INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Nanotechnology

ISSN online: 1741-8151 - ISSN print: 1475-7435
https://www.inderscience.com/ijnt

An improved model for unsupervised voice activity detection

Shilpa Sharma, Rahul Malhotra, Anurag Sharma

DOI: 10.1504/IJNT.2023.10056468

Article History:

Received:	10 October 2021
Last revised:	26 November 2021
Accepted:	27 December 2021
Published online:	31 May 2023

An improved model for unsupervised voice activity detection

Shilpa Sharma*

Computer Science and Engineering, CT Group of Institutions,
Jalandhar, India

and

Lovely Professional University,
144411, India
Email: shilpa13891@gmail.com

Rahul Malhotra

Electronics and Telecommunication Engineering,
CT Group of Institutions,
Jalandhar, 144020, India
Email: blessurahul@gmail.com

Anurag Sharma*

Department of Computer Science & Engineering,
GNA University,
Phagwara, 144401, India
Email: er.anuragsagmail.com
*Corresponding authors

Abstract: The antique way to express our self is speech and nowadays speech is being used in many applications especially in machine communication. As the application of speech is increasing at rapid rate, therefore various techniques are evolving to separate out the speech signals from audio signal which is mixture of noise and speech. The method to distinguish voice and noise is known as voice activity detection. This method is gaining huge popularity as it removes background noise and acceptable approach in the area of speech coding, audio surveillance and monitoring. In this manuscript, hybrid model of unsupervised classifier is investigated. The proposed approach is tested at different levels of noise signal and overlap window size. To validate the proposed approach, a comparison with existing artificial neural network and support vector machine (SVM) is presented. The outcomes of the proposed method are observed better than the existing methods with the accuracy of 99.73% along with better SNR of 25.61 dB. Also proposed model LFV-KANN efficiently handles increase in noise power by hybridisation of two classifiers: ANN and K-means clustering.

Keywords: voice activity detector; artificial neural network; SVM; support vector machine; K-means; unsupervised learning; machine learning; TIMIT database.

Reference to this paper should be made as follows: Sharma, S., Malhotra, R. and Sharma, A. (2023) 'An improved model for unsupervised voice activity detection', *Int. J. Nanotechnol.*, Vol. 20, Nos. 1/2/3/4, pp.235–258.

Biographical notes: Shilpa Sharma received her BTech in Computer Science from Lovely Institute Technology in 2009 and MTech in computer Science Engineering from DAVIT, Punjab, India. She has 12 years of experience as an Assistant Professor. Her area of interests includes data mining, soft computing, computer networking and artificial intelligence. She has published various articles in national/international conferences and journals. She has guided more than 10 MTech, students. She published around 20 research papers in national, international journals and, conferences.

Rahul Malhotra obtained his UG and PG degree in Electronics and Communication Engineering and received PhD degree from IKG PTU, Jalandhar, Punjab. He is currently working as Professor in CT Group of Institutions, Jalandhar, India. His research areas include image processing, IoT, signal processing, machine learning, text mining. He has more than 18 years of teaching experience. He has published various articles in national/international conferences and journals. He has guided more than 95 MTech students and seven Research Scholars. He published around 100 research papers in national, international journals and conferences.

Anurag Sharma obtained his UG and PG degree in Electronics and Communication Engineering and received PhD degree from National Institute of Technology, Jalandhar, India in 2019. He is currently working as Professor in the Faculty of Engineering and Design Automation, GNA University, Phagwara, India. He has more than 18 years of teaching experience. He has published various articles in national/international conferences and journals. He has vast research and industrial experience in the fields of biomedical engineering, data sciences, image processing, machine learning. He has guided more than 90 MTech students and more than five Research Scholars. He published around 100 research papers in national, international journals and conferences.

1 Introduction

Speech is an ancient mode of communication, and it is now being employed in a wide range of applications, including machine communication and biometric reconstruction, among others. As the use of speech in more and more applications grows at an exponential rate, new approaches are being developed to distinguish speech signals from audio signals that are a mixture of noise and speech. Voice activity detection is the term used to describe the mechanism of distinguishing between voice and background noise (VAD). This method is gaining widespread acceptance since it effectively eliminates background noise and is an acceptable solution in the fields of speech coding, audio surveillance and monitoring, and speaker language recognition, among other applications.

In the previous work, various problems related to VAD have been addressed. The VAD generally achieved by feature extractions and distinct model. The VAD can be categorised as supervised or unsupervised speech detection. The supervised VAD approach is less complex but more accurate as compared to unsupervised techniques as it requires specialised data which is not labelled. However, various techniques had been developed for both the approaches such as global thresholding scheme with a fuzzy entropy tool [1], GROA which was an integration of the ride optimisation algorithm and grasshopper optimisation algorithms in tuning the optimal weights of SVM [2], acoustic decoy using deep learning [3], support vector machine (SVM) classifier with a radial basis function (RBF) kernel [4], adaptive blind source separation (BSS) algorithm for acoustic noise reduction and speech amplification [5], vector quantisation-based self-adaptive VAD [6], energy-based VAD, Gaussian mixtures VAD, LRT-based VADs and sequential GMM-based VAD [7]. The other techniques are based on discrete Fourier transforms (DFT) coefficient based statistical models. Based upon the previous work done by other authors on the feature extraction techniques and classifiers in VAD is compiled in the form of Table 1.

Table 1 Compiled data on previous work done in VAD

Author (et al.)	*Features extraction*	*Classifier(s)*	*Accuracy (%)*
A. Mohamoud	MFCCs, LPC, energy, and pitch	K-NN	80
Vlasenko	MFCCs and log energy	HMM	81
Wua	Spectrum, formant, and pitch	GMM, SVM	76
Neiberg	Pitch, MFCCs	GMM	93
Vernendis	Formats, pitch, energy	HMM	54
Altun	Pitch, energy, MFCCs, and LPC	Multi-class SVM	80
Luengo	Energy, pitch	voice detection GMM	92
Sidorova	Formants, intensity, and pitch	ANN	79
Nicholson	Speech power, pitch, LPCs	Neural Network	50
Hetan	TEO-PWP and MFCCs	GMM and PNN	54
W. Kwon	Pitch, energy, formant, and MFCC	Gaussian and SVM	41
Haq	Pitch, energy, duration, and MFCC	MLB	53
Emerich	MFCCs and 7 statistical moments	SVM	80
Pathak	LPCs	Neural Network	46
Iliou	Energy, formant, and MFCC	SVM	74
T.L. Nwe	LFPC, LPCC, and MFCC	HMM	78
Yamada	Frequency and bandwidth	Neural Network	70
L. Pao	LPC and MFCCs	K-NN	80
Petrushin	Pitch, formants, bandwidth	K-NN	70
H. Kao	Pitch, energy, formant, and MFCC	SVM	90
Han	F0, energy, duration, model jitter	ANN	78
Saidatul	Energy	ANN	78
Mokhayeri	PD, ECG, and PPG	HMM	82
H. Costin	HRV and MV of ECG	SVM	89
	Minimum distance MV frequency		
Wijsman	HR, ECG, SC, and EMG	K-NN	78
Choi	PDM and PSD features	K-NN	83
L. He	TEO-PWP and MFCCs	GMM and PNN	96

Besides these approaches, some other methods have also been investigated to exploit the variability of noise and voice properties. Additionally, few researchers have used the integration of multiple features with the help of linear combination, principal component analysis, and linear discrimination analysis. The major limitation of supervised VAD method is requirement of abundant trained dataset which may lead to mismatch between the trained and untrained testing data [8]. Therefore, unsupervised approaching is attaining the attention of the researchers in this domain and various approaches have been development. All the techniques are developed to improve the voice detection process.

The initial phase of all the previously proposed approaches is to use a window size of audio frame and use feature extraction method to detect the presence of noise in given speech frame. However, the widow size selection processor for getting optimal widow size has not been explored. Hence in this work we considered the overlap window size is considered for identify a to investigate the improvement in the voice frame detection process. The scheme is evaluated in terms of efficiency of voice detection, probability of noise presence, accuracy, and time consumption.

