Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.13, No.3, 2022

Voice Activity Detection: Fusion of Time and Frequency Domain Features with A SVM Classifier

Sheriff Alimi* Oludele Awodele
Department of Computer Science, Babcock University, Ilesan-Remo, Nigeria
*E-mail mail of the corresponding author: alimi0356@pg.babcok.edu.ng

Abstract
Voice activity detection (VAD) discriminates between segments of an audio signal that has speech content from the ones with either noise or silence. It is deployed as the front-end of some speech processing applications such as voice recognition, and speaker recognition to improve their performance in terms of accuracy and efficiency. It is also used in the communication system to bring about efficient utilization of transmission bandwidth by ensuring only segments of the audio signal with voice activity are encoded and transmitted. In this work, the VAD algorithm was implemented using a features-fusion strategy. In the pre-processing stage, contents outside the human auditory frequency range were removed with the aid of a digital Butterworth bandpass filter. The signal was then fragmented into frames from where time-domain features (zero-crossing, standard deviation, normalized envelope, kurtosis, skewness, and root-mean-square energy.) and frequency-domain features (13MFCCs) were extracted and then combined to form a feature representation of each frame. Recursive feature elimination was applied to the dataset to reduce the features to seven (7) which was used to train a Support Vector Machine (SVM) to be able to distinguish between voiced and unvoiced frames. A State-of-art performance was recorded by this simple SVM-based VAD system with an accuracy of 100%, recall of 100%, precision of 100% and F1 score of 100% which is at par with similar implementations which utilizes a complex architecture of deep neural network with high computational cost and training time.
Keywords: Voice activity detection, fusion strategy, support vector machine, frequency domain features, time domain features
DOI: 10.7176/CEIS/13-3-03
Publication date: May 31st 2022

1. Introduction

In speech signals, some segments have voice activity, and utterances while the other sections have no such activity and are considered silent parts of the signal. In some speech processing applications, it is important to be able to distinguish between the silence and voiced sections, most times the whole speech signal is partitioned into a much smaller unit called a frame whose size is determined by application type and in situations where transforming the discrete-time signal to frequency domain is required, the size is chosen to optimize the performance of discrete Fourier transform (DFT) operation. It then becomes imperative for such an application to be able to decipher at frame level if there is voice activity or not.

In a typical speech conversation, it was discovered that the speaker only talks for 40% of the time while the remaining 60% of the time the speaker is idle, the absence of speech activity, the idle part which lacks human utterances considered silence (Krishnakumar & Williamson,2019; Bäckström,2017).

Voice Activity Williamson, (AD) is primarily the analysis of audio, and speech signals to determine the regions with an utterance (Lavechin et al., 2020; Bäckström,2017), so the VAD algorithm function as a discriminator in identifying the speech part of an audio signal and eventual discard region of silence.

VAD has been tremendously used as the front-end of so many speech processing applications such as speaker recognition, speech recognition, speech enhancement, gender identification and age identification(Mohammed and Hassan, 2020)VAD has significantly helped the back-end applications to improve their performance accuracy and overall processing time(Dey et al., 2019) as silent segments are never passed to them for processing.

In digital telephony like GSM technology, VoIP technology and other related communication systems where there is always a contest for the bandwidth available for the transfer of information from one end of the communication to another, it is a big waste, encoding and transferring the silent part of a talk which stands at 60% of the conversation period over a contested transmission media. This is inefficient utilization of scarce communication resources. Many speech processing applications such as speaker recognition, speaker verification, automatic speech recognition, emotion recognition and gender detection deal with classification problems; the use of the silent sections of the speech in both training and validating such systems will yield unsatisfying accuracies. To address these problems, voice activity detection (VAD) will be very useful in discriminating between sections with utterances and those without, to discard the silent ones so that voiced ones are passed for further processing by the back-end system. The resultant effect of the introduction of VAD is that it brings about efficient utilization of transmission bandwidth and improves the accuracies of speech processing applications

mentioned earlier.

