INTERNATIONAL JOURNAL OF SYSTEMS APPLICATIONS, ENGINEERING & DEVELOPMENT
DOI: 10.46300/91015.2020.14.21 Volume 14, 2020

A Novel Approach to Increase the Efficiency of a Multi-lingual Real-time Speaker Identification System

Alaka Pradhan
alakapradhan111@gmail.com
Dept. of Instrumentation and Electronics Engineering, College of Engineering and Technology, Bhubaneswar, India

Susanta Kumar Sarangi
susantasarangi@soa.ac.in
Dept. of Electronics and Communication Engineering, ITER, SOA Deem to be University, Bhubaneswar, India

Kanhu Charan Bhuyan
kcbhuyan@cet.edu.in
Dept.of Instrumentation and Electronics Engineering, College of Engineering and Technology, Bhubaneswar, India

Abstract— Nowadays, the real-time speaker recognition system is very popular due to its cost-effective nature. However, it is a very challenging one to produce a more efficient speaker identification system. In our work, we work on a multi-lingual real-time speaker identification system. We work in a novel way to enhance the efficiency of the said system. We take some real speech signals and use different speech enhancement methods and our proposed voice activity method (VAD) to enhance the efficiency of said system. By doing so, we increase the accuracy of the said system relatively by 2% as compared to existing methods.

Keywords— Speaker Identification system, VAD, speech enhancement method, MFCC, SFCC

I. INTRODUCTION

Speech technology and systems with human-computer interactivity have always indicated secure, stable, and remarkable enlightenment and uptrend over the last two decades. Nowadays, we can use speech signals for various household activities due to the natural and cost-effective nature of voice signals [1]. Due to the rich information, it contains about the particular person and its environmental condition and also due to its complicated signal produced as a result of various transformations occurring at several different levels i.e., acoustic, semantic, linguistic, and articulatory. Speech processing is the extrication or uprooting of the required information from a speech signal, which is to be digitalized and processed through a computer [2]. As a speech signal is a carrying message of diverse information, the speech processing field has different applications depending on the kind of information we are interested in.

Speaker recognition, also called as speaker identification (SID), deals with the task of recognizing or identifying a person with characterized voices. Analysis, coding, and recognition are the main areas of research in the speech processing system. As a subtask under the speaker identification system, speaker verification is about verifying whether a speech segment is coming from a specific speaker or not [3]. The commercialization of speech technologies helps the machines to respond correctly and provide useful services in multidimensional requirements. For the case of our research, here Multi-lingual (multiple languages based speech signals i.e., in English, Hindi, and Odia) Speaker Identification will be the area of focus throughout. When it comes to the name real-time speaker identification system, the system will work under the real-time environmental conditions of the speaker itself, which will be a real practical service towards mankind and our society as extended to three different languages [4].

The main problem which we will face in case of a real-time speaker identification system is that, the effect of noise is high due to real-time data. Noise is random in nature and has no useful information for our speech rather it decreases the degree of identification accuracy

and efficiency. Keeping this on mind we have reviewed various speech enhancement techniques [17-23], which we can use in our existing real-time speaker identification system to overcome the problem of noise effect so that we can increase the accuracy and efficiency of our system. In an in-depth study of said techniques, we found the Voice Activity Detection (VAD) method has a vital role in enhancing the speech signals, which is common to all the methods.