Initial GSM 729 [9] specifications defined the VAD module for low sampling speech coding, which was less resilient to noise than later versions. The robustness to noise is also a significant step, as it has the potential to increase the effectiveness of automatic speech recognition when operating in a noisy setting. Many techniques have been presented in order to keep up with the state of the art in speech activity detection. The majority of these proposed algorithms are distinguished on the basis of the features that are included in them [9]. Out of all the characteristics, short term energy, MFCC, pitch and low frequency variability have risen in favour due to their simplicity. However, they are effective in noisy environments, these features have been used in this research [10]. All of these changes took place between the years 2000 and 2021. All of these experiments increased the robustness against a wide range of situations while reducing complexity and increasing overall efficiency. Additionally, combinations of different features-based VAD algorithms have been tested, for example, CART [11] and ANN [12], however the complexity of these algorithms has increased. Additionally, certain attempts have been initiated for noise characterisation, resulting in the development of enhanced speech spectra obtained from Wiener filtering based noise statistics [13], which are now in use. Because they were designed under the assumption of stationary noise, these approaches are more sensitive to variations in SNR. Other studies have also reported on the use of noise estimation and adaptation to improve the robustness of VADs while adding additional computing complexity to the equations. Additionally, the ETSI AMR [13] and the AFE have been employed as VAD approaches for a variety of algorithms that have been presented. Due to the increase in the number of applications of VAD, many new characteristics are being introduced, such as wavelet converted images, spectral information, and wavelet energy entropy ratio [14]. These characteristics are simple to implement and more ideal for random signal processing, but they have the problem of being unable to determine the end point of a voice signal efficiently in a noisy environment. In addition, a teager energy (TE) based on power spectral density is calculated in order to improve the judgement performance. The presence of IS-127 has been identified as a characteristic of detection in VAD. The comparison of VAD techniques is shown in Table 2.

Table 2 VAD techniques comparison

Technique	*Pros*	*Cons*	*Solutions*
Multimodal VAD	highly effective in noisy and transient environment	Incorporation of the video signal	Multimodal Compact Bilinear Pooling (MCBP)
Hidden Markov models (HMMs) based Gaussian mixture	Simple design and its practical design	Inefficient approach for nonlinear functions	Neural network based VAD
Wavelet transformation, spectral entropy, and wavelet energy entropy ratio	Easiness in realisation and more appropriate for random signal processing	Limitation to detect end point of speech signal efficiently under noisy environment	A power spectral density based Teager Energy (TE)
GMM-based VAD	Integration of log likelihood ratio and short time energy based voice activity detection feature.	Unsupervised VADs only	Self-adaptive VAD based on vector quantisation scheme
Short term energy and zero crossing rate	Less complexity	Ineffective in noisy environment	Auto-correlation based function, Mel-frequency, delta line spectral frequencies and features based higher order statics.
CART, ANN	effective in noisy environment	Augmented the complexity	Wiener filtering

The rest of the manuscript is organised as follows: Section 2 describes the proposed work to followed by the methodology of proposed algorithm that is used to achieve the higher accuracy and less time consumption at different overlap window size in Section 3. The outcomes for the proposed methods are illustrated in Section 4. Section 5 is concluding the manuscript.

2 Proposed work

The speech reorganisation is the oldest method to express our thoughts and now this field in gaining attentions of many researchers around the globe due to wide range of applications. Therefore, number of efforts has been made to improve speech signal transmission and remove unwanted signals. The method to extract speech signals from the from a voice signal is known as voice activity detection and have two essential part such as feature extraction and discrimination. Hence, previously various study, methods and schemes have been proposed on these two essential parts of VAD. Some of them are energy based feature extraction, pitch, MFCC, autocorrelation-based features [10], line spectral frequencies, periodicity-based features [15], linear prediction residual, statistical model, Gaussian mixture model, hidden Markov Model and multi-layer perception (MLP) [15]. Most of these methods have opted a step of sequences as taking sample and sampling at the rate of 8000 per samples followed framing of sample. Mostly the window is approximately taking 30 frames per window. Afterwards, frames are shifted at the fix window size of 40% of actual window. The research work began with cutting the speech data signal into frames as shown in Figure 1 or we can say window before analysis of signal whether it is speech or non-speech; the frame size is 10–30 ms, and frames can be overlapped normally; the overlapping region ranges from 0 to 50% of the frame size. In this research, we have used hamming windowing technique because at the endpoints of the window, there is an amplitude discontinuity. Windowing the signal data ensures that the ends of the signal line up while keeping everything smooth; this considerably lowers 'spectral leakage'. The side lobe suppression of the Hamming window is somewhat better than that of other window approaches. And we used two existing techniques VAD-ANN and VAD-K means.

Figure 1 Windowing overlapping (see online version for colours)

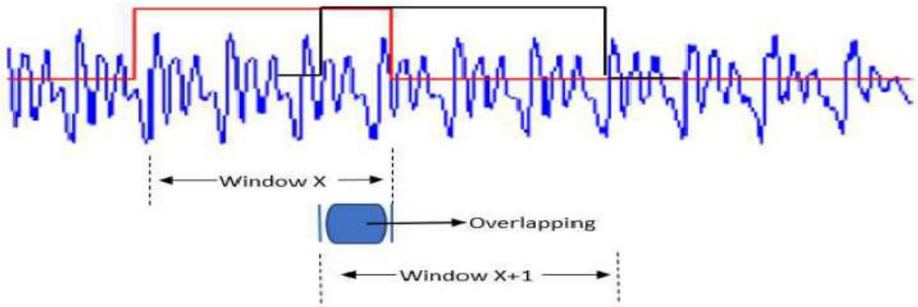
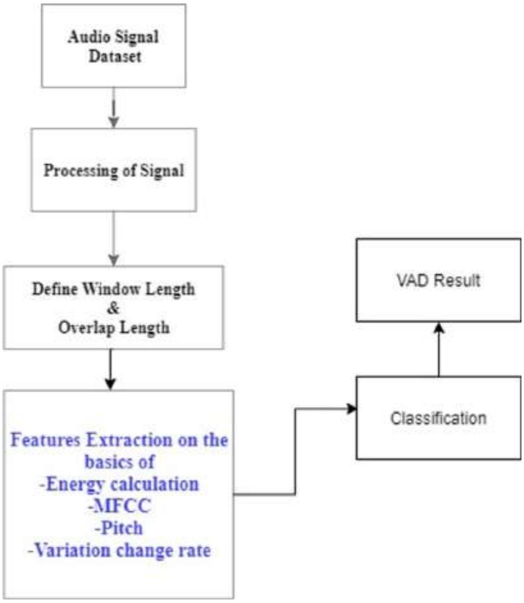


Figure 2 VAD model with the combination of four feature extraction techniques (see online version for colours)


Audio Signal Dataset
Processing of Signal
Define Window Length & Overlap Length
Features Extraction on the basics of -Energy calculation -MFCC -Pitch -Variation change rate
Classification
VAD Result

Then feature extraction method is used and number of researchers has proposed a diverse range of feature extraction methods. Generally, MFCC, pre-emphasis, frame blocking window, DFT spectrum, Mel spectrum, DCTC and Dynamic MFCC are commonly evaluated features reported in previously work. But in this work, we brought a novelty with the implementation of MFCC, Energy, Pitch and new parameter, i.e., low frequency variability as shown in Figure 2.

2.1 Extraction and detection of speech frames in unsupervised speech signal

Speech signals provide a richness of information from a variety of sources, including the message, speaker identity, communication language, and background environment. Various methods for detecting speech fragments in input signal data have been proposed over the years. Short-time energy, zero crossing rate, linear prediction, pitch analysis, cepstral coefficients, and other features are extracted using VAD methods. Extracting features that highlight the characteristics of the intended source is one of the most critical difficulties in voice processing. In this research, we used the following feature extraction VAD methods and these methods divides the given speech signals into voiced and unvoiced classes in a straightforward and quick manner. The method is based on a mix of signal energy estimations and low frequency variability computations. Figure 3 and 4 are illustrating confusion matrices.

1 *Low frequency variability (LFV)*: According to many experiments done on speech segments, it is found that voice segments found in low frequency. In this manuscript, we taken all signal segments which are found in the low frequency as a voice because our ears are more sensitive in the low frequency. The low frequency variability, LFV, is defined by Equation xxii as The low frequency variability, lfv, is defined by equation (i) as,

$$\text{lfv= sum\_of } (|sgnx\ (m)| - |\ sgnx\ (m-1))|/w(n-m) \tag{i}$$

where

$$sgnx\ (m) = 1 \text{ when } x\ (m) \geq 0$$

$$-1 \text{ when } xm < 0\ w\ (n) = 0.5N,\ 0 \leq n \leq N - 10, \text{ otherwise } 0$$

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

The absence or presence of speech in the input signal is indicated by LFV. The frame is regarded unvoiced if the LFV is large, and it is deemed voiced if the rate is low.

It is an experimental proved that speech segment found at low frequency. So, we labelled all speech segment as voice under low frequency irrespective of energy of the speech segments.

2 *Signal energy*

Another metric utilised in the classification of voiced and unvoiced segments is short-time energy calculation. If the incoming frame's energy is strong, it is categorised as a voiced frame; if the energy is low, it is defined as an unvoiced frame [16]. The frame's short-time energy, *x m*, indicated by En, is defined by the equation (ii)

$$\text{En} = [x\ (m) * h(n - m)]^2 \tag{ii}$$

where,

$$h(n) = 0.54 - 0.46 \cos (2\pi n/N - 1),\ 0 \leq n \leq N - 10, \text{ otherwise } 0.$$

3 *MFCC*: MFCCare the Mel frequency cepstral coefficients. MFCC takes into account human perception for sensitivity at appropriate frequencies by converting the conventional frequency to **Mel Scale**, and are thus suitable for speech recognition tasks quite well (as they are suitable for understanding humans and the frequency at which humans speak/utter) [16].