Quite a lot of different algorithms have been utilized in achieving voice activity detection starting with adaptive thresholding, spectral subtraction(Pang, 2017) and a more recent approach which involves extraction of features from the speech signal and training of machine learning models such as chain of Deep Neural Networks(Krishnakumar & Williamson, 2019) and the ensemble of SVMs(Dey et al., 2019) to function as classifiers. Time-domain features such as zero crossing, envelope energy etc. have been used as features for training machine learning classifiers which result in satisfactory performance on clean signals, however, the performance degrades with a signal that has a low signal-to-noise ratio (SNR); incorporation of spectral features have proven useful in such situation(Dwijayanti et al, 2018).

The VAD algorithms based on chains of Deep Neural Networks yield good results, but their architectures are complex which translates to high computational costs. Therefore, there is a need to explore other approaches that give a state-of-art performance while relying on the use of a simple architecture thereby eliminating the challenge of demand for high compute resources.

The aim here is to develop a simple voice detection algorithm that functions as a binary classifier that distinguishes between frames with and without voice activity in a speech signal, and the algorithm should be robust to noise outside the human hearing frequency range. This will be realized using a band-pass filter, and extraction of time and frequency domain features from each labelled frame with SVM for the classification.

It is pertinent then that the VAD algorithm should be simple without undue complexity, with high accuracy and resilient to noise. The VAD algorithm of this work matches the above-described attributes expected of a typical VAD system.

2. Review of Related Work

There are quite a good number of research done in the area of voice activity detection using various techniques and algorithm. This section focuses on the review of such work.

Elton et al, (2016), leveraging spectral noise removal techniques where the noise was estimated during speech inactive period couple with fuzzy entropy for feature extraction yielded a state-of-art performance with SVM as a binary classifier. Accuracy of 93% was achieved in this work with noisy speech signals where the signal-to-noise ratio ranged between -10dB and 10dB.

Lee and Ellis, (2006), in their VAD implementation, the autocorrelation function (ACF) for each sub-band in a frame was computed and summed up to form a summarized autocorrelation function (excluding the aperiodic ones). The summarized autocorrelation function (SAC) serves as the basis for classification and this VAD attained an accuracy of 88%. In this work, an assumption was made that the long-time average of the speech autocorrelation function is close to the noise autocorrelation function, and this might not be true all the time. An unsupervised approach was adopted by Park et al, (2017) in the design and implementation of a VAD system. The voice detection algorithm is based on the thresholding of fractal dimension derived using Katz algorithm. The method achieved an average accuracy of 90.45% with audio signals across three types of noise (white noise, car noise and babble noise) with different signal-to-noise-ratio (SNR). The Katz algorithm requires that the complete audio signal is available before computation and detection can be performed and this makes this unsupervised voice activity detection unsuitable for real-time continuous audio signals.

Mohammed & Hassan,(2020) also used a supervised model for distinguishing between voiced and unvoiced frames. Time-domain features were extracted from each frame and a modified fuzzy-c was used to cluster the features into three groups (two for voiced and one for unvoiced frame features). Finally, three ANNs were used for classification. Pang, (2017), implemented Spectrum Energy Based Voice Activity Detection where the signal is separated into two bands, the Low-Frequency band (LFB) and High-Frequency Band (HFB). The total spectrum energy in the LFB and HFB for each of the frames is calculated, EnL and EnH respectively. A five-point moving average filter is applied to EnH which is the noise signal. Mean EnH is subtracted from EnL of each frame and any frame with a residual value greater than a threshold value of 25 is considered a voice frame otherwise it is a silence frame.

Automated extraction of features from raw waveform in voice detection systems and proper handling of domain misclassification was the focus of work done by (Lavechin et al, 2020) based on an end-to-end adversarial neural network architecture. Raw-waveform data are fed into Sincnet CNN which extracts features used to train two different networks simultaneously and the backpropagation from the networks is used to update the learning of the SincNet. The first network is for voice detection and comprises two LSTM and three feed-forward. The second network is for domain detection and comprises an LSTM, temporal pooling and feed-forward.