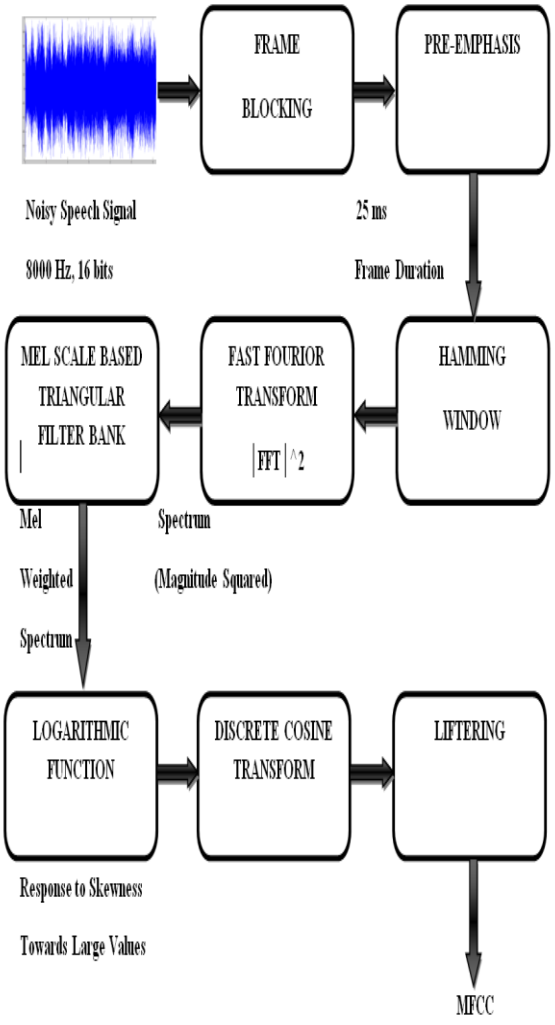


Fig. 1 general process of MFCC

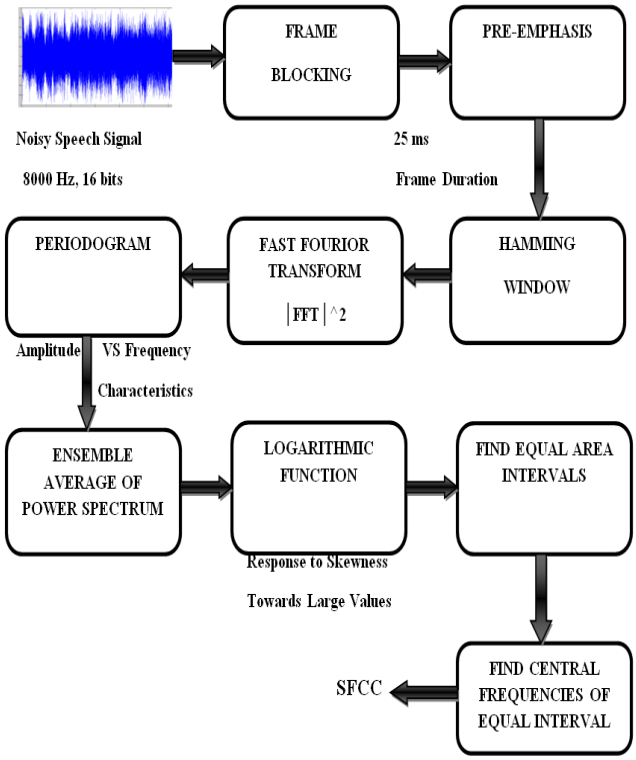


Fig. 2 General process of SFCC

So, in-depth learning about the existing VAD method gave us a detailed of its working. After that, we modified this existing VAD to get a better VAD, which is successfully implemented in the speech enhancement techniques, whose result showed better performance as compared to the existing one.

II PROPOSED SYSTEM INTRODUCING SPEECH EMHANCEMENT TECHNIQUE AND IMPROVED VOICE ACTIVITY DETECTION METHOD

The proposed approach contains the following parts: (i) feature extraction from the speech signal, (ii) speech enhancement technique used at the testing phase, (iii) improved voice activity detection algorithm, (iv) classification of features. The respective components are discussed in detail, as in the sub-sections below.

A. Feature extraction methods (MFCC and SFCC):-

As we know Mel frequency cepstral coefficient (MFCC) is very popular feature extraction method [5-9] and in recent years speech-signal-based frequency cepstral coefficient (SFCC) [4], [10-12] is doing well in speech

based different applications. We use both the methods in this paper. We need to extract the features both at the training phase and testing phase of a speaker identification system. Here at the testing phase, before extracting features, we will make the real-time speech signals clean by introducing a speech enhancement technique with a modified VAD algorithm. The general process to obtain the MFCC features is as depicted in Figure 1.

To start the process of feature extraction first, we need to analyze the voice after taking an input through a microphone from a speaker. We have collected the input data from speakers in the type of .wav file. The system's design then involves manipulating the input signal, i.e., keeping the sampling frequency to 8000 Hz, changing the bit to 16bit, and then making the channel mode to monotype. Different operations are performed on the input signal at different levels, i.e., the signal is first windowed and then goes for the pre-emphasizing process. The mathematical equation involved in this process is given below:

$$Y[n] = X[n] - a\,X[n-1] \quad (1)$$

Where Y[n] is the output signal and X[n] is the input signal to be pre-emphasized, and the value of $a = 0.95$ which means it makes 95% of any one sample is pre-assumed to originate from the previous sample. After that, the hamming windowing is subsequently applied onto each frame for its smoothening. Fast Fourier transform (FFT) is applied on each frame to transform from time to frequency domain and subsequently mapped onto the mel-scale. The relation for mel-scale to frequency scale and vice-versa is represented by:
The formula for converting from frequency to Mel scale is:

$$m = M(f) = 1125\ \ln(1 + f/700) \quad (2)$$

To go from Mel back to frequency:

$$f = M^{-1}(m) = 700(e^{\frac{m}{1125}+1}) \quad (3)$$

Where m Stands for Mel scale value and f Stands for frequency in hertz. As the speech signals, which we are dealing with, are non-stationary so, it is pronounced that they keep changing with time. It may be a change in voice or change in frames. Therefore, we need to add features related to the change in cepstral features over time. There are two types of feature adding modes, such as delta or velocity features, and the other one is delta-delta or acceleration features. For this work, we have considered 12 delta features and 39 delta-delta features where delta features represent the change between frames in the corresponding cepstral or energy feature. In contrast, the delta-delta features represent the change between the corresponding delta features. The energy in a frame for a signal x in a window from time sample t1 to time sample t2 is represented at the equation below:

$$\text{Energy} = \sum x^2[t] \quad (4)$$

$$d(t) = \frac{c(t+1) - c(t-1)}{2} \quad (5)$$

Where x= input signal, t= time which ranges from t1 to t2 and $d(t)$= time domain derivative term.

At the final step, we convert the log Mel spectrum back to time using Discrete Cosine Transform (DCT). The result is called the Mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. The first 'D' components of DCT represent a compacted MFCC vector of the corresponding frame. Denoting the output of the filter bank by Ek (k=0, 1... K), the MFCCs are calculated as,

$$Cn = \sum_{k=1}^{k} (\log E_K) \cos\left[\frac{n(k-0.5)\pi}{k}\right], n = 0,1 \ldots D \quad 6)$$

Where D= number of MFCC coefficients, K = number of Mel-scaled filters. Now, coming to the next feature extraction method i.e., SFCC, the general process to obtain its features is depicted in Figure 2.