4 *Pitch*: The fundamental frequency of a male voice is 125 Hz, 200 Hz for a female voice, and 300 Hz for a child's voice. Remember that the fundamental frequency changes with the pitch of the speaker's words, so consider it a range rather than an exact figure. The speed at which a speaker's vocal folds (also known as vocal cords) move, as well as their size and how they're used, all contribute to these variances [17].

Afterwards, neural network approaches are applied to evaluate the VAD which is further verified by testing and training. And we got better result in ANN as compare to SVM.

Figure 3 Performance matrices 1 based upon window overlapping (see online version for colours)

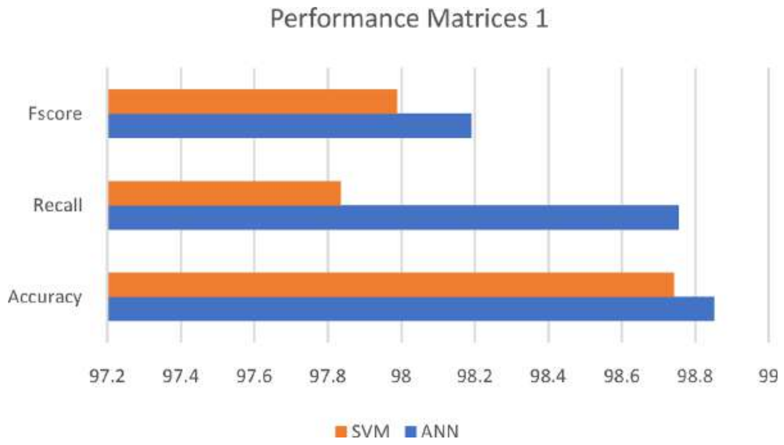
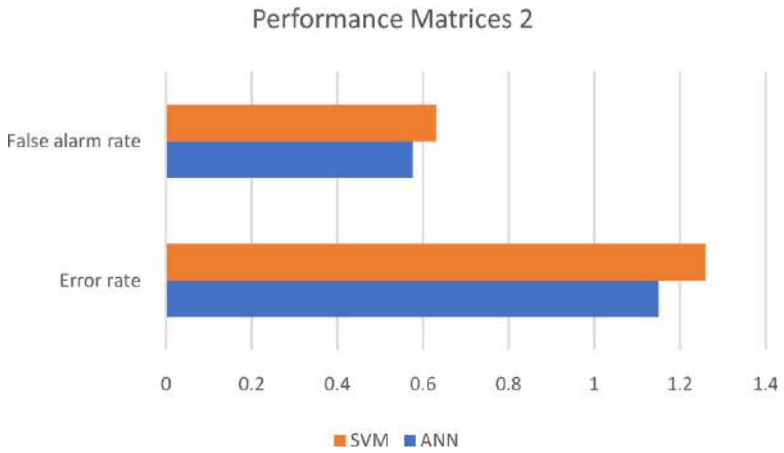


Figure 4 Performance matrices 2 based upon window overlapping (see online version for colours)



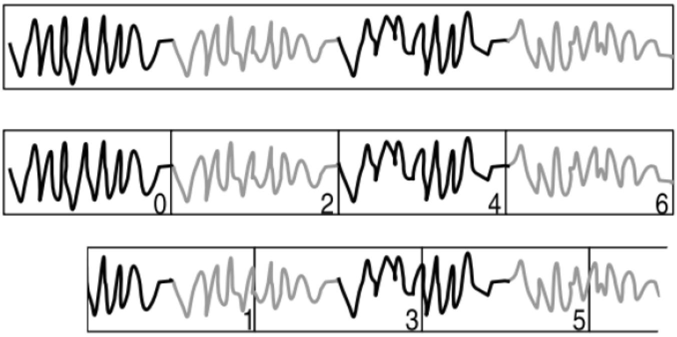
Further, in our research, first we labelled the data as we have worked on unsupervised dataset. We used updated VAD-ANN method to bring the novelty in our research where we used energy-based feature extraction with integration of MFCC, pitch [18] and new method introduced, i.e., low frequency variability. And we got better results as shown in Table 3.

Table 3 Performance matrix based upon SVM, ANN and ANN updated

Performance *Parameters*	*ANN*	*SVM*	*ANN updated*
Accuracy	98.8506	98.7411	**99.6169**
Recall	98.7539	97.8343	**99.5471**
Fscore	98.1895	97.9875	**99.3875**
Error rate	1.1494	1.2589	**0.38314**
False alarm rate	0.57471	0.62945	**0.19157**
SNR	19.345	18.9451	**24.1497**

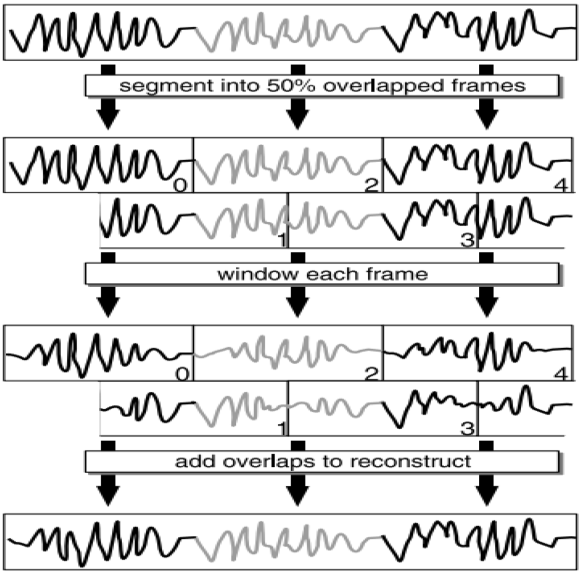
Figure 5 is a demonstration of an entire audio recording (upper waveform) divided into two distinct sequences of analysis windows (two lower waveforms) having 50% overlapped frames such that no short-term auditory characteristic is concealed by spanning the analysis window boundary: At least one of the two analysis streams will appear to be unbroken.

Figure 5 Window and Window overlapping



The most of pervious researcher is on segmentation or feature extraction method and overlooked the effect of mean and variance of audio signal as well as overlap window size optimisation. The window size affects the speed of processing of speech signals and can impact the overall processing time of the VAD method. It is obvious that higher window size gives better output but there is a tradeoff between the processing time and window size [18]. Generally, window size is assumed as 40% for acceptable output. With window overlapping concept, there is no loss of any information. Window overlapping and reconstructing of signal is shown in Figures 5 and 6.

Figure 6 Window overlapping and reconstructed signal


segment into 50% overlapped frames
0
2
4
1
3
window each frame
0
2
4
1
3
add overlaps to reconstruct

As there is requirement to find an optimal window size with lesser processing time, hence investigation is reported at for various overlap window size and different noise level. In this work, a wide of window sizes ranging from 0% to 50% at the step size of 2% are considered to find out the effect of overlap window size. The investigation revealed that the 30–35% window can be considered to perform VAD as it is lower than 40%, hence have faster processing. A well verified dataset of TIMIT is used in this experiment having 156 samples. Additionally, the investigations are reported by employing unsupervised VAD method. Further, the other steps are executed with same manner that is opted by various researchers and reported in this report [19]. Also, we worked on the improvement of SNR ratio with the help of K-means clustering as well as for classifications. The confusion matrices are shown in Table 4.

Table 4 Performance matrix based upon K-means classifier

Performance *Parameters*	*K-means*
Accuracy	99.7263
Recall	99.5072
Fscore	99.5606
Error rate	0.27367
False alarm rate	0.13684
SNR	25.6158

3 Methodology

This work is focused on the accurate and faster unsupervised VAD method. Earlier, the experiment executed by the researcher of this domain has considered that harmonic level or fractal dimension to classify the presence of speech or absence of speech in an audio signal. In this manuscript, mean and variation of signal along with other more three feature extractions methods are considered to achieve the detection of speech in an audio signal. According to best of our knowledge this is a first attempt in the direction of unsupervised VAD. The first step is to consider a frame level speech signal form an utterance which is classified to ‘a’ number of frames in each signal followed by the filtration of low- and high-level signals. For this simulation we have considered a fixed frame size of 480 frame/s and window size is ratio of number of frames to frame shift i.e., 488 frames/s [20]. Further, the means of this entire audio signal is calculated by taking the summation of audio signal to number of frames. Similarly, the variance of the signal is calculated to set a threshold value for the detection of speech signal. If the mean and variance of the signal is greater than the calculated threshold value, than speech signal is present otherwise no speech signal is present n the audio signal. The algorithm for the proposed approach is given in Table 5.

