Stuttered Speech Recognition And Classification Using Enhanced Kamnan Filter And Neural Network

[1]B.Vaidianathan, [2]S.Arulselvi, [3]B.Karthik
{vaidia5000@gmail.com1, arulselvi2003@gmail.com2, karthikguru33@gmail.com3}

Research Scholar, Electronics and Communication Engineering, BIST, Bharath Institute of Higher Education and Research, Chennai, India.[1],
Associate Professor, Electronics and Communication Engineering, BIST, Bharath Institute of Higher Education and Research,Chennai,India.[2,3]

Abstract. Stuttering or stammering assessment is one of the vital factors in speech recognition algorithms. To reconstruct the stuttered speech into spontaneous speech it is necessary to detect and correct the features influencing the speech signal. In this paper the speech signal is processed based on the disturbances created by acoustic effects like pauses and noises made both externally and internally. To eliminate the effects of noise on speech signal an Enhanced Kalman Filter is introduced here and its performance along with various filters are studied and compared based on the parameters like Mean Square Error (MSE), Mean Absolute Error (MAE), SNR ratio, Peak Signal to Noise ratio and Cross correlation. Then based on the extracted features classification of the speech signal is carried out using Convolutional Neural Network (CNN) algorithm of Deep learning technique.

Keywords: Stuttering Enhanced Kalman Filter, Mean Square Error, Mean Absolute Error, Signal to Noise Ratio, Convolutional Neural Network.

1 Introduction

Speech is the most habitual way of communication for human beings. It is the physiological movement of air through vocal chords, tongue, jaw, teeth, lips, and palate to produce sound. When a person is unable to produce speech sounds fluently or perfectly which makes the listener helpless to recognize the words then it is termed as speech disorder. Almost seventy billion people undergo speech disorders like stuttering, mumbling, apraxia, cluttering, dysarthria, lisping, whispering etc., These disorders are due to damage to brain's motor nerves for speech, paralysis of speech muscles, Cleft Lip and Palate and some more.

Stuttering is a disorder found in speech pathology and commonly seen more in male compared to females. The normal flow of verbal communication is disturbed by occurrences of repetitions, long silences and prolongations of sounds. Interjections, revisions, incomplete phrases, repetition, Prolonged sounds, broken words are most prevalent stuttering types.

In recent years research for signal processing in acoustic science is popular as variety of application including medical application, robotics, home automation, defense, machine translation requires speech recognition. In speech recognition, digitizing of the signal is performed by separating from noise, detecting and comparing the phoneme to predict the word to determine the whole speech sequence.

I3CAC 2021, June 07-08, Chennai, India

DOI 10.4108/eai.7-6-2021.2308659

2 Literature Survey

Sakshi Gupta et al.,(2020) has structured an automatic speech disorder recognition method which is used to disclose the disorders like prolongation of a word and also repetition of a syllable, word or phrase from the speech given. Here the database for input speech signal was obtained from the archive of a university. These signals are pre-processed and segmented considering the above disorders. Then the extraction of both static and dynamic sound's features are done by using the WMFCC feature extraction algorithm. Next it is preceded with the classification of the signal by using the Bi-LSTM network of deep learning. Finally the proposed model is compared with the unidirectional LSTM model and the results portrays that the proposed method as highly accurate. [1]

The technology that is used to convert an audio speech into text information is called as Automatic Speech Recognition (ASR) system. But the dysfluencies like addition or elongation of words or syllables in the speech becomes a hurdle for the ASR system. For the proper function of this system two main operations namely classification and testing of the input have to be done. Here **Girirajan** et al., (2020) has proposed LSTM algorithm for the classification of the signal in MATLAB platform. The paper's main focus was to detect the mention abnormalities in the speech and classify them as normal and abnormal speech. This shows an slight hike in terms of efficiency when compared to other popular classification algorithms. [2]

Here **Mohammed Sidi Yakoub** et al.,(2020) formulated a novel approach for improving the detecting process of dysarthric speech. The EMDH technique is applied in the pre processing of the signal for determining and choosing the mode functions from the disturbed data to remodel the original signal. Then the features withdrawn are utilized for classification by merging the algorithms EMDH and CNN together. Compared to the standard system this new approach shows satisfactory results. [3]