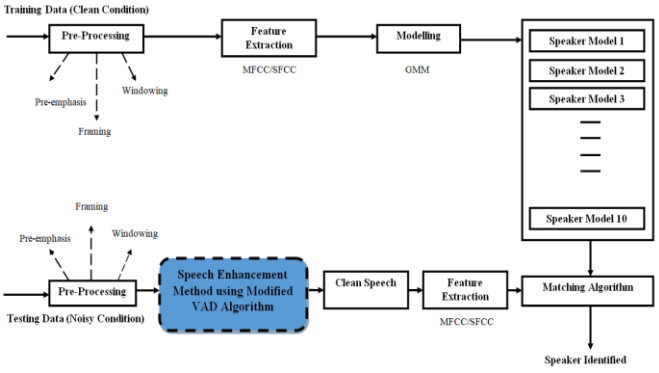


Fig. 3 proposed system

The blocks equal to MFCC, as described above, plays the same function here for SFCC. Only the filter bank here is replaced by speech-signal based triangular filter bank and unlikely in MFCC, after Fourier transform is done, Periodogram is found showing the characteristics of the signal plotted against frequency and amplitude to represent the power spectrum at each frequency. From the power spectrum, we then assembled the average, which is a set of very large numbers. Further to response to the skewness of those large numbers, a logarithmic function has been taken. Then equal-area intervals are classified, and the center frequencies of those equal-area intervals are the SFCC coefficients.

B. System Modeling Using Gaussian mixture models

The training and test feature vectors are directly compared with each other with the assumption that either one is an imperfect replica of the other. This module classifies extracted features according to the individual speakers whose voices have been stored. The recorded voice patterns of the speakers are used to derive a classification algorithm. The training phase is to estimate the parameters of the probability density function from a training sample. Matching is usually done by evaluating the likelihood function of the test utterance with respect to the model [13] [14].

The purpose of voice modeling is to build a model that captures these variations in the extracted set of features. In this work, based on the statistical approach, we focus on the speaker recognition in independent mode of the text [15-16] by the Gaussian mixture. Gaussian mixture models (GMMs) are parametric representation of a probability density function. When trained to represent the distribution of a feature vector, GMMs can be used as classifiers. GMMs have proved to be a powerful tool for distinguishing acoustic sources with different general properties

C. Speech Enhancement Techniques

MFCC features are the more commonly used and robust technique for feature extraction in presently available speaker identification systems, especially in clean environments. But its performance degrades in a noisy environment. Our work in this paper is entirely focused on overcoming the weaknesses of MFCC in noisy speech, which directly hampers the system performance. Thus, we have introduced the speech enhancement method into both MFCC and SFCC methods to enhance their performance in application to the real world. Regarding the working principle of any speech enhancement method, we found some common and important phases of the whole enhancement in all the techniques discussed in [17-23]. The proposed method and an outline of these important blocks constituting the required system is given in figure 3 and 4, respectively.

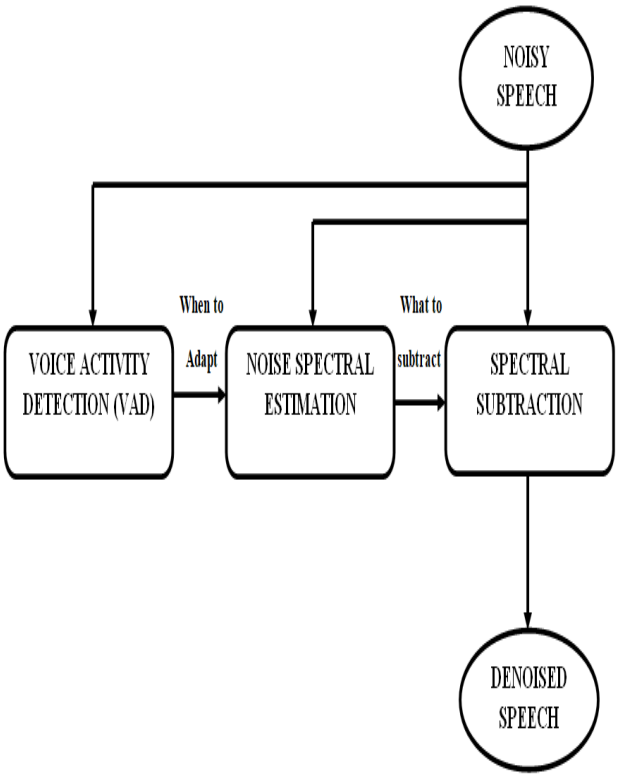


Fig. 4 outline of a common speech enhancement system

The real-time speech signal which contains noise and is denoised by first going through a VAD, whose work is to detect the activity of noise, whether there is speech activity or non-speech or silence. For our system to work efficiently and keep the system's computational time low, we consider these silences or non-speech parts to be noise and eliminate it. By removing the silence parts from the original speech, now the resulted

speech is somehow clear and sensible. But it is not able to remove the whole noise presence. So, the resulted VAD output signal is then tested by processing into seven different enhancement methods [17-23] and the method providing the best result [22] is then used in increasing the accuracy of the speaker identification system. To choose the best one among the seven, we have tested each one with a noisy speech signal (speech signal containing traffic noise), and the simulation result is shown below in Figures 5 and 6, respectively.