Table 5 Proposed algorithm

```
----------------------------------------------------------------------------------------------------------
Sig: original audio signal to process
S: Sub audio sample
Fs: Audio frequency
N: Window length
Os: Overlapping size
Inppath: File path of audio signal
$S_{Freq}$: Signal in terms of frequency
E: Energy of audio sample
V: Variation of frequency components
LFV: Low frequency variability
$E_{th}$: energy Threshold
$LFV_{th}$: Low frequency variability threshold
----------------------------------------------------------------------------------------------------------
Sig, Fs= readaudio(Inppath)
Count=length(Sig);
S=Sig of length N*Fs with overlapping size: Os,
for i within range of Count
p1: Energy Calculation
$E(n) = \sum_{m=0}^{N-1} S^2(m)$
            where $S$ if given by
```

```
        $S = \left[s(m)*w(n-m)\right]^2$ ,
where, $w$ is hamming window function given by
        $w(n) = 0.54–0.46*\cos(2n\ n/ (N–1)),\ 0 < n < N–1$,
        $w(n) = 0$, otherwise
        where $N$ is number of samples in window with overlapping ‘Os’, ‘n’ is frame shift
p2: Low frequency variability (LFV)
$V(m) = \mathrm{sign}\,(S_{Freq}(m) - S_{Freq}(m-1))$ ,
where, $S_{Freq}$ is signal’s frequency sample near to 0 representing low frequency components
LFV($n$) = $\sum_{m=0}^{N-1} V(m) > 0$ , $\begin{cases} V(m) = 1, s(m) \geq 0 \\ V(m) = -1, s(m) < 0 \end{cases}$
p3: Decision Phase:
$E_{th} = 0.1$
$LFV_{th}$ = mean(LFV($n$)) + std(LFV($n$))
if $E(n) > E_{th}$ and $>LFV(n) > LFV_{th}$
        Decision=Voiced
else
        Decision=Non-Voiced
end
featureSet=Extract Features() # function to extract features
createDataset(featureSet,Decision) # Storage of extracted information to form final dataset
Move to next sub sample
end
Final Dataset with features as inputs and decision as Labels.
```

Further to examine the effect of overlap window size, ranges of different size are considered ranging for 0 to 25. The observations are recorded at diverse range of noise level varied from 5 dB to 25 dB. Further, to consider the worst conditions, random noise is introduced to check at effectiveness of the proposed method in real time scenarios. In this research work, we trained our network also because we are working on unsupervised dataset.

3.1 Experiential learning and testing

The weights in a neural network must be altered in order for the network to accurately acquire the desired output, and this must be done in order to ensure that the network is operating at maximum efficiency. There are three fundamental approaches to teaching a neural network new information. There are three types of learning: supervised learning, reinforced learning, and unsupervised learning.

3.2 Learning under supervision

The training data for neural networks that apply supervised learning must be organised as input vectors, and the data must be organised as input vectors. Next, a target vector is required, which determines how successfully each input is learned and serves as a guide for making adjustments to weight values in order to remove errors.

A single run through the network is used to calculate the output of a feedforward neural network for a specific pattern that has been entered. The term 'supervised learning' refers to the most fundamental class of machine learning algorithms. In order to function properly, these algorithms, as their name implies, require direct supervision [21].

The algorithm in this type of learning is spoon-fed data that has been labelled/ annotated by humans, which is called reinforcement learning. This data contains information about the types of objects of interest as well as their locations.

The algorithm learns from the annotated data after the training phase is completed and predicts the annotations of new data that was previously unknown to the algorithm [22] after the training phase is completed. In this section, we'll look at some of the most often used supervised learning algorithms.

The algorithms are used are neural networks, decision trees, random forest, K-nearest neighbours, linear regression, logistic regression, and SVMs [23].

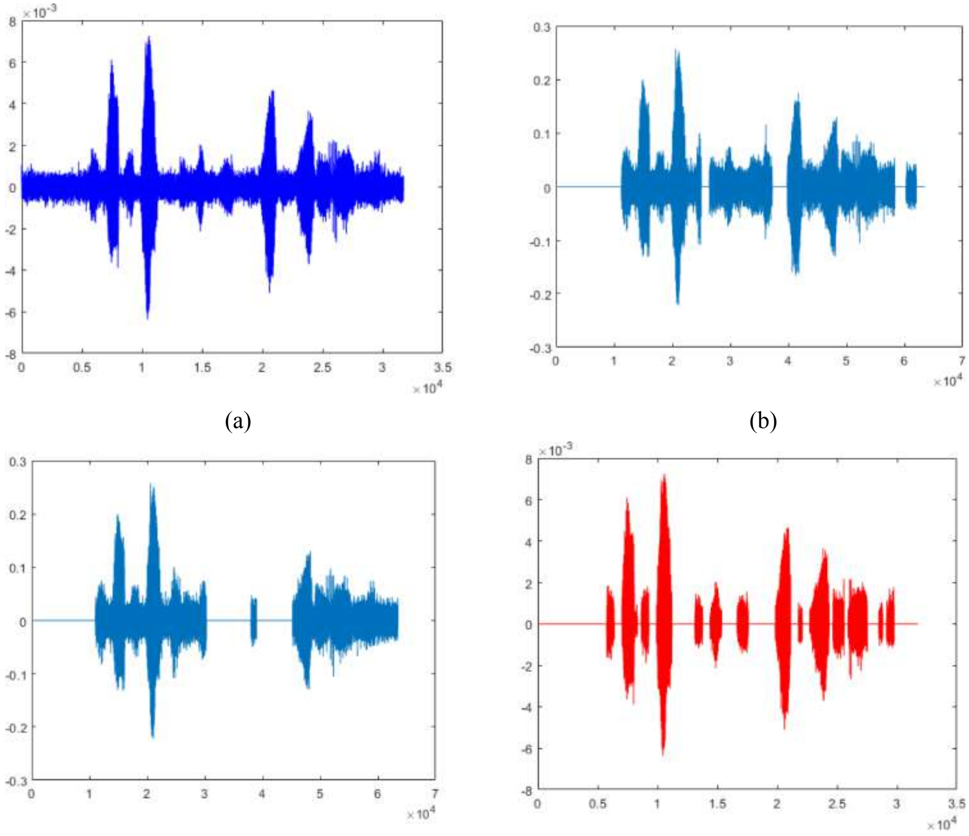
3.3 Unsupervised learning is a type of learning that occurs without supervision

In a neural network that does unsupervised learning, there are no target outputs or signals that indicate if a particular operation was successful or unsuccessful. The network is in charge of finding patterns and regularities in the data that is being fed into it. The patterns in input data can then be used by a network to detect the same patterns in new input data or classify them as a new output class once the network has learned them. Unsupervised learning can be accomplished through the use of the competitive learning rule. In accordance with this notion, a network can contain an input layer that accepts data and a competitive layer of neurons that compete with one another. Each neuron in this network will attempt to respond to specific parts of incoming data, but only one neuron will be activated at a time [12]. Because it is significantly more difficult to achieve than voice recognition, this method of learning is most appropriate for considerably more complicated issues.

Without the assistance or participation of a human, the algorithm attempts to learn and discover valuable properties of classes from the annotated data that has been provided to it [17]. The apriori algorithm, K-means clustering, and other unsupervised learning techniques are examples of common unsupervised learning approaches.

Additionally, the time calculation has been recorded to fastness of proposed method of unsupervised VAD and shown in Figure 7.

Figure 7 Comparison of speech signal with 5 dB Noise signal: (a) SVM, (b) ANN, (c) updated ANN and (d) proposed algorithm with K-means where *x* and *y*-axis represents energy and frames (see online version for colours)



(a)

(b)

(c)