Krishnakumar and Williamson, (2019) explored Boosted Deep Neural Network, a promising architecture for voice activity detection. The input feature for the voice signal is the multi-resolution cochleagram (MRCG) and two layers of boosted deep neural Network(bDNN) were used. The DNN were either based on an ensemble of CNN, LSTM or dilated CNN. The bDNN architecture based on LSTM has the best performance of 91.4% and

90.7% on seen and unseen noise data respectively and then followed by the dilated CNN architecture.

Zazo et al, (2016), evaluated the performance of the Convolutional, Long Short-Time Memory (CLDNN) VAD algorithm which uses raw waveform as input against VAD algorithms based on standard DNN and LSTM with log-mel as input features. The dataset for the research comprises clean and noisy voice recordings. The raw waveform was applied to the CLDNN architecture for training the model which also automatically extracts needed features for the classification exercise. The CLDNN trained from raw waveform had the best performance of false alarm of 4.1% fixing false rejection at 2% on noisy data. Next in line in terms of performance respectively are the CLDNN and LSTM trained with log-mel features.

Dwijayanti et al,(2018), proposed a deep neural network (DNN)-based VAD with the fusion of log power spectral and speech dynamics, the result of their study shows that VAD with log power spectral had better performance compared to the use of MFCCs and MFCCs combined with delta, and delta-delta in both clean and noisy signals.

To explore the performance of the VAD algorithm designed based on two strategies :(1) Fusion of feature vectors and (2) fusion of decisions, Drugman et al, (2016), trained ANN with the result of the fusion of source-related and filter-based features. Also, the two features were separately used to train two different ANNs and the final decision is based on the geometrical mean of the predictions of the two classifiers. The feature fusion and decision fusion of MFCC and source-related features yield an F1-score of 93.6% and 94.8% 8%. In like manner, Sadjadi and New features (the two being excitation-based features) yield f1-score of 94.9% and 95.3%. The result shows clearly that the strategy of decision fusion outperforms that of feature fusion.

In designing a VAD algorithm based on adaptive thresholding that incorporates a false acceptance rate, Chelloug and Farrouki, (2019) considered the first 100ms frames as silent frames which are used to compute threshold values for discriminating between voiced and voice-free frames. The threshold value is updated via adaptive learning. A correct classification rate of 84.6% was achieved by the algorithm with a false acceptance rate of 9.7% and a true rejection rate of 5.7%.

The focus of the work done by Tan et al, (2020), was to develop an unsupervised VAD method, that generalizes well to real-world data, irrespective of whether the data is corrupted by stationary or rapidly changing additive noise. The energy level and pitch are used to determine if a segment (group of consecutive frames) should be considered a voice segment or not. In the second stage, the decision on each frame is based posterior of SNR weighted energy of the frame.

Smartphones are ubiquitous and can be interface with many hearing devices, hence, the need for smartphones VAD. This is the basis of the convolutional neural network App for real-time voice activity detection developed on the smartphone by Sehgal and Kehtarnavaz, (2018). Voice signal is fragmented, and each frame is converted to log-mel filter-bank energy images which were used to train CNN and the CNN based VAD achieved Speech Hit Rate (SHR) and Noise Hit Rate (NHR) of 91.3% and 99% for Real-time application.

Makowski and Hossa, (2020), implemented a voice activity detection algorithm based on quasi-quadrature filters and GMM decomposition for speech and noise. Envelopes of narrow frequency bands were computed using a complex filter and weighting of the envelopes was performed to minimize the contribution from noisy bands. The envelope's mean, variances and moment are the statistics used for thresholding in making the decision. The VAD algorithm was compared with a similar algorithm using false rejection and false acceptance probabilities as well as cost function at different SNR, this VAD algorithm recorded a better performance on all the mentioned metrics.

An ensemble of five SVM-VAD was trained with extracted features from voice frames by Dey et al, (2019) and the VAD distinguishes voice segments into two classes, one that contains speech and the second one contains noise, or silence, or music. The prediction accuracy of 87.4% was achieved with the ensemble of five SVM.