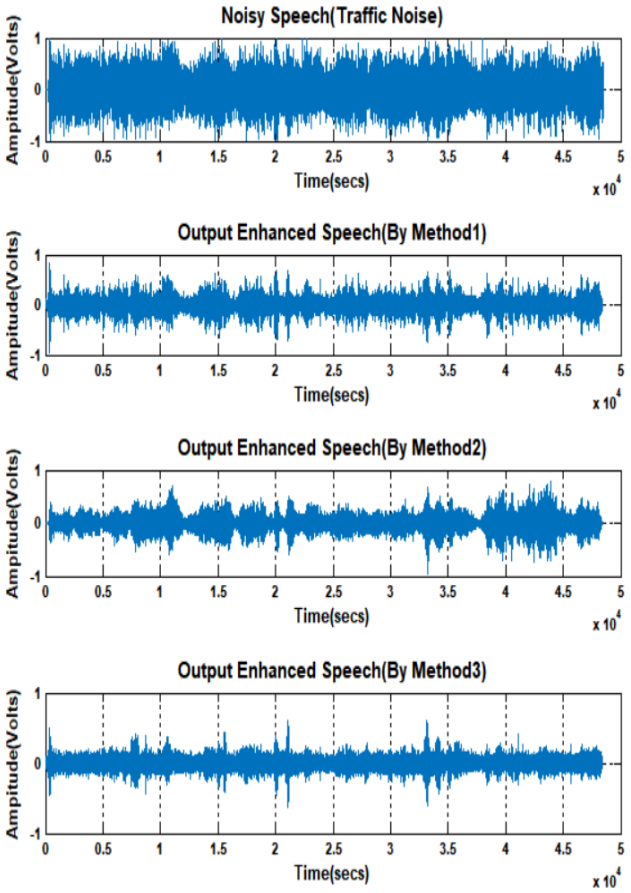


Fig. 5 experimental results showing input noisy signal affected by traffic noise and output enhanced speech signals using method 1 to 3.

Our input speech signal is a real-time signal i.e., the signal recorded near the traffic area, so the noise present in the speech signal is called traffic noise. The simulation of all the above methods was done using MATLAB software.

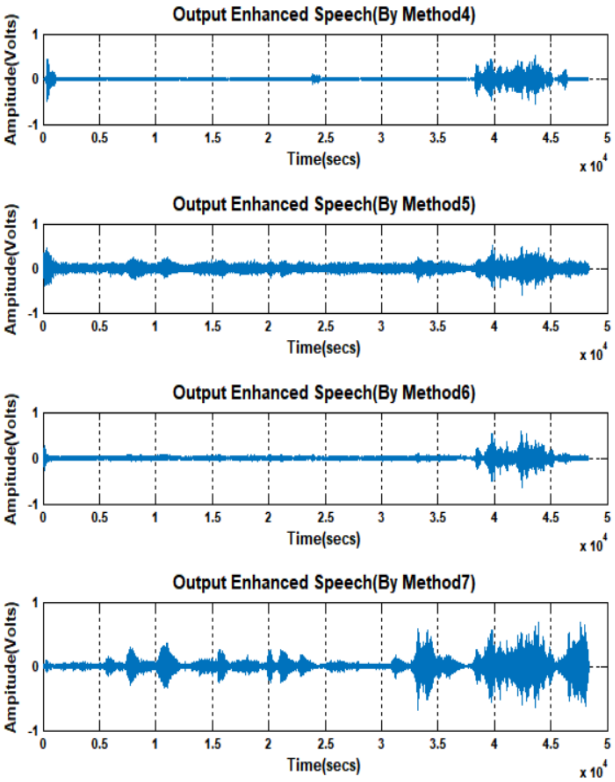


Fig. 6 experimental results showing output enhanced speech signals using method 4 to 7.

C. Comparison between the Existing Voice Activity Detection and Proposed Improved Voice Activity Detection Algorithm

While going through a rigorous study of those seven speech enhancement methods, it was found that they all consist of similar VAD algorithms. As the role of a VAD in speech enhancement system [24], [25] and [25] are vital, we focused on this exiting VAD algorithm. We have modified the threshold value and the VAD decision process of the existing VAD algorithm, such that it gave us a better result as compared to the existing one. A conventional VAD system [27] is presented below in Figure 7:

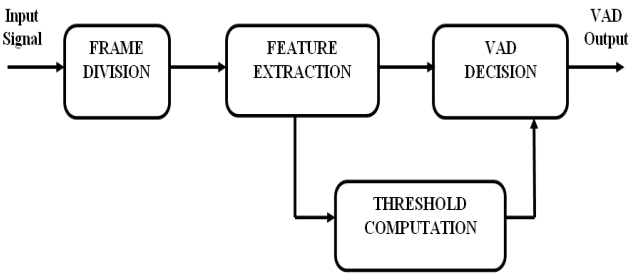


Fig. 7 block diagram of a VAD System

As per the above block diagram, the real-time speech signal is first processed by framing, followed by windowing. Then each frame is processed to extract the useful and valuable features from each frame. After features are extracted, the data of each feature is stored. Then, we need to set some threshold values for each data extracted from the feature extraction process. After this process, the values are fed to the voice activity detection block to make the decision based on its algorithm to detect the speech parts and the non-speech parts of the frame so that afterward, we will eliminate that non-speech magnitude part from each frame.

Comparison is made based on the threshold parameter i.e., the spectral distance calculation. Terms to get familiar with 1st are:

1. Signal: Input noisy signal
2. Noise: Estimated noise spectrum magnitude.
3. Noise margin: Assumed to be equal to 3.
4. Noise flag: Counts the noise present, and its initial value is set to 0.
5. Speech flag: Counts the speech presence.
6. Noise counter: Counts the presence of noise periods, initially set to 0.
7. Hangover: Assumed the value to be 8.
8. Distance: Distance is the measure of the mean of spectral distance.
9. Spectral Distance Existing (SP_E): Spectral distance for existing VAD method.
10. Spectral Distance Modified (SP_M): Spectral distance for the modified VAD method.

The equations used to design the spectral distance value which acts as the threshold value for the VAD method to decide for speech enhancement method (method 1 to 7 as described in the paper [17-23]) is given below:

Equations involved in the algorithm of the *Existing VAD* method:

$$Spectral\ Distance\ Existing(SP_E) = 20 \times (log10(signal) - log10(noise)) \quad (7)$$

Where signal=the input noisy speech signal

$noise$ = estimated noise power spectrum value

Assumptions taken: if the spectral distance value comes to be less than 0 then for those frames spectral distance is assumed to be 0.