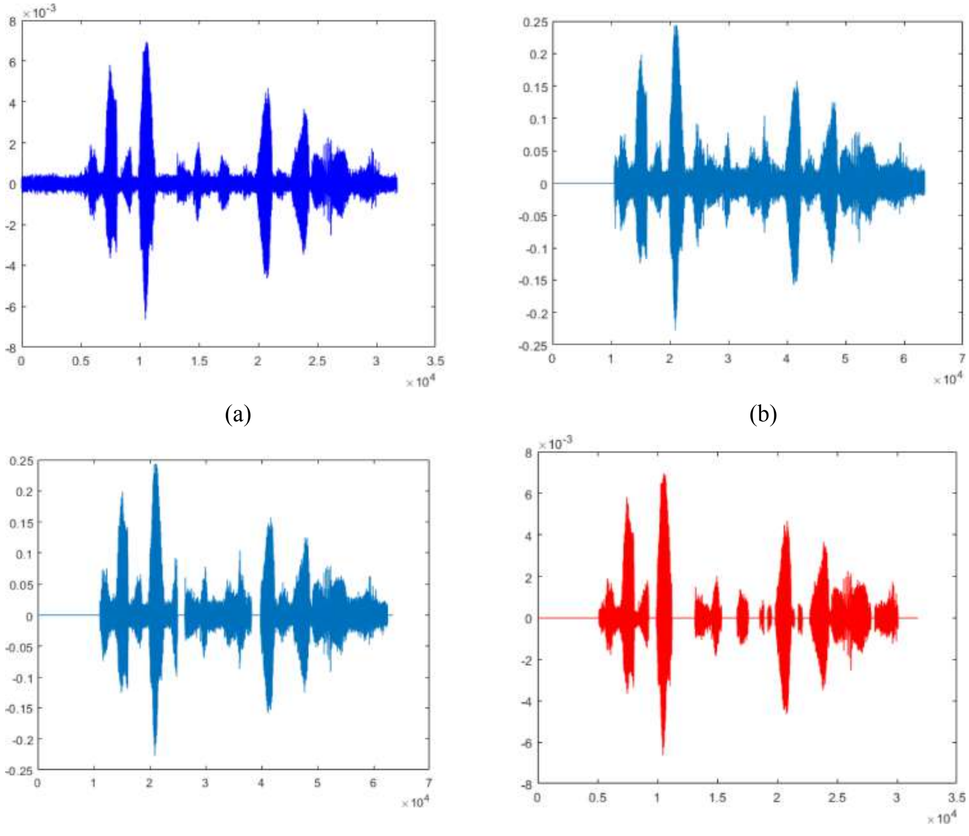
(d)

4 Results and discussion

In this section, the outcomes of the proposed methods have been presented. The result of proposed algorithm is compared with the existing method such as ANN and SVM where the results published by the other authors. Figure 7 illustrates the comparison of voice sample at 5 dB noise levels. To validate the proposed algorithm, simulation for MFCC and reference article is presented. During this evaluation it has been found that proposed approach is out performing compared to earlier published work.

Similar investigations have been reported with noise level 10 dB, 15 dB, 20 dB, 25 dB and at random noise level. The outcomes of the experiments are presented in Figures 8–11, respectively. In every case it has been observed that the proposed signal is more effective and efficient for the detection of speech signal in audio signal.

Figure 8 Comparison of speech signal with 10 dB Noise signal: (a) SVM, (b) ANN, (c) updated ANN and (d) proposed algorithm with K-means where *x* and *y*-axis represents energy and frames (see online version for colours)



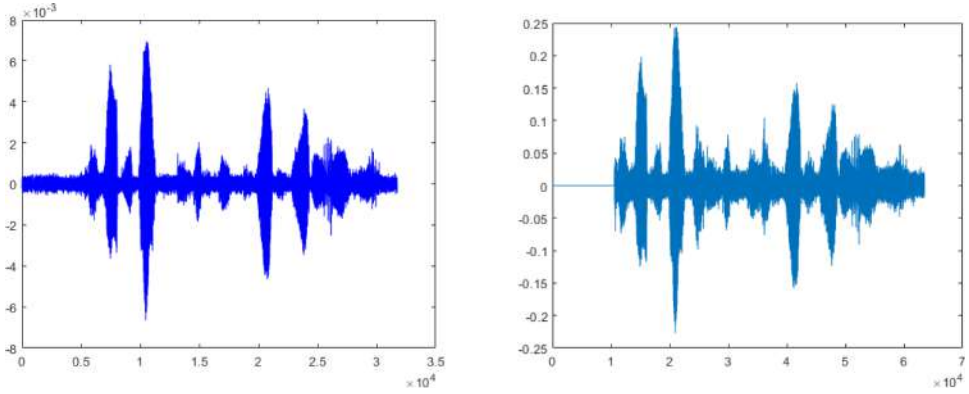
(a)

(b)

(c)

(d)

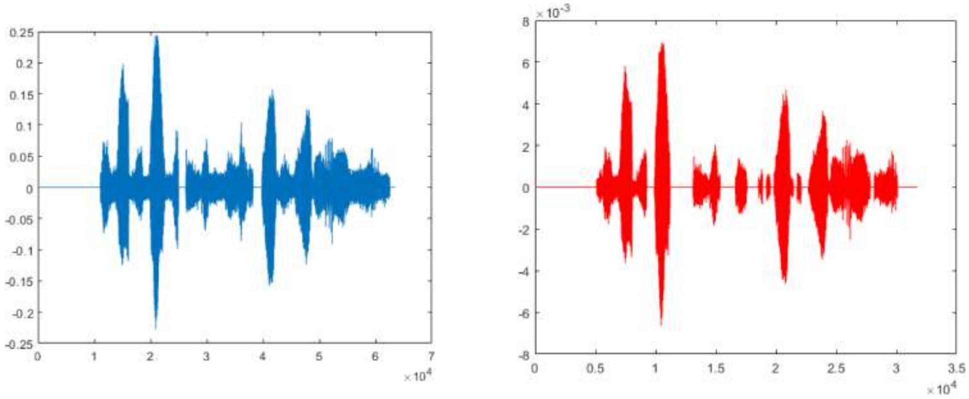
Figure 9 Comparison of speech signal with 15 dB Noise signal: (a) SVM, (b) ANN, (c) updated ANN and (d) proposed algorithm with K-means where *x* and *y*-axis represents energy and frames (see online version for colours)



(a)

(b)

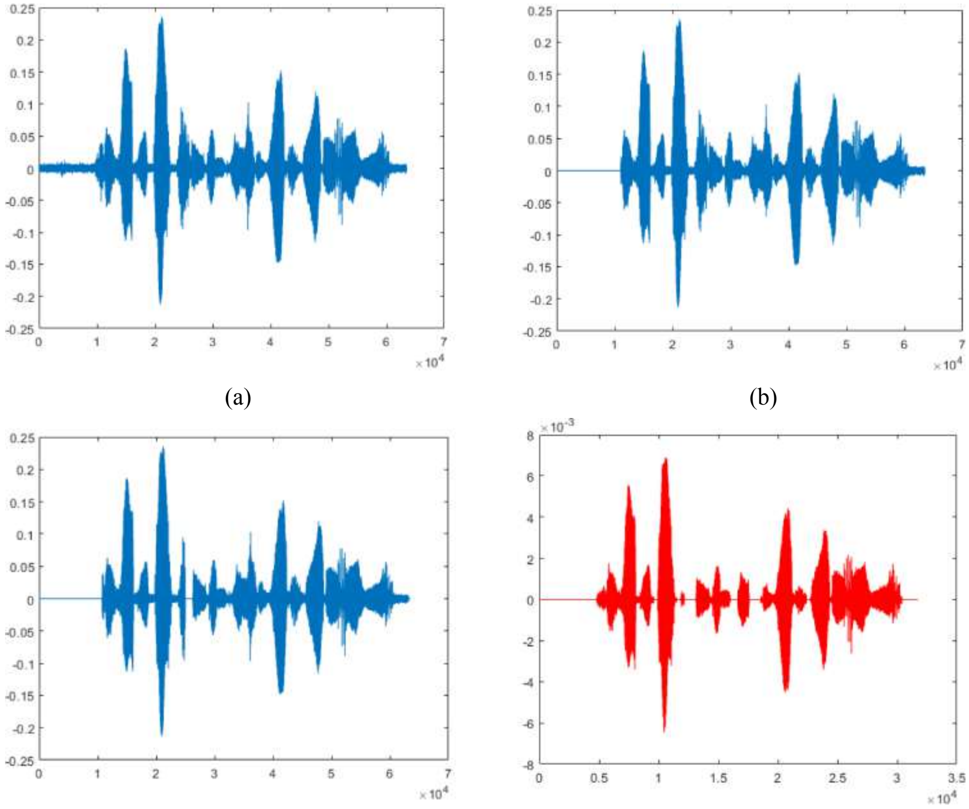
Figure 9 Comparison of speech signal with 15 dB Noise signal: (a) SVM, (b) ANN, (c) updated ANN and (d) proposed algorithm with K-means where *x* and *y*-axis represents energy and frames (see online version for colours) (continued)



(c)

(d)

Figure 10 Comparison of speech signal with 20 dB noise signal: (a) SVM, (b) ANN, (c) updated ANN and (d) proposed algorithm with K-means where *x* and *y*-axis represents energy and frames (see online version for colours)