Arjun et al., (2019), devised a method which is used to correct the stutters found in a speech signal. To avoid the recurrence of same word, the speech is sampled into individual words by using appropriate thresholding and speech energy techniques. For discarding the long pauses between the words, the speech signal is segmented into frames of 50ms windows. By using the LPC and MFCC methods every two subsequent frames are checked for feature extractions, here it is the similarity index. Based on correlation of these extracted details, normal pauses are retained and others are discarded to get speech signal in appropriate format. This type of speech signal processing method is regarded as simple and robust as it utilizes the thresholding and correlating concepts. [4]

Numerous speech enhancement algorithms have been designed to transform a corrupted speech signal into an intelligent form of signal. Mostly the corruption might have been caused by various types of noises. **Nasir Saleem** et al., (2019) have described about a study describing about the potentiality of various algorithms in enhancing the speech signal by removal of the distortions. Here the unsupervised type of SCS algorithms like WF, SigSub, MMSE, EMD etc.. are considered for surveying. This survey concludes that these algorithms potentiality is better in noise reduction to show the speech quality than in the speech intelligence. When the estimation of the speech is carried out too hard by these methods it may result in the content loss, so further research in this area is required. [5]

Selvaraj et al., (2019) suggested a method to enhance the speech signal which shows adverse effect by the non-stationary noises low SNR. Here the sliding window concept is combined with the EMDH algorithm to form the SWEMDH method, which is used to improve the speech enhancement process. The intrinsic mode function identified by the Hurst

exponent model is corrupted by the noises which results in the complexity of signal reconstruction. So the analysis of EMD is preceded by selecting the sliding window with respect to time frame. Here the sifting iteration's count is calculated by decomposing the consecutive windows and finding the mean sifting steps. Hence the time complexity is improved by the proposed system. [6]

3 Proposed System

The need of computer-aided speech recognition systems has led to the development of many algorithms using various methodologies. This paper focuses on improving the filtration of noise from the original speech signal by introducing an Enhanced Kalman filter and classifying the signal by employing CNN algorithm using Matlab. The process of assessing and classifying of the disfluency in speech signal is initialized by extracting the stuttered voice signal.

A.Preprocessing: The voice signal is always accompanied by noise which results in degradation of the signal. The factors that create interference may be due to disorders in organs, acoustic background, sensor positions, reverberation effect etc. Here various filters are surveyed along with the proposed Enhanced Kalman filter.

***EMD Filter*:** Empirical mode decomposition filter is a non – stationary and nonlinear adaptive method which does not require any basic function. The decomposition of the signals into Intrinsic Mode Functions (IMF) is obtained by detecting the local maxima and minima followed by its envelope value and then subtracting the mean of the envelope extremities from the input data. This is conducted until non IMF residual value is attained referred to as the sifting process. The EMD is expressed as total of IMF and residue.

$$r_n + \sum_{i=1}^{n} \mathrm{imf}_i(t) = x(t) \qquad \text{-----(1)}$$

where r – residue, i- index of mode. This is used to remove low noise frequencies.

DWT Decomposition filter: The discrete wavelet transformation is based on a series of filters. The signal 'x' is dilated by sub sampling and passed through low pass filter where it is convolved with the impulse response 'g' as follows

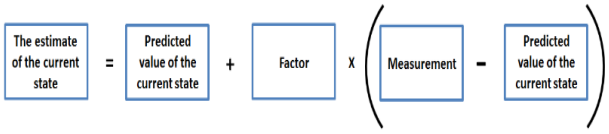
$$y[n] = x[n] * g[n] = \sum_{k=-\infty}^{\infty} x[k]g[n-k] \qquad \text{-----(2)}$$

Then it is decomposed by passing the high pass filter 'h'. The two filters are combined together to be called as quadrature mirror filter. Half the frequencies would have been filtered by now so by using Nyquist rule half of the samples can be removed. The filter output is sub sampled by two the process is repeated with new series of filters.

DWT Adaptive filters: The DWT shifts and dilates signals into a small number of coefficients of large magnitude. Dilations follow from the "discarding" features of wavelets, assures that the low-order polynomial signal's wavelet coefficients as zero. Practically, the signal will not be polynomial alone but, approximated by a polynomial function. So the DWT

is modified to adapt to match the signal. This adaptive DWT have the potential to improve the transforming process of denoising by providing higher efficiency and less computation.