Then we need to find the mean of the spectral distance calculated and it is given by the equation as follow:

$$mean = \frac{SP_E}{total\ number\ of\ frames} \quad (8)$$

Then assign this mean value to distance (Dist):

$$Dist = mean \quad (9)$$

Now comparing with Noise margin:

$$NoiseFlag = \begin{matrix} 1, if\ Dist < Noise\ margin \\ 0, else \end{matrix} \quad (10)$$

$$Noise\ Counter = \begin{matrix} Noise\ counter + 1, if\ Dist < Noise\ margin \\ 0, else \end{matrix} \quad (11)$$

Now we have to detect noise only periods and attenuate the signal:

$$Speech\ Flag = \begin{matrix} 0, if\ Noise\ \ counter > Hangover \\ 1, else \end{matrix} \quad (12)$$

Equations involved in the algorithm of the *Modified VAD* method:

$$Spectral\ Distance\ Modified\ (SP_M) = (20[log_{10}(signal)^2 - log_{10}(noise)^2]) \quad (13)$$

Where signal is the input noisy speech signal and noise estimated noise power spectrum value.

Assumptions taken: if the spectral distance value comes to be less than 0 then for those frames spectral distance is assumed to be 0.

Then we need to find the mean of the spectral distance calculated and it is given by the equation as follow:

$$mean = \frac{SP_M}{total\ number\ of\ frames} \tag{14}$$

Then assign this mean value to distance (Dist):

$$Dist = mean \tag{15}$$

Now comparing with Noise margin:

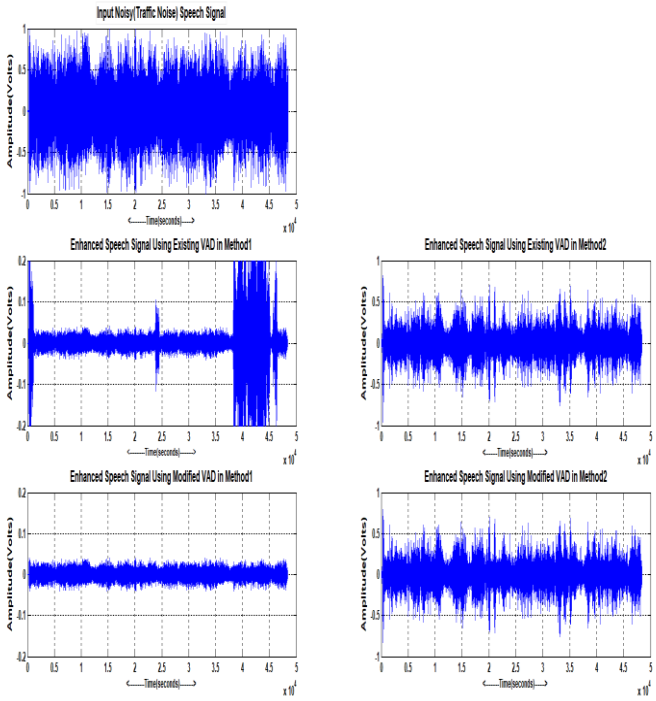
$$NoiseFlag = \begin{matrix} 1, if\ Dist > Noise\ margin \\ 0, else \end{matrix} \tag{16}$$

$$Noise\ Counter = \begin{matrix} Noise\ counter + 1, if\ Dist > Noise\ margin \\ 0, else \end{matrix} \tag{17}$$

Now we have to detect noise only periods and attenuate the signal:

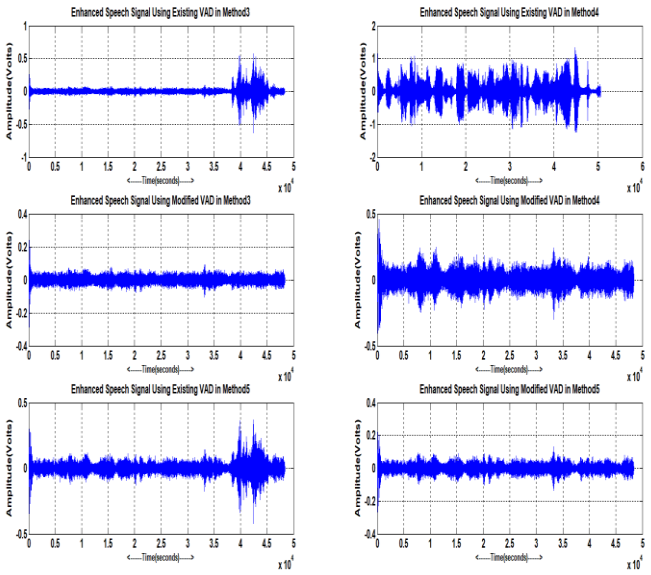
$$Speech\ Flag = \begin{matrix} 0, if\ Noise\ counter < Hangover \\ 1, else \end{matrix} \tag{18}$$

The above-said algorithms are processed in the Matlab software, taking a noisy input signal i.e., a real-time speech signal recorded near a traffic area. The output of each enhancement method using the existing and new modified VAD algorithm is clearly shown below. It is visible that the new proposed VAD method is giving a better-enhanced output signal compared to the existing one.