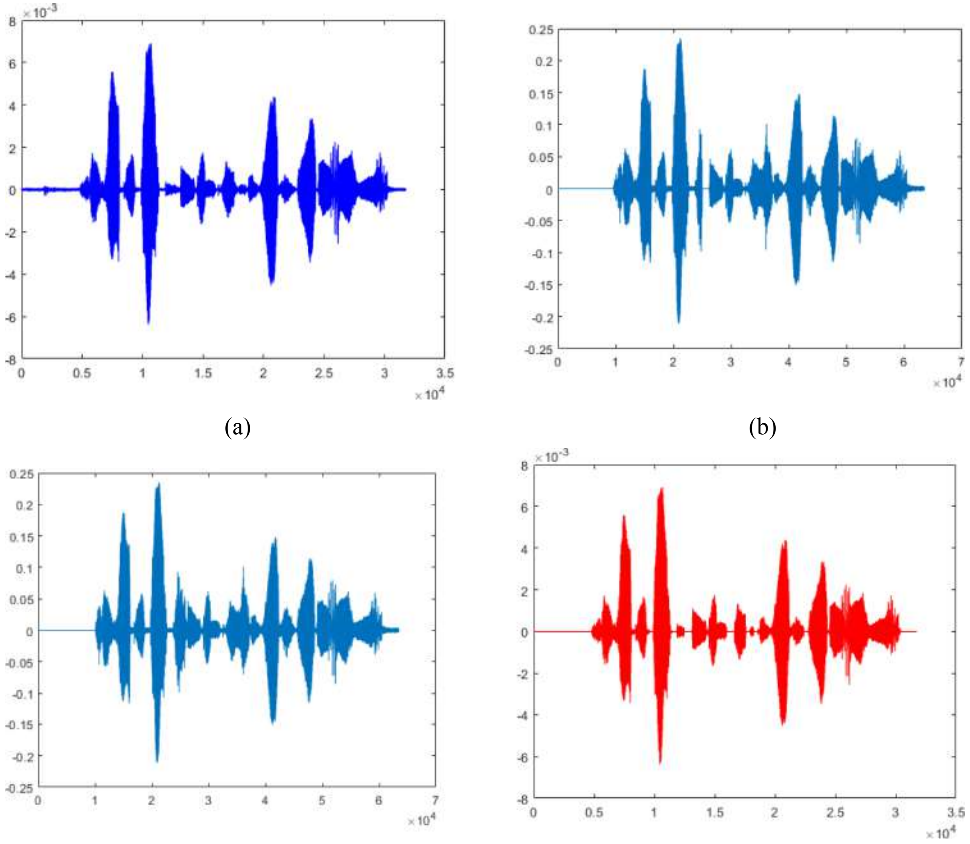
(a)

(b)

(c)

(d)

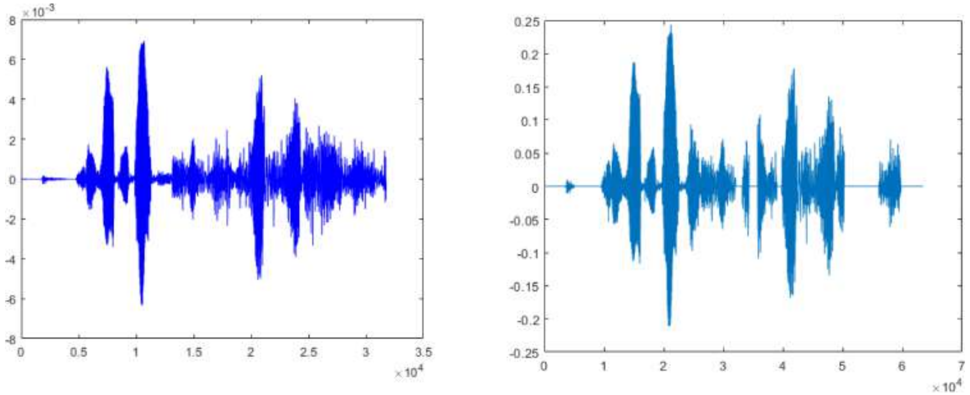
Figure 11 Comparison of speech signal with 25 dB Noise signal: (a) SVM, (b) ANN, (c) updated ANN and (d) proposed algorithm with K-means where *x*-axis represents energy and *y*-axis represents frames (see online version for colours)



(a) (b)

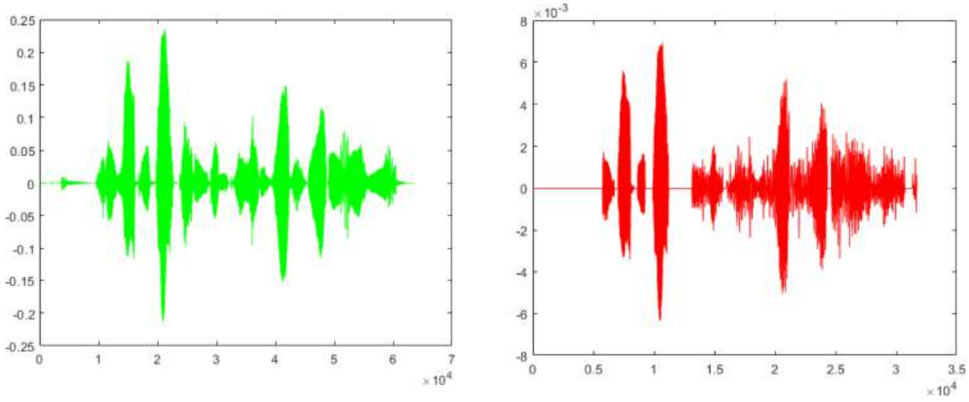
(c) (d)

Figure 12 Comparison of speech signal with random noise signal: (a) SVM, (b) ANN, (c) updated ANN and (d) proposed algorithm with K-means where *x*-axis represents energy and *y*-axis represents frames (see online version for colours)



(a) (b)

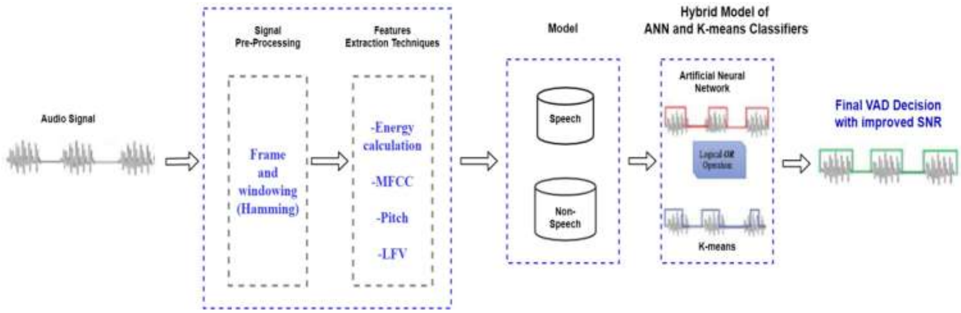
Figure 12 Comparison of speech signal with random noise signal: (a) SVM, (b) ANN, (c) updated ANN and (d) proposed algorithm with K-means where *x*-axis represents energy and *y*-axis represents frames (see online version for colours) (continued)



(c) (d)

To improve the signal to noise ratio, we used new technique, i.e., hybridisation model of artificial neural network and k-means classifiers as shown in Figure 13. Here, in this approach we used logical or operation where it will consider as segment is voice either detected by ANN or K-means classifiers and result is shown in Figure 12. So, we got improved complex matrices.

Figure 13 An improved model for unsupervised voice activity detection (see online version for colours)


Audio Signal
Signal Pre-Processing
Frame and windowing (Hamming)
Features Extraction Techniques
-Energy calculation
-MFCC
-Pitch
-LFV
Model
Speech
Non-Speech
Hybrid Model of ANN and K-means Classifiers
Artificial Neural Network
Logical OR Operator
K-means
Final VAD Decision with improved SNR

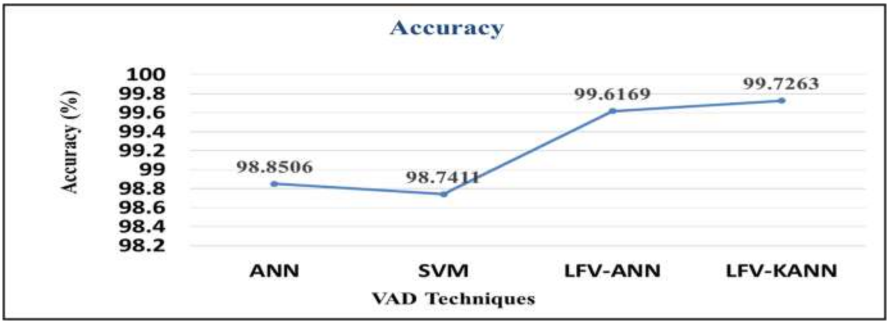
Further, the investigations are carried out for the complexity matrices of SVM, ANN, Updated ANN and proposed approach with K-means. Following are the performance matrices that we have taken to analyse the optimal window length and window overlapping size:

