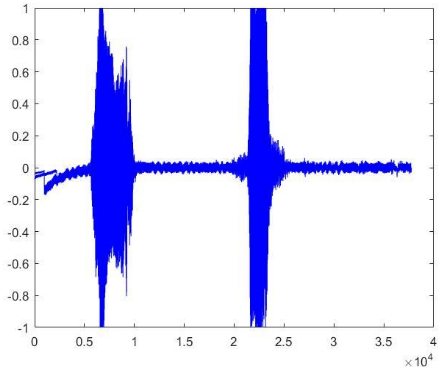

Figure 2: Speech signal before stop gap removal

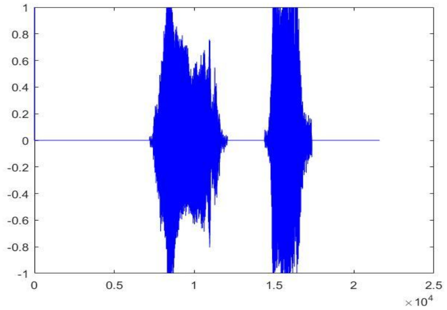

Figure 3: Speech signal after stop gap removal

In order to extract feature sequences MFCC feature extraction is applied and to translate huge amount of vector from a multidimensional space to a predefined number of clusters using VQ. Each of clusters is defined by its central vector or centroid. According to the Euclidian distance function, K-Means algorithm clusters the data in to K groups then assigns objects to their closest cluster gives. Experimental results are conducted by varying the weights, step size and window size. Result are tabulated separately by varying window size as shown in Table 1. Lower the weight higher the recognition rate by removing the stop gaps. Table 2 presents the results by varying step size and window size simultaneously. Smaller the step and window size the recognition rate is more. Smaller the weight, step size and window the stop gaps can be identified and removed better since signals are better analysed at all the levels of MFCC feature extraction.

Table 1: Influence of weight parameter on recognition rate

Weight(w)	Recognition rate (%)
30	50
25	50
20	50
15	50
10	70
5	**80**

Table 2: Effect of step and window size on recognition rate.

	step size	window size	Recognition rate (%)
Weight=5	0.1	0.1	50
	0.1	0.05	50
	0.05	**0.05**	**80**

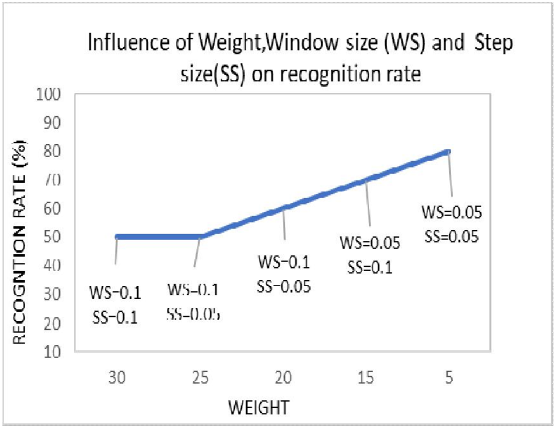

Figure 4: Influence of Weight, Window size (WS) and Step size (SS) on recognition rate.

Table 3: Influence of Different spectral parameters on stop gap removal process (Frame duration 25 milliseconds)

	Parameter	No of stop gaps	No of voiced frames	Percentage of stop gap removal
Frame duration 25 milli seconds **Total no of frames =83**	ZCR	45	38	45.7831
	Centroid	81	2	2.4096
	Energy	43	40	48.1928
	Entropy	66	17	20.4819
	ZCR+Centroid + Energy +Entropy	26	57	68.6747

Table 4: Influence of Different spectral parameters on stop gap removal process (Frame duration 250 milliseconds for detail frame analysis)

	Paramet er	No of stop gap	No. of voiced frames	Frame No. of Voiced region
Frame duration 250 ms **TotalNo.of frames=8**	Centrid	3	5	1,2,5,7
	Energy	2	6	2,3,4,5,6
	ZCR	7	1	7
	Entropy	3	5	3,4,5,6,
	ZCR+ Centroid + Energy +Entropy	1	7	1,2,3,4,5, 6,7

Table 5: Performance analysis of spectral parameters in stop gap removal in terms of signal length

Parameters (frame duration 250 milli seconds)	Signal length before stop gap removal (in milli seconds)	Signal length After stop gap removal (in milli seconds)
ZCR	92160	44100
Entropy		11025
Energy		55125
Centroid		44100
ZCR+Centroid+Energy +Entropy		77175

Table 1 and 2 shows the effect of weight, step and window size parameters on recognition rate. Figure 4 shows the influence of weight, window size and step size feature parameters in increasing the stuttered speech recognition rate by removing stopgaps with non-overlapping, frames with 8kz sampling frequency with weight =5, step=0.01 and window=0.01 with 12MFCC will results in better recognition rate. Performance evaluation of spectral parameters on stop gap removal has been tabulated in Table 3 and 4. Table 5 tabulates assessment of performance of spectral parameters and their contribution towards stop gap removal.

6.CONCLUSION AND FUTURE ENHANCEMENT

In this work, the influence of combination of spectral parameters helps in the improvement of stuttered speech recognition and FCM performing better in clustering to improve recognition rate. Stuttering is automatically removed by spectral features. Stop gap removal process is affected by weight, step and window size parameters. The system can be further improved for removal of stop gaps or prolongation disorders based on pitch and epoch features using PSOLA and ESOLA methods for disorder speech signals.

REFERENCES

1.Dr.M.A.Anusuya,Dr.S.K.Katti, “Speaker Independent Kannada Speech Recognition using Vector quantization “,2012

2. Theodoros Giannakopoulos, “A method for silence removal and segmentation of speech signals, implemented in MATLAB”

3. 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.

4. 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.

5. E. Noth, H. Niemann, T. Haderlein, M. Decher, “Automatic stuttering Recognition using hidden Markov models", Proceedings of the International Conference on Spoken Language Processing, Vol. 4, pp. 65,68, 2000.

6. A. Czyzewski, A. Kaczmarek, B. Kostek, “Intelligent processing of stuttered speech", Intelligent Information Systems, Vol. 21, 2003.

7. V.NaveenKumar, “Design and Implementation of sailent pause stuttered speech recognition system”, 2015.

8. Kirill Sakhnov,”Dynamical Energy-Based Speech/Silence Detector for Speech Enhancement Applications”,2009

9. Muhammad Asadullah, Shibli Nisar “A silence removal and endpoint detection approach for speech processing “, 2016

10. Pierre Arbajian,” Segment-Removal Based Stuttered Speech Remediation”,2014

11. T. Giannakopoulos, “Study and application of acoustic information for the detection of harmful content, and fusion with visual information,” Ph.D. dissertation, Dept of Informatics and Telecommunications, University of Athens, Greece, 2009.

12. PuneetKumarMongia and R.K.Sharma,” Estimation and Statistical Analysis of Human Voice Parameters to Investigate the Influence of Psychological Stress and to Determine the Vocal Tract Transfer Function of an Individual”,2014

13. K. K. L. B. Adikaram,” Non-Parametric Local Maxima and Minima Finder with Filtering Techniques for Bioprocess”, 2016.

14. K.B.Drakshayini, Dr. Anusuya M.A “Vector Quantization for stuttered speech Recognition”, Journal of Data Mining and Management [ISSN: 2456-9437], Vol 4,May 2018.

15. 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”,2008

16. H.Y.Vani, Dr.M.A. Anusuya and Dr.M.L. Chayadevi“Fuzzy Clustering Algorithms - Comparative Studies for Noisy speech signals”, 2019