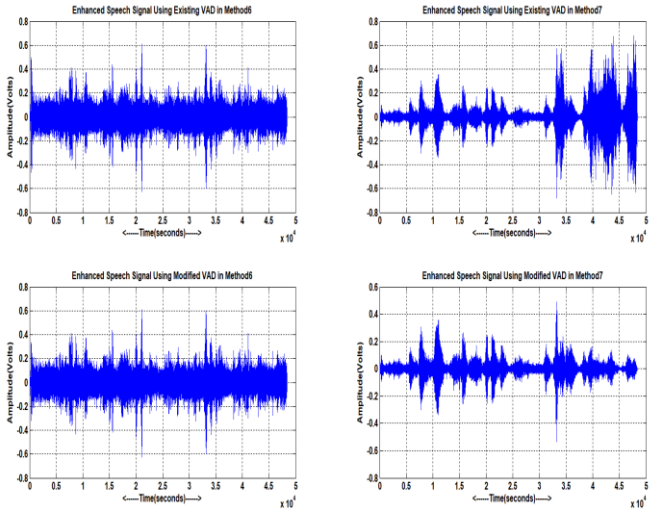
(a) (b)

Fig. 8 (a) Experimental Results Showing Output Enhanced Speech Signals Using Existing VAD, (b) Showing Output Enhanced Speech Signals Using Proposed VAD (applied to speech enhancement methods 1 & 2).



(a) (b)

Fig. 9 (a) Experimental Results Showing Output Enhanced Speech Signals Using Existing VAD, (b) Showing Output Enhanced Speech Signals Using Proposed VAD (applied to speech enhancement methods 3, 4&5).



(a) (b)

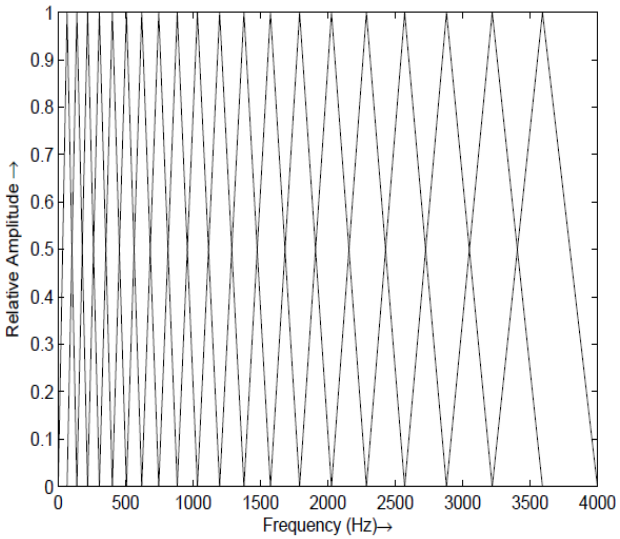
Fig. 10 (a) Experimental Results Showing Output Enhanced Speech Signals Using Existing VAD, (b) Showing Output Enhanced Speech Signals Using Proposed VAD (applied to speech enhancement methods 6 & 7).

Both the VAD algorithms are processed in the Matlab program and applied to the earlier described seven speech enhancement techniques individually and compared experimentally the output results (shown below). Hence, the best among the seven enhancement techniques (i.e., the method [22] that gave better output using the modified VAD algorithm is selected for use in the real-time speaker identification system to increase its accuracy.

Table 1: ***Input Data for Training Phase***

No. of speakers	Time of Recording speech signals	Language	Distance from microphone	Channel type	Duration of speech signal(minutes)	Sampling Frequency(Hz)	Type of Environment
1	DAY	ENGLISH HINDI ODIA	NORMAL	MONO	2-3MIN	8000	CLEAN
2	NIGHT	ENGLISH HINDI ODIA	NORMAL	MONO	2-3MIN	8000	CLEAN
3	DAY	ENGLISH HINDI ODIA	NORMAL	MONO	2-3MIN	8000	CLEAN
4	DAY	ENGLISH HINDI ODIA	NORMAL	MONO	2-3MIN	8000	CLEAN
5	NIGHT	ENGLISH HINDI ODIA	NORMAL	MONO	2-3MIN	8000	CLEAN
6	DAY	ENGLISH HINDI ODIA	NORMAL	MONO	2-3MIN	8000	CLEAN
7	DAY	ENGLISH HINDI ODIA	NORMAL	MONO	2-3MIN	8000	CLEAN
8	DAY	ENGLISH HINDI ODIA	NORMAL	MONO	2-3MIN	8000	CLEAN
9	DAY	ENGLISH HINDI ODIA	NORMAL	MONO	2-3MIN	8000	CLEAN
10	NIGHT	ENGLISH HINDI ODIA	NORMAL	MONO	2-3MIN	8000	CLEAN

(a)



(b)

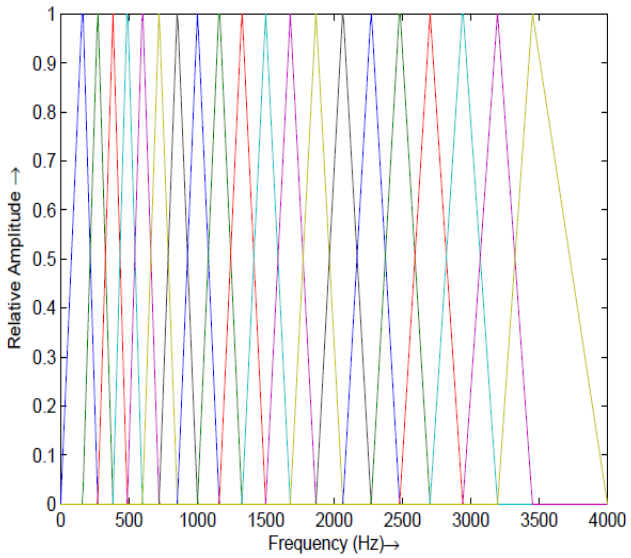


Fig. 11 canonical *filter bank structure of (a) MFCC & (b): SFCC.*

III. EXPERIMENTAL FRAME WORK

i. DATABASE:-

We have recorded the real-time speech signals at various noisy environments having different noise strength at different time periods with three different languages (i.e., English, Hindi, and Odia) which is given below:

Above shows the collection of samples for the training phase in clean condition recorded from 10 different speakers with three different languages i.e., in English, Hindi, and Odia for time duration between 2-3 minutes in each language with a constant sampling frequency of 8000Hz.