- Accuracy
- Recall
- Fscore

- Error rate
- False alarm rate

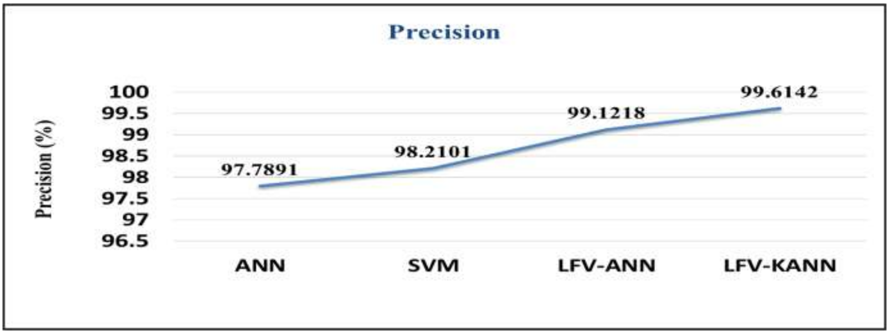
Improved model of unsupervised voice activity detection has been proposed. Labelling is done with the help of speech feature extractions techniques. Trained the network further using two classifiers ANN and K-means clustering. Now, there is compassion between existing VAD techniques with the improved model of VAD. Later, different amount of noise is added on improved VAD model to find out the precision, accuracy, f score, error rate, SNR, FAR and recall. Figure 14 represents the accuracy percentage comparison, Figure 15 represents the precision percentage comparison, Figure 16 represents the recall percentage comparison, Figure 17 represents the F-score percentage comparison, Figure 18 represents the error percentage comparison, Figure 19 represents the FAR percentage comparison, Figure 20 represents the SNR percentage comparison, i.e., performance comparison at different rate of noise added.

Figure 14 Performance metric – accuracy (see online version for colours)



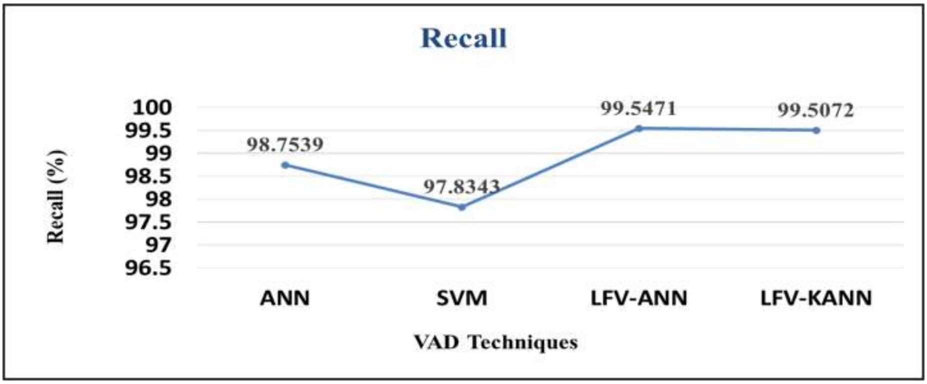
As shown in Figure 14, it concludes that proposed model LFV-KANN outperforms existing techniques in terms of accuracy.

Figure 15 Performance metric – precision (see online version for colours)



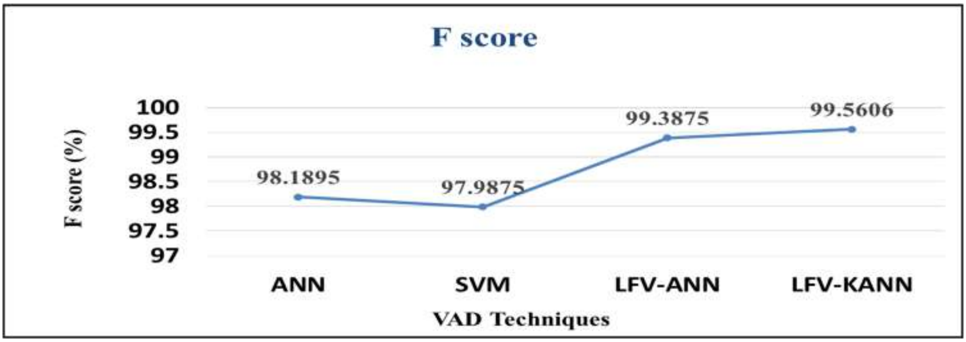
As shown in Figure 15, it concludes that proposed model LFV-KANN outperforms existing techniques in terms of precision.

Figure 16 Performance metric – recall (see online version for colours)



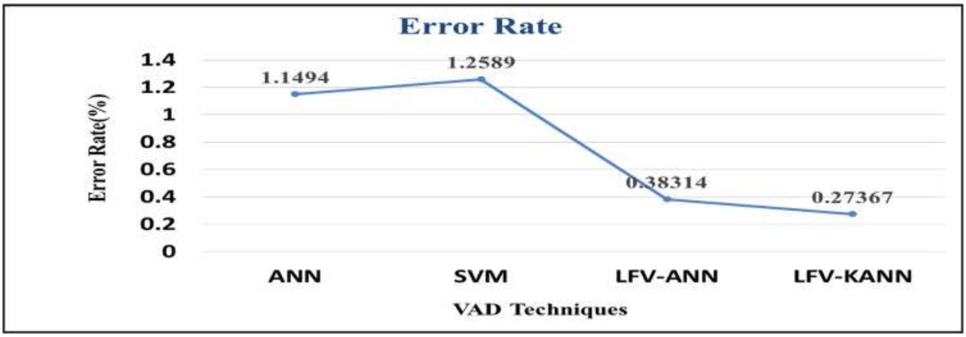
As shown in Figure 16, it concludes that proposed model LFV-KANN outperforms existing techniques in terms of Recall.

Figure 17 Performance metric – F-score (see online version for colours)



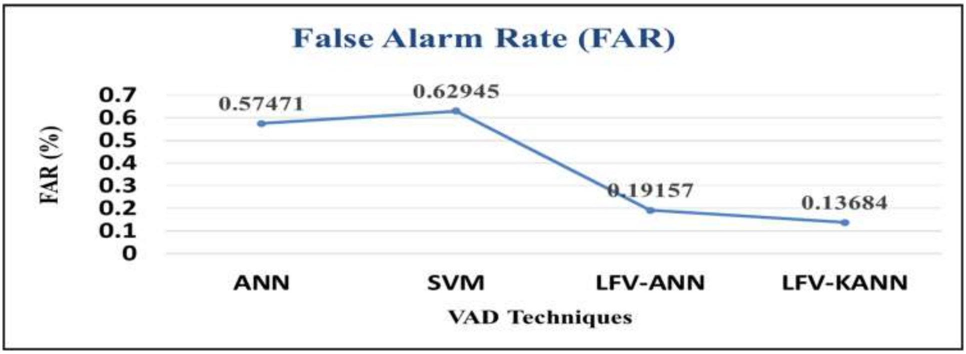
As shown in Figure 17, it concludes that proposed model LFV-KANN outperforms existing techniques in terms of F score.

Figure 18 Performance metric – error rate (see online version for colours)



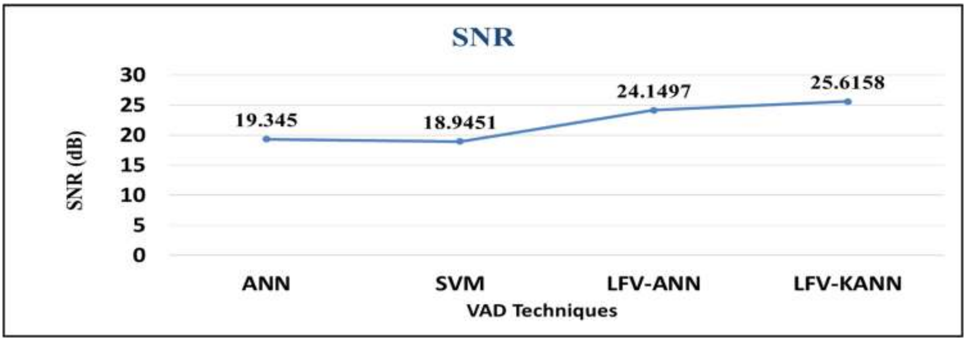
As shown in Figure 18, it concludes that proposed model LFV-KANN outperforms existing techniques in terms of error rate.

Figure 19 Performance metric – FAR (see online version for colours)



As shown in Figure 19, it concludes that proposed model LFV-KANN outperforms existing techniques in terms of false alarm rate.

Figure 20 Performance metric – SNR (see online version for colours)



As shown in Figure 20, it concludes that proposed model LFV-KANN outperforms existing techniques in terms of SNR.

Table 6 concludes that proposed model LFV-KANN outperforms existing techniques in terms of accuracy, precision, recall, F score, error rate, false alarm rate and SNR. Table 7 illustrate concludes that proposed model LFV-KANN efficiently handles increase in noise power by hybridisation of two classifiers: ANN and K-means clustering.