Kalman filter: It is a recursive approach used to estimate hidden variable depending on the inaccurate measurements over time with statistical noise. It requires little computational power and has multiple applications which include navigation and control, time series analysis of signal processing. It works on two steps: Prediction – to provide estimate of current variables and Updation – of these values by weighted average of the outputs of next measurement. The equation for this updation state means as follow.



---(3)

Enhanced Kalman Filter: The proposed filter is the modified version of the KF. It is not confined to linearity but can work on nonlinear functions. It linearizes the signal for the current state estimate and use linear Kalman filter to predict the next estimate. The state transition and observation models are differentiable functions and can be expressed as,

$$\boldsymbol{x}_k = \boldsymbol{f}(\boldsymbol{x}_{k-1}, \boldsymbol{u}_k) + \boldsymbol{w}_k$$
$$\boldsymbol{z}_k = \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{v}_k \qquad \text{----------(4)}$$

where, w_k – process noise, v_k - observation noise, u_k – control vector

The prediction of state from the previous estimate is calculated by function 'f' and the calculation of predicted measurement is done with the help of function 'h'. As the application of these functions to the covariance is not possible the Jacobian of each step is analyzed with current predicted states.

B. Sampling: Sampling or segmenting is a process to divide a continuous speech into smaller units like words, syllable etc., The sampling of speech signal is based on the stuttering rate of various stuttering disorders. Prolongation (lengthy occurrence of a word), Repetition (repeating of syllable or word), Long pause (long silence between words) are the disorders considered in this process. The preprocessed signals are marked and segmented based on the disorder by adding, cancelling and dismissing some words. Automatic segmentations can be carried out by using any of the method like Fourier Transform, Short Term Energy, Minimum Phase Group Delay Method, Word Chopper, Wavelet Method and some more.

C. Feature Extraction: The objective is the conversion of acoustic signal into recognizable sequence of acoustic feature vectors. The segmented sample of the stuttered speech signal is arranged into differential sets for training, validation, and testing. The features are extracted based on the following parameters to calculate the efficiency of the algorithm:

Mean square Error: Error signal (e) is the difference between the input signal x(t) and the reconstructed signal. The sum of the squared average of the error signal is termed as MSE. It is expressed as

$$\text{MSE} = \frac{1}{l}\sum_{i=1}^{n} e_i^2, \quad \text{where} \quad e = \hat{x}(t) - x(t). \qquad \text{----(5)}$$

Mean Absolute Error: It is the amount of error between the input and reconstructed signal.

$$\mathrm{MAE} = \frac{1}{l}\sum_{i=1}^{n} e_i. \quad \text{--------(6)}$$

Signal to Noise Ratio: It is defined as the ratio of the speech signal to the noise present in the speech signal.

$$\mathrm{SNR\ [dB]} = 10\log_{10}\left(\frac{P_{\mathrm{signal}}}{P_{\mathrm{noise}}}\right) \quad \text{------(7)}$$

Peak Signal to Noise Ratio: It is the proportion of maximum attainable signal power to the distorting noise power and can be expressed in terms of MSE.

$$\mathrm{PSNR\ [dB]} = 10\log_{10}\frac{255^2}{\mathrm{MSE}}. \quad \text{-------(8)}$$

Cross correlation: It is used to determine the rate of relationship between two different entities f and g is given as

$$(f \star g)[n] \triangleq \sum_{m=0}^{N-1} \overline{f[m]}\, g[(m+n)_{\mathrm{mod}\ N}] \quad \text{------(9)}$$

D. Classification: The classification of the speech signal is performed by the widely popular Convolution Neural Network architecture. The application of CNN to acoustics modules enhance the performance of the process compared to various other classification algorithms. The CNN is a nonlinear function framed with several layers that includes convolutional, hidden and pooling. The features are extracted by the initial convolutional layers. From the featured information the pooling layer down sample the data by keeping important values and discards the insignificant values. Next the estimation of the class conditional probability is obtained by CNN, which is used to find the emission scaled-similarity to classify the data based on the value. The advantage of using CNN in speech recognition are based on the properties weight sharing, filtering and pooling which are used to enhance the overall performance of the system.

4 Results and Discussions

Now let us see the results that is been obtained from a stuttered speech. Here a stuttered speech is taken as input signal. Then the preprocessing stage takes place where the speech signal will be analyzed and the noise present in the signal has to be removed. Here our proposed Enhanced Kalman filter is used and the noise present in the signal is completely removed. Then according to the concept the pause, external parameters which affects the signal is found and eliminated. At last we will be using Convolution Neural Network where the signal will be classified and the stuttered word which is been pronounced by the particular person is analyzed and classified correctly.