Table 2: *Input Data for Testing Phase*

No. of speakers	Time of Recording speech signals	Language	Distance from microphone	Channel type	Duration of speech signal(minutes)	Sampling Frequency (Hz)	Type of Environment
1	DAY	ENGLISH HINDI ODIA	NORMAL	MONO	2-3SEC	8000	MACHINE
2	DAY	ENGLISH HINDI ODIA	NORMAL	MONO	2-3SEC	8000	WIND
3	DAY	ENGLISH HINDI ODIA	NORMAL	MONO	2-3SEC	8000	FAN
4	DAY	ENGLISH HINDI ODIA	NORMAL	MONO	2-3SEC	8000	MUSIC
5	NIGHT	ENGLISH HINDI ODIA	NORMAL	MONO	2-3SEC	8000	CROWD
6	NIGHT	ENGLISH HINDI ODIA	NORMAL	MONO	2-3SEC	8000	VOLVO
7	DAY	ENGLISH HINDI ODIA	NORMAL	MONO	2-3SEC	8000	MACHINE
8	NIGHT	ENGLISH HINDI ODIA	NORMAL	MONO	2-3SEC	8000	RAIN
9	DAY	ENGLISH HINDI ODIA	NORMAL	MONO	2-3SEC	8000	CROWD
10	NIGHT	ENGLISH HINDI ODIA	NORMAL	MONO	2-3SEC	8000	TRAIN

The table 1 and 2 shows the collection of real-time samples from the same speakers who have already enrolled for speaker identification during the training phase with three different languages i.e., in English, Hindi, and Odia. We have recorded the samples at different environmental conditions like, for example, we have recorded during rain, recorded at the crowdie place, recorded near factories where we got the noise of machines and wind. Also, we have recorded near the railway station to get the noise of trains, etc. for a duration of almost 10 seconds from which we have considered only 2 to 3 seconds for our testing phase.

ii. SCORE CALCULATION:-

We need to calculate the score in terms of percentage, which is defined as the accuracy rate. Here in our work we will calculate the accuracy of real-time speaker identification system using MFCC and SFCC individually and compare both of them in two different conditions i.e., accuracy before using speech enhancement and that of after using speech enhancement and another comparison will be done between the speech enhancement using existing VAD and speech enhancement using our modified VAD.

The mathematical equation for score calculation is given as:

$$Score\ of\ identification(Accuracy) = \left[\frac{Number\ of\ utterances\ correctly\ identified}{Total\ number\ of\ utterances\ under\ test}\right] \times 100 \qquad (19)$$

i. *COMPARISION OF MFCC & SFCC:-*

Comparison between MFCC and SFCC is done based on different conditions such as:

Accuracy of MFCC (varying model order considering: before using speech enhancement and after using speech enhancement condition).

Accuracy of SFCC (varying model order considering: before using speech enhancement and after using speech enhancement condition).

Table 3: ***MFCC & SFCC Accuracy (With 39 features)***

Model Order	No of Features (Kept Constant)	Feature Extraction Methods	Accuracy (Before Speech Enhancement) %	Accuracy (After Speech Enhancement) %
2	39	MFCC	63.33	66.67
		SFCC	65	70
4	39	MFCC	65	76.67
		SFCC	73.33	80
8	39	MFCC	73.33	80
		SFCC	76.67	83.33
16	39	MFCC	80	83.33
		SFCC	83.33	86.67

Above data shows clearly that after using speech enhancement in the system (with the existing VAD), the performance of the system has increased at an average of 3% (at each model order) in case of MFCC and an average of 5% in case of SFCC (at each model order). It is also seen experimentally that SFCC is giving better performance than MFCC even in the normal system condition (i.e., before speech enhancement). Figure 10 shows the canonical filter bank structures of MFCC and SFCC. From these figures, we can see that in the case of MFCC, at the F1 frequency zone i.e., between 0 to 1000 Hz, there are more no. of filters, and cepstral features are there, but in higher frequency regions F2 and F3 the no. of filters decreased. But in the case of SFCC, between 0-1000 Hz, the no. of filters are less, and in higher frequency regions, the no. of filters are more. So, more no. of cepstral features mean more speakers are tend to be identified. So, SFCC is found to give better performance than that of MFCC.

ACCURACY COMPARISION OF EXISTING VAD & PROPOSED VAD:-

Graphically already, we have shown the comparison between both VAD methods in the above section. Here we have compared both the algorithms based on the percentage of accuracy that our real-time multilingual speaker identification system is giving, when examined

through speech enhancement technique with the existing VAD and modified VAD.

Already the percentage of accuracy with existing VAD is provided in table 3. Here the same value is compared with our proposed VAD method.

Table 4: ***Comparison of Accuracy between Exiting VAD Method & New Proposed VAD Method***

Model Order	Feature Extraction Method Used	Accuracy With Existing VAD %	Accuracy With Modified VAD %
2	MFCC	66.67	70
	SFCC	70	73.33
4	MFCC	76.67	83.33
	SFCC	80	86.67
8	MFCC	80	86.67
	SFCC	83.33	90
16	MFCC	83.33	90
	SFCC	86.67	93.33

From the above-provided data, it is clear that our proposed VAD is giving an increased average percentage of 4% (in case of MFCC) and an average increase percentage of 7%(in case of SFCC) as compared to existing VAD.

IV. CONCLUSION

We have used different speech enhancement methods to reduce the noisy part of the speech signal. After that, we proposed modifying the VA method to enhance the speaker identification system's efficiency. By doing so, we increase the system accuracy by relatively 2% than the existing MFCC and SFCC based system. In the future, the better classifier with feature extraction methods overall will increase the system in a better way.