Table 6 Performance metric comparisons

Parameters *Metrics*	*ANN*	*SVM*	*LFV-ANN*	*LFV-KANN*
Accuracy	98.8506	98.7411	99.6169	99.7263
Precision	97.7891	98.2101	99.1218	99.6142
Recall	98.7539	97.8343	99.5471	99.5072
Fscore	98.1895	97.9875	99.3875	99.5606
Error rate	1.1494	1.2589	0.38314	0.27367
False alarm rate	0.57471	0.62945	0.19157	0.13684
SNR	19.345	18.9451	24.1497	25.6158

Table 7 Performance metric during different noise added

Performance *Measurement*			*LFV-KANN*		
Accuracy	98.6864	99.1527	99.7263	99.8358	99.7263
Precision	98.127	99.4316	99.5152	99.7905	99.6142
Recall	98.127	99.4316	99.5152	99.7905	99.6142
Fscore	98.3158	99.3355	99.5671	99.7369	99.5606
Error rate	1.3136	0.43788	0.27367	0.1642	0.27367
False alarm rate	0.65681	0.21894	0.13684	0.082102	0.13684
SNR	*18.7578*	*23.5674*	*25.6158*	*27.839*	*25.6158*
Noise added (dB)	**5**	**10**	**15**	**20**	**Actual dataset**

5 Conclusion

It has been observed that our proposed algorithm's is dependent of overlap window size in line to extraction features as it is influenced by the overlap window size. According to result shown in previous section, proposed model LFV-KANN outperforms existing techniques in terms of accuracy, precision, recall, F score, error rate, false alarm rate and SNR. Along with this, the proposed model LFV-KANN efficiently handles increase in noise power by hybridisation of two classifiers: ANN and K-means clustering. This proposed model can be used in the future with advanced technologies in a variety of application scenarios such as smart homes, healthcare automation, smart city or industry-driven projects. Additionally, various techniques can be used to enhance the results further such as multiobjective evolutionary optimisation [24], deep learning [25], and genetic algorithm based deep learning, [26] etc.

References

1 Elton, R.J., Mohanalin, J. and Vasuki, P. (2021) 'A novel voice activity detection algorithm using modified global thresholding', *Int. J. Speech Technol.*, Vol. 24, No. 1, pp.127–142.

2 Singh, S.P. and Jaiswal, U.C. (2021) 'Audio classification using grasshopper□ride optimization algorithm□based support vector machine', *IET Circuits Devices Syst.*, Vol. 15, No. 5, pp.434–447.

3 Kwon, H., Yoon, H. and Park, K.W. (2020) 'Acoustic-decoy: detection of adversarial examples through audio modification on speech recognition system', *Neurocomputing*, Vol. 417, pp.357–370.

4 Mohamed Ismail Yasar Arafath, K. and Routray, A. (2019) 'Automatic measurement of speech breathing rate', *2019 27th European Signal Processing Conference (EUSIPCO)*, pp.1–5, doi: 10.23919/EUSIPCO.2019.8902730.

5 Zoulikha, M. and Djendi, M. (2018) 'A new robust forward BSS adaptive algorithm based on automatic voice activity detector for speech quality enhancement', *Int. J. Speech Technol.*, Vol. 21, pp.1007–1020.

6 Zhigang, Z. and Junqin, H. (2015) 'An adaptive voice activity detection algorithm', *Int. J. Smart Sens. Intell. Syst.*, Vol. 8, No. 4, pp.2175–2194, https://doi.org/10.21307/ijssis-2017-848

7 Hsieh, C.H., Feng, T.Y. and Huang, P.C. (2009) 'Energy-based VAD with grey magnitude spectral subtraction', *Speech Commun.*, Vol. 51, No. 9, pp.810–819.

8 Wang, Y. and Lee, L. (2014) 'Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning', *IEEE/ACM Trans. Audio Speech Lang.*, Vol. 9290, No. c, pp.1–16, https://doi.org/10.1109/TASLP.2014.2387413

9 Mustafa, M.K., Allen, T. and Appiah, K. (2015) *Research and Development in Intelligent Systems XXXII*, https://doi.org/10.1007/978-3-319-25032-8

10 Sunil Kumar, S.B. and Sreenivasa Rao, K. (2016) 'Voice/non-voice detection using phase of zero frequency filtered speech signal', *Speech Commun.*, Vol. 81, pp.90–103, https://doi.org/10.1016/j.specom.2016.01.008

11 Amardeep, A. (2013) *Methods for Improving Voice Activity Detection in Communication Services*, http://www.diva-portal.org/smash/record.jsf?pid=diva2:588802%0Ahttp://uu.diva-portal.org/
smash/get/diva2:588802/FULLTEXT01.pdf

12 Pannala, V. and Yegnanarayana, B. (2021) 'A neural network approach for speech activity detection for Apollo corpus', *Comput. Speech Lang.*, Vol. 65, p.101137, https://doi.org/10.1016/j.csl.2020.101137

13 Devi, T.M., Kasthuri, N. and Natarajan, A.M. (2013) 'Environmental noise reduction system using fuzzy neural network and adaptive fuzzy algorithms', *Int. J. Electron.*, Vol. 100, No. 2, pp.205–226, https://doi.org/10.1080/00207217.2012.687192

14 Joseph, S.M. and Babu, A.P. (2016) 'Wavelet energy based voice activity detection and adaptive thresholding for efficient speech coding', *Int. J. Speech Technol.*, Vol. 19, No. 3, pp.537–550, https://doi.org/10.1007/s10772-014-9240-x

15 Benzvi, D. and Shafir, A. (2019) 'An ICA algorithm for separation of convolutive mixture of periodic signals', *2018 IEEE International Conference on the Science of Electrical Engineering in Israel, ICSEE 2018*, Vol. 2, No. 4, pp.273–283, https://doi.org/10.1109/ICSEE.2018.8646002

16 Meduri, S.S. and Ananth, R. (2011) *A Survey and Evaluation of Voice Activity Detection Algorithms*, Medieteknik.Bth.Se, http://medieteknik.bth.se/fou/cuppsats.nsf/all/a1e356336cee2e3ac125799800566259/$file/BTH2011_Meduri.pdf

17 Korkmaz, Y. and Boyacı, A. (2020) 'Unsupervised and supervised VAD systems using combination of time and frequency domain features', *Biomed. Signal Process. Control*, Vol. 61, pp.1–8, https://doi.org/10.1016/j.bspc.2020.102044

18 Hajarolasvadi, N. and Demirel, H. (2019) '3D CNN-based speech emotion recognition using k-means clustering and spectrograms', *Entropy*, Vol. 21, No. 5, p.479, https://doi.org/10.3390/e21050479

19 Elton, R.J., Vasuki, P. and Mohanalin, J. (2016) 'Voice activity detection using fuzzy entropy and support vector machine', *Entropy*, Vol. 18, No. 8, https://doi.org/10.3390/e18080298

20 Nóbrega, R. and Cavaco, S. (2009) 'Detecting key features in popular music□: case study – singing voice detection', *Second International Workshop on Machine Learning and Music at ECML-PKDD'09*, pp.7–12, http://www.ecmlpkdd2009.net/wp-content/uploads/2008/09/machine-learning-and-music.pdf#page=13

21 Ghahabi, O., Zhou, W. and Fischer, V. (2018) 'A robust voice activity detection for real-time automatic speech recognition', *Proc. of ESSV 2018*, p.644283, http://essv2018.de/wp-content/uploads/2018/03/30_OmidGhahabi_ESSV2018.pdf

22 Gelly, G. and Gauvain, J.L. (2018) 'Optimization of RNN-based speech activity detection', *IEEE/ACM Trans. Audio Speech Lang.*, Vol. 26, No. 3, pp.646–656, https://doi.org/10.1109/TASLP.2017.2769220

23 Hou, Q., Li, C., Kang, M. and Zhao, X. (2021) 'Intelligent model for speech recognition based on SVM: a case study on English language', *J. Intell. Fuzzy Syst.*, Vol. 40, No. 2, pp.2721–2731, https://doi.org/10.3233/JIFS-189314

24 Kaur, M. and Singh, D. (2021) 'Multiobjective evolutionary optimization techniques based hyperchaotic map and their applications in image encryption', *Multidimension. Syst. Signal Process.*, Vol. 32, No. 1, pp.281–301, https://doi.org/10.1007/s11045-020-00739-8

25 Kaushik, H., Singh, D., Kaur, M., Alshazly, H., Zaguia, A. and Hamam, H. (2021) 'Diabetic retinopathy diagnosis from fundus images using stacked generalization of deep models', *IEEE Access*, Vol. 9, pp.108276–108292, doi: 10.1109/ACCESS.2021.3101142

26 Singh, D., Kumar, V., Kaur, M., Jabarulla, M.Y. and Lee H-N. (2021) 'Screening of COVID-19 suspected subjects using multi-crossover genetic algorithm based dense convolutional neural network', *IEEE Access*, Vol. 9, pp.142566–142580, doi: 10.1109/ACCESS.2021.3120717.