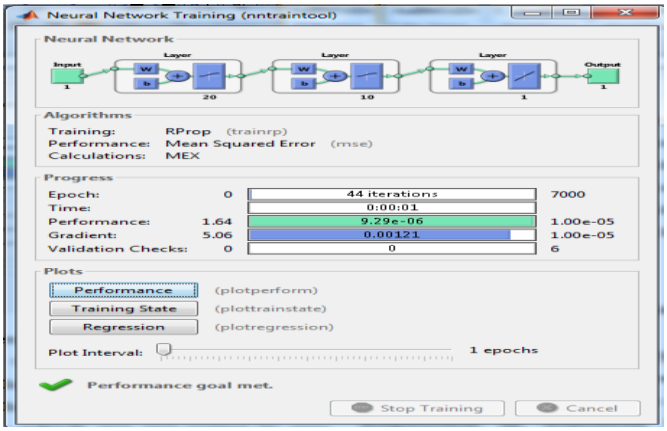


Figure 1: Neural Network Performance

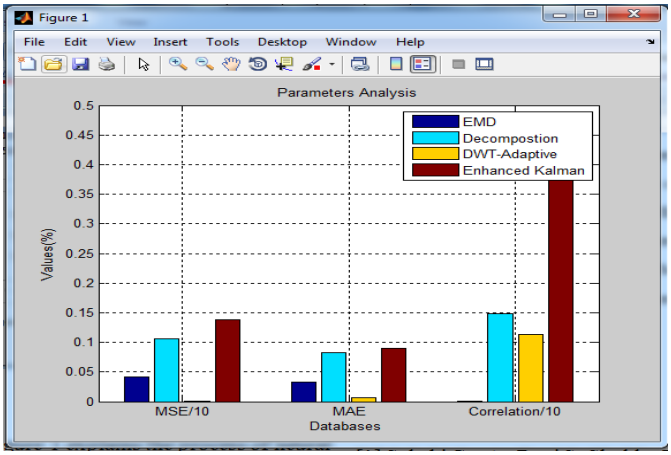


Figure 2: Parameter Analysis

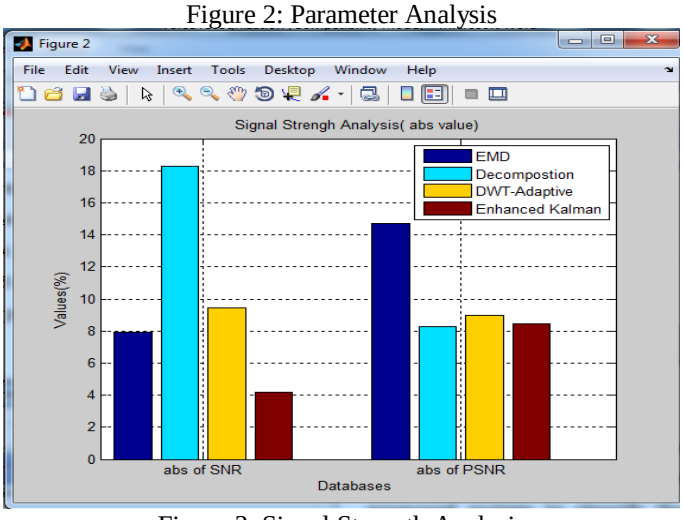


Figure 3: Signal Strength Analysis

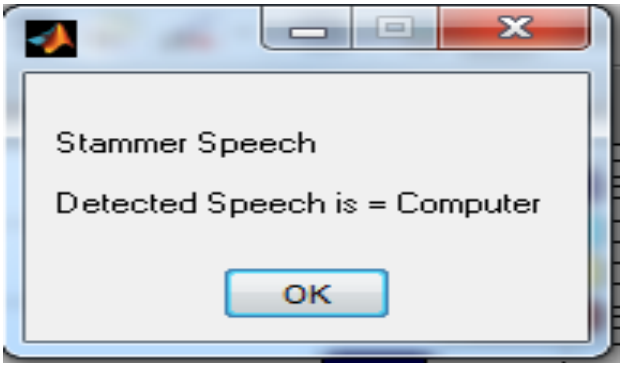


Figure 4: Classified Stuttered word

The above figure 1 explains the process of neural network performance that is been used in our proposed system to classify the stuttered words. Figure 2 represents the bar chart of parameter analysis of all 4 filters which we are using in our research. The parameters such as MSE, MAE and correlation have been measured and compared with existing filters. It states that the results obtained from Enhanced Kalman filter is good than other filters. Figure 3 represents the signal strength analysis where we will calculate the Absolute Value of SNR and PSNR. It is found that the absolute value obtained from Enhanced Kalman filter is better than other existing filters. The last Figure 4 represents the output of stuttered word that is been pronounced by the concern person.