V. REFERENCES

[1] Helander, Martin G., ed. Handbook of human-computer interaction. Elsevier, 2014

[2] Mohamed Faouzi, Ben Zeghibaa, "Joint Speech And Speaker recognition," in *IDIAPRR*, February 2005, pp. 05-28.

[3] Prateek Srivastava, Reena Panda & Sankarsan Rauta, "A Novel, Robust, Hierarchical text-independent Speaker recognition Technique," *An International Journal (SPIJ)*, 2012.

[4] S. Priyadarshini, S. K. Sarangi and K. C. Bhuyan, "A Novel Approach To Enhance The Efficiency Of Real-Time Speaker Identification System," 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), Bhubaneswar, India, 2018, pp. 52-54.

[5] Namrata Dave, "Feature Extraction Methods LPC,PLP and MFCC In Speech Recognition," *International journal for advance research in engineering and technology*, vol. 1, no. VI, July 2013.

[6] M. Barik, S. Kumar Sarangi and S. Kumar Sahu, "Real-time speaker identification system using cepstral features," 2016 2nd International Conference on Communication Control and Intelligent Systems (CCIS), Mathura, 2016, pp. 89-93.

[7] Md. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition ," in *Speech Communication*, 2012, pp. 543-565.

[8] SAYF A. MAJEED, HAFIZAH HUSAIN, SALINA ABDUL SAMAD, TARIQ F. IDBEAA, "Mel frequency cepstral coefficients(MFCC) feature extraction enhancement in the application of speech recognition: A Comparison Study," *Journal of Theoretical and Applied Information Technology*, vol. 79, no. 1, September 2015.

[9] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Cefficient (MFCC) and Dynamic Time Warping (DTW) Technques," *Journal of Computing* , vol. 2, no. 3, March 2010.

[10] K. Paliwal, B.Shannon, J. Lyons and K. wojcicki, "Speech-signal-based frequency warping," in *IEEE Signal Process. Lett.*, 2009, pp. 319-322

[11] Susanta Kumar Sarangi, Goutam Saha, "A Novel Approach in Faeture Level for Robust Text-Independent Speaker Identification," in *4th International Conference on Intelligent Human Computer Interaction*, Kharagpur, December 2012, pp. 27-29.

[12] Paul, Dipjyoti, Monisankha Pal, and Goutam Saha. "Spectral features for synthetic speech detection." *IEEE journal of selected topics in signal processing* 11.4 (2017): 605-617.

[13] V. Srinivas, Ch. Santhi Rani, P.Hema kumar, "Novel Speaker Recognition System using GMM," *International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)*, vol. 4, no. 9, September 2017.

[14] D.A.Reynolds and C.Richard, "Robust text-Independent speaker identification using Gaussian mixture speaker models," in *IEEE Transaction on speech and audio processing*, January 1995.

[15] Abhilasha Sukhwal, Mahenfra Kumar, "Comparative Study between different Classifiers based Speaker Recognition System using MFCC for Noisy Envionment," in *International Conference on Green Computing and Internet of Things (ICGCIoT)*, Kota, 2015.

[16] Rania Chakroun, Leila Beltaifa Zouari, Mondher Frikha and Ahmed Ben Hamida, "Improving Text-independent Speaker Recognition with GMM," in *2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, Tunisia, March 2016, pp. 21-24.

[17] STEVEN,F.BOLL, "Supression of Acoustic Noise in Speech Using Spectral Subtraction," in *IEEE TRANSACTIONS ON ACOUSTIC, SPEECH, AND SIGNAL PROCESSING*, vol. 2, 1979

[18] M.Berouti, R.Schwartz and J.Makhoul, "ENHANCEMENT OF SPEECH CORRUPTED BY ACOUSTIC NOISE," in *Bolt Bernek and Newman Inc.*, Cambridge.

[19] Y. Ephraim and D.Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," in *IEEE Trans.*, vol. 32, 1984, pp. 1109-1121.

[20] Scalart, Pascal. "Speech enhancement based on a priori signal to noise estimation." *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Vol. 2. IEEE, 1996.

[21] Boh Lim Sim, Yit Chow Tong, Joseph S.Chang and Chin Tuan Tan, "A Parametric Formulation of the Generalized Spectral Subtraction Method," , vol. 6, JULY 1998.

[22] Sunil,D. Kamath and Philipos, C.Loizou, *A MULTI-BAND SPECTRAL SUBTRACTION METHOD FOR ENHANCING SPEECH CORRUPTED BY COLORED NOISE*, 3rd ed. Boca Raton: Electrical Engineering Handbook.

[23] Israel Cohen, "Speech Enhancement Using a Noncausal A Priori SNR Estimator," in *IEEE SIGNAL PROCESSING LETTERS*, SEPTEMBER 2004.

[24] X.Bao, J. Zhu,N. Chen, "A Robust Voice Activity Detection Method Based on Speech Enhancement," in *IET Intelligent Signal Processing Conference*, December 2013, pp. 1-4..

[25] Y. Dongwen, Y. Yonghong,D. Jianwu,F. K. Soong, "Voice Activity Detection based on an unsupervised learning framework," in *IEEE Transactions on Audio Speech and Language Processing*, November 2011, pp. 2624-2633.

[26] S. W. Chin,K. P. Seng,L.Ang, "Improved voice activity detection for speech recognition system," in *International Computer Symposium*, December 2010, pp. 518-523.

[27] J. Sohn,N. S. Kim, and W. Sung, "A Statistical Model-Based Voice Activity Detection," in *IEEE Signal Processing Letters*, 1999, pp. 1-3.