Park et al, (2017) implemented voice activity detection based on spectral energy. This work used computed spectral energy per frame for the frequency range associated with humans and, this was made possible with the short-time-Fourier transform (STFT). The spectral energy threshold was the basis for classifying frames as either voice frames or silent frames.

A voice activity detection algorithm based on long-term pitch information was implemented by Yang, He, Qu, & Zhang, (2016). In this system, the long-term pitch divergence (LTPD) is computed from features extracted after the decomposition of an audio signal into 88 frequency bands by multi-rate filter banks. The computed LTPD and SNR are the basis for the classifications

From the review of the literature on VAD algorithms, many of the realizations were done using complex architectures of Deep Neural Networks ranging from cascaded networks to ensemble networks, of course, good performances were obtained most with a trade-off of high computational cost. This is one of the gaps that demand attention.

In this current work, an attempt will be made to achieve performance that is at par with these complex realizations by using a simple architecture with relatively low computational cost.

3. Methodology

In the design and implementation of the voice activity detection system of this work, four major processing stages are considered which are the pre-process, features extraction, features selection, SVM training and, SVM classifier performance evaluation stages as depicted in figure 1 below.

The voice signals used in this work are obtained from recordings of one of the authors and two of his female colleagues that gave their consent to use their voice recordings for this research.

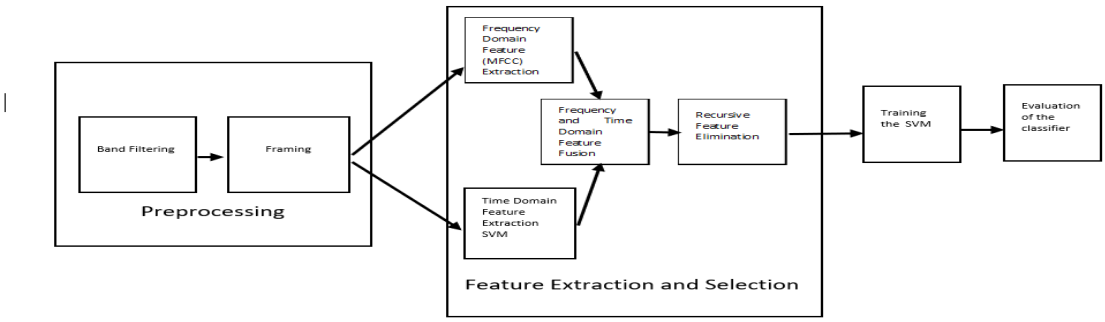

Figure 1. Implementation Methodology

3.1 Pre-Processing Stage

The audio signal is a digital signal obtained by the sampling of a continuous-time signal representation of human utterance at an agreed frequency, in this work, the sampling rate of 44,100KHz (f_s) was used in the speech recordings. The digital speech signal is then passed to a digital processing system which in our case is the VAD system.

The first activity performed on the audio signal in the VAD pipeline is usually pre-emphasis which intends to increase the signal to noise ratio as well as the intelligibility and fidelity of the signal. In this work, a different approach was taken which is the use of a digital Butterworth Band Pass IIR filter to allow only human speech with a frequency range 300Hz to 3400Hz. The upper frequency is set at 3000Hz to give an allowance of the transition frequency to take care of the 3400Hz. Butterworth filter is chosen because the bandpass is maximally smooth, with no ripples. The application of the filter is to make the VAD system robust and immune to noise which is usually the high-frequency component of the speech signal.

The second step in the pre-processing stage is the framing which is breaking or segmenting the whole speech signal into smaller and equal sizes. Each segment called frame has the same number of samples. The number of samples per frame is usually in the range of 512,1024 etc. because Discrete Fourier Transform (DFT) is more efficient when processing samples is in the power of 2 (2^N)).In this work, 512 samples are chosen per frame.

3.2 Features Extraction Stage

Two sets of features are considered in this VAD implementation, the time and frequency domain features which are the combination of low and mid-level features abstraction.

3.3. Extraction Of Time-Domain Features

For each of the frames, statistical information was obtained/ computed which are the standard deviation, skewness and kurtosis.

In addition to the above, three more features were calculated for each frame to make a total of six (6) time-domain features per frame and these are (1) zero-crossing, (2) energy envelope and (3) root-mean-square energy. Equation (1) through to equation (6) depicts how the time-domain features are generated.

Standard Deviation(std):

$$std_i = \sqrt{\frac{\sum_{n=1}^{N} s_i(n) - \bar{s_i}}{N}} \quad (1)$$

$where\ N = 512\ , s_i\ is\ the\ i^{th}\ frame\ and\ n = 1,2, \ldots 512$

Root-mean-square energy (RMS):

$$RMS_i = \sqrt{\frac{1}{N} \sum_{n=1}^{N} s_i\,(n)^2} \quad (2)$$

Zero Crossing Rate Feature (Zcr):

$$Z_{cr\_i} = \frac{1}{2}\sum_{n=i}^{N} | sgn(s_i(n) - sgn(s_i(n+1)| \quad (3)$$

$$sgn():$$
$$s_i(n) > 0 \rightarrow +1;$$
$$s_i(n) < 0 \rightarrow -1$$
$$s_i(n) = 0 \rightarrow 0$$

Skewness (Sk):

$$Sk_i = \frac{\sum_{n=1}^{N}(s_i(n)-\bar{s}_i)^3}{(N-1)*std_i^3} \quad (4)$$

Kurtosis (Kurt):

$$Kurt_i = \frac{\sum_{n=1}^{N}\frac{(s_i(n)-\bar{s}_i)}{N}}{std_i^4} \quad (5)$$

Normalized Amplitude Envelope (NAE)

$$NAE_i = \frac{max\,(s_i(n))}{max\,(s)} \quad (6)$$

It is normalized by dividing the maximum amplitude value in each frame by the maximum amplitude value in the entire speech signal to limit the effect of an outlier.

The time-domain features are computed for both the voice and unvoiced frames of the speech signals.

3.4. Extraction of Frequency Domain Features

The feature from the frequency domain that is considered for this VAD implementation is the Mel Frequency Cepstral Coefficient (MFCC). The first step in generating the above coefficients is the Discrete Fourier transformation of the product of each frame and a window function whose operation is expressed in equation (7). The window function of choice is Hann, represented by equation (8).

$$s(k) = \sum_{n=1}^{N} s_i(n)w(n)\, e^{\frac{-j2\pi nk}{N}} \quad (7)$$

where k=1, 2…. N

$$w(n) = 0.5[1 - \cos\left(\frac{2\pi n}{N-1}\right)] \quad (8)$$

where n=1, 2……. N

The Discrete Fourier Transform described above is simply a short-time Fourier transform (STFT). The power spectral is estimated for each of the speech frames by expression in equation (9) below:

$$P_i(k) = \frac{1}{N}| s_i\,(k)|^2 \quad (9)$$

The computed power spectral is multiplied with the Mel filter bank and the coefficients are added up. The described operation is carried out with thirteen (13) different Mel filter banks to generate thirteen (13) coefficients that represent the energies of the 13 filter banks. Log of the 13 energies is then taken and Discrete Cosine Transform of the output as computed which resulted into 13 Mel-frequency-Cepstral Coeffects (MFCC) which represents MFCC.

Thirteen (13) MFCCs are generated for each of the frames and combining these with six (6) time-domain features results in nineteen (19) features per frame, for both frames with and without voice activity.

3.5. Recursive Feature Elimination

The combination of the time and frequency domain representations resulted in 19 features representing each of the frames. Recursive feature elimination (RFE) was applied to select high priority features that contribute significantly to separating the two classes of data. With RFE, the number of features was reduced from 19 to 7.

3.6 SVM Training

The primary objective of this stage is to train a Support Vector Machine (SVM) with feature vectors of both the voice and silence frames for it to become a binary classifier.

SVM is a large margin, linear regression and classification tool in machine learning which maximizes prediction accuracy and at the same time prevents over-fitting of data. It searches for boundary hyperplane between the voice feature vectors x^+ and the silence feature vectors x^- that yield the highest separation between the two.

The hyperplane equations for classes are represented below where y_i is +1 for voice feature vectors and -1 for silence feature vectors

$$y_i = +1; wx^+_i + b \geq 1 \quad (10)$$
$$y_i = -1; wx^-_i + b \leq -1 \quad (11)$$

The W *and* b are the weight and constant of the equations. Generalizing equations (10) and (11) yields:

$$y_i(wx_i + b) \geq 1 \quad (12)$$

In like manner, extended manipulations of equations (10) and (11) result into:

$$x^+ - x^- = \frac{2}{w} \tag{13}$$

Further refinement of equation (12) and (13) result into constrain optimization problem below:

$$minimize\ f(w,b) = \frac{||w||^2}{2}$$

$$subject\ to: g(w,b) = y_i(wx_i + b) \geq 1$$

With the Lagrange multiplier, the constraint optimization problem can be resolved which is already implemented in software libraries such as sci-kit-learn for python. The training of the SVM model extracts the values of w and b which serves as the basis for the classification of a new data sample.

3.7 Performance Evaluation of The SVM Classifier

The speech features data set is split into the training and validation sets in a ratio of 70% to 30% respectively. The training set is used to train the SVM classifier while the validation set is used to obtain the performance of the VAD system.

The metric used for the evaluation are precision, accuracy, recall, F1

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1-score} = \frac{2*Precision*Recall}{Precision + Recall}$$

where *TP* = True Positive, *TN* = Time Negative, *FP* = False Positive, *FN* = False Negative.

4. Result and Analysis

The first stage of the VAD algorithm is the application of the digital bandpass filter with lower and upper cutoff frequencies of 300Hz and 3400Hz respectively to remove noise outside the human auditory frequency range. Figure 2 and figure 3 show the graph of a sample speech signal before the application of the digital filter in the time and frequency domain respectively while figures 4 and 5 show the two domains' representation of the same signal after the filtering process.

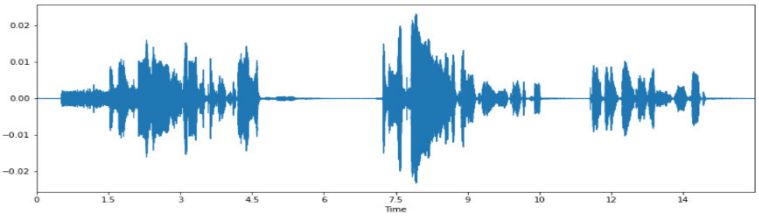

Figure 2 Sample signal time-domain representation before applying filter

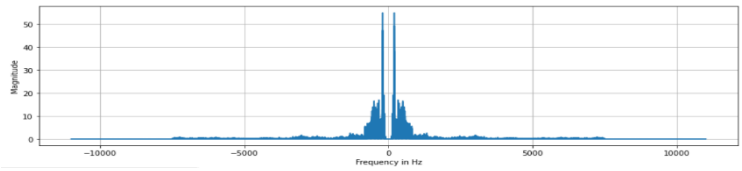

Figure 3 Sample signal frequency-domain representation before applying filter

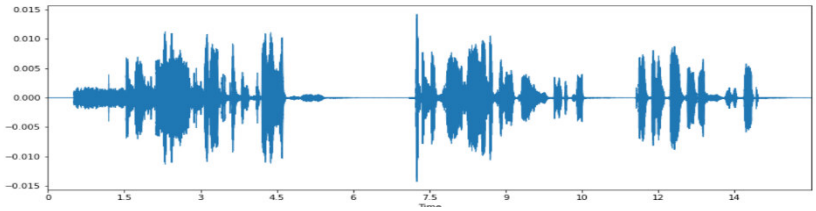

Figure 4 Sample signal time-domain representation after filtering operation

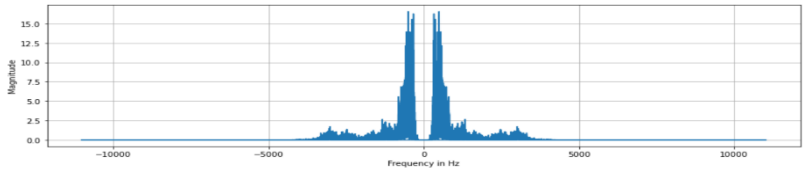

Figure 5 Sample signal frequency-domain representation after filtering operation

There is a significant change in the signal after passing it through the bandpass filter by comparing the two time-domain representations. The spectrum analysis also clearly revealed that frequency components of the speech above the higher cutoff frequency (3400Hz) of the filter were removed which are assumed to be the noise contained within the human auditory frequency range.

A total of 314 data points were extracted from 314 visually inspected frames which comprise 187 feature vectors from the frames without voice activity while 127 are for the voiced ones. Table 2 shows samples of time-domain features which are the standard deviation, skewness, kurtosis, zero-crossing, energy envelop and the root mean square for voiced and unvoiced classes. Table 1 contains sample feature data for the frequency domain from mfcc0 to mfcc12 for the two classes.

Table 1 Mel Frequency Cepstral Coefficients of voiced and unvoiced sample frames

Class	mfcc0	mfcc1	mfcc2	mfcc3	mfcc4	mfcc5	mfcc6	mfcc7	mfcc8	mfcc9	mfcc10	mfcc11	mfcc12
1	-667.547	-654.498	-524.643	-626.501	-599.677	-625.068	-491.274	-601.924	-595.792	-588.206	-616.326	-588.668	-611.483
1	-627.435	-602.438	-635.504	-601.593	-598.48	-606.385	-630.874	-544.127	-523.223	-559.475	-561.048	-522.743	-545.407
1	109.6918	140.343	98.79312	138.513	120.3517	132.581	88.31181	140.4578	126.2435	114.4579	147.8204	116.1384	134.5401
1	135.894	120.4886	150.518	124.9583	116.5196	127.8496	138.8621	101.1753	96.45069	117.0992	113.6568	100.9746	108.0538
1	5.916163	14.6907	-77.8704	8.544289	6.88729	20.75451	-73.1382	1.788214	0.016243	6.495839	-2.12252	0.404058	16.76525
1	9.401464	-1.71211	6.411694	12.86418	5.360177	14.39612	15.85955	8.774739	5.36725	12.41436	11.18975	3.076547	11.1984
1	57.86197	29.79846	45.54201	29.66286	28.14723	36.2601	34.42569	19.7766	12.51178	29.94729	28.69293	25.94044	39.80795
1	39.62717	49.64317	38.81763	42.40513	40.71951	39.86903	51.70235	37.86803	31.41098	35.7858	34.81285	27.9739	38.60461
1	11.78587	22.33723	19.11372	-15.6676	1.3885	-30.061	34.09326	-4.4374	-4.02866	6.219887	-7.34015	8.356093	-10.2445
0	-855.77	-953.988	-973.159	-991.461	-985.88	-992.709	-995.179	-990.464	-1005.08	-995.435	-985.707	-861.373	-884.414
0	-910.603	-938.893	-965.684	-878.06	-924.322	-921.606	-844.371	-884.957	-918.743	-937.151	-940.844	-985.482	-985.974
0	-983.604	-978.857	-1002.42	-982.789	-982.05	-969.848	-996.353	-1016.99	-1014.18	-1026.74	-1034	-1025.14	-1042.45
0	-1045.55	-1057.84	-1058.87	-1041	-1038.31	-1020.21	-1030.5	-1022.82	-989.362	-986.836	-1015.75	-1040.55	-1057.57
0	-1055.25	-1058.73	-1059.84	-1046.87	-1064.93	-1037.57	-1025.67	-1025.48	-1008.99	-1013.31	-988.962	-980.615	-971.766
0	-981.762	-980.215	-979.22	-974.739	-967.125	-968.23	-981.616	-980.807	-985.399	-988.451	-989.428	-980.032	-988.935
0	-984.958	-990.756	-985.537	-975.938	-987.826	-984.155	-981.836	-988.827	-921.759	80.35852	37.28705	28.21441	22.67187
0	22.81503	12.92874	16.41576	17.83514	21.02428	21.46605	-1.48347	-37.6423	-41.7277	-10.7674	-3.16223	-15.2786	-12.0428
0	-24.8718	-7.41521	-49.2044	-40.2998	-19.0842	-29.7032	12.51252	11.00104	21.54072	23.18946	28.06202	13.21343	28.63843

Table 2 Time Domain features of both voiced and unvoiced sample frames

std	skewness	Kurtosis	Zero-crossing	Energy-envelop	RMS	Class
0.02978	-0.05051668	-1.148985	0.015625	0.050915498	0.02978125	1
0.031358	-0.05725393	-1.24215	0.013671875	0.050719671	0.03136364	1
0.034659	-0.1445414	-1.132414	0.015625	0.05303698	0.03469888	1
0.03125	-0.15248829	-1.075058	0.015625	0.053722382	0.03151448	1
0.037792	-0.11929356	-1.244198	0.013671875	0.05617024	0.03781194	1
0.034931	-0.14918895	-1.039718	0.015625	0.053330723	0.03529259	1
0.032682	-0.34921852	-0.730927	0.013671875	0.068768561	0.03323885	1
0.038227	-0.23290113	-1.20428	0.013671875	0.059270862	0.03822747	1
0.03741	0.176758811	-1.267081	0.015625	0.064492963	0.03773385	1
0.038853	-0.18968654	-1.036787	0.015625	0.065765858	0.03901434	1
0.000246	-0.02819782	0.6858659	0.115234375	0.000783315	0.0002461	0
3.56E-05	0.09513294	-0.115889	0.25	9.79E-05	3.63E-05	0
4.47E-05	-0.50149185	0.4710221	0.227005871	9.79E-05	4.52E-05	0
4.07E-05	-0.33187223	0.1034547	0.19140625	9.79E-05	4.07E-05	0
3.82E-05	-0.48541221	-0.134864	0.154296875	9.79E-05	3.92E-05	0
0.000502	-0.05862364	0.9240443	0.359375	0.002154117	0.00050158	0
0.00022	-0.06153521	-0.061567	0.384765625	0.000652763	0.00021972	0
0.000221	0.108978756	0.184226	0.390625	0.000815954	0.00022116	0
0.00015	-0.05229124	0.1031211	0.376953125	0.000489572	0.00015048	0

Recursive Feature Elimination was applied as means of feature selection which reduces the number of features to seven (7). The seven features are: - Standard deviation, Energy Envelop, Root Means Square, MFCC1, MFCC2, MFCC7 and MFCC8. Figures 6 and 7 below are plots of standard deviation vs mfcc2 and

energy envelope vs mfcc8 respectively of voiced and unvoiced frames.

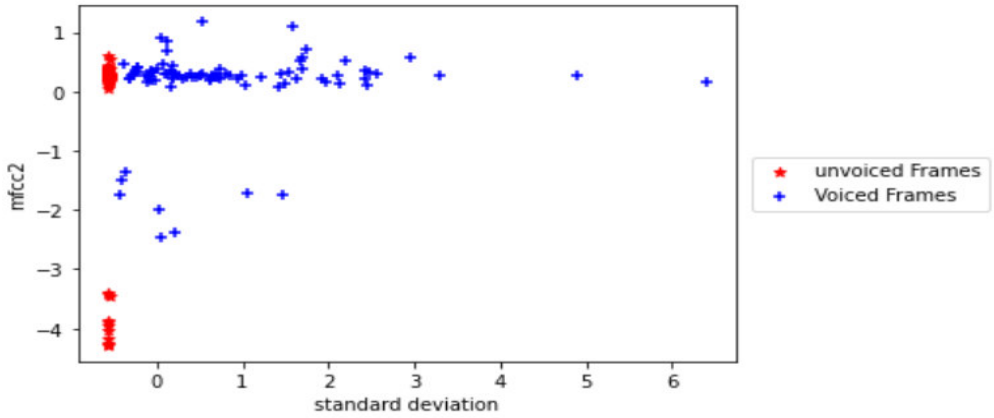

Figure 6 Standard deviation vs Mfcc2 feature Separation Space