Conclusion

In this paper an automated method to recognize and classify the stuttered speech signal is described based on Convolutional Neural Network by introducing an Enhanced version of Kalman filter (EKF) for nonlinear noise removal. The contamination of the stuttered speech signal by noise is filtered by using various types of filters and their performance are compared with the described EKF. From the obtained Matlab simulated results based on various parameters like MSE, PMSE, SNR, PSNR the proposed Enhanced KF executes satisfactory results.

References

[1] Sakshi Gupta, Ravi S. Shukla, Rajesh K. Shukla, Rajesh Verma, 2020, “Deep Learning Bidirectional LSTM based Detection of Prolongation and Repetition in Stuttered Speech using Weighted MFCC”, International Journal of Advanced Computer Science and Applications, Vol. 11, No. 9.

[2] S.Girirajan, R.Sangeetha, T.Preethi, A.Chinnappa, 2020, “Automatic Speech Recognition with Stuttering Speech Removal using Long Short-Term Memory (LSTM)”, International Journal of Recent Technology and Engineering, ISSN: 2277-3878, Volume-8 Issue-5

[3] Mohammed Sidi Yakoub, Sid-ahmed Selouani, Brahim-Fares Zaidi and Asma Bouchair, 2020, “Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network ”, EURASIP Journal on Audio, Speech, and Music Processing, https://doi.org/10.1186/s13636-019-0169-5

[4] Arjun K N, Karthik S, Kamalnath D, Pranavi Chanda, Shikha Tripathi, 2019, "Automatic Correction of Stutter in Disfluent Speech", Third International Conference on Computing and Network Communications, Procedia Computer Science, Vol. 171,pp. 1363–1370

[5] Karthik, B., Kumar, T.K., Dorairangaswamy, M.A., Logashanmugam, E., Removal of high density salt and pepper noise through modified cascaded filter, Middle - East Journal of Scientific Research, 2014, 20(10), pp. 1222–1228

[6] Nasir Saleem, Muhammad Irfan Khattak, Elena Verdú. 2019, "On Improvement of Speech Intelligibility and Quality: A Survey of Unsupervised Single Channel Speech Enhancement Algorithms", International Journal of Interactive Multimedia and Artificial Intelligence, (2019), http://dx.doi.org/10.9781/ijimai.2019.12.001

[7] Selvaraj Poovarasan‡, Eswaran Chandra 2019, "Speech Enhancement Using Sliding Window Empirical Mode Decomposition and Hurst-based Technique", Archives of Acoustics Vol. 44, No. 3, pp. 429–437, DOI: 10.24425/aoa.2019.129259

[8] Adappa S Angadi, Adokshaja Kulkarni, Ravi A Gadad, K. Sridhar 2020, "Survey on Efficient Signal Processing Techniques for Speech Enhancement", International Research Journal of Engineering and Technology, Vol. 7 Issue: 1, E-ISSN: 2395-0056

[9] Karthik, B., Krishna Kumar, T., Vijayaragavan, S.P., Sriram, M., Removal of high density salt and pepper noise in color image through modified cascaded filter, Journal of Ambient Intelligence and Humanized Computing, 2020

[10] Çakır, Emre, Parascandolo, Giambattista, Heittola, Toni, Huttunen, Heikki, Virtanen, Tuomas, 2017, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 25, doi - 10.1109/TASLP.2017.2690575

[11] G. Manjula, M. Shiva Kumar, 2016, "Overview of Analysis and Classification of Stuttered Speech", International Journal of Industrial Electronics and Electrical Engineering, ISSN: 2347-6982 Vol-4, Issue-7.

[12] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, 2014, "Convolutional Neural Networks for Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol.22, No. 10

[13] Karthik, B., Kiran Kumar, T.V.U., Noise removal using mixtures of projected gaussian scale mixtures, World Applied Sciences Journal, 2014, 29(8), pp. 1039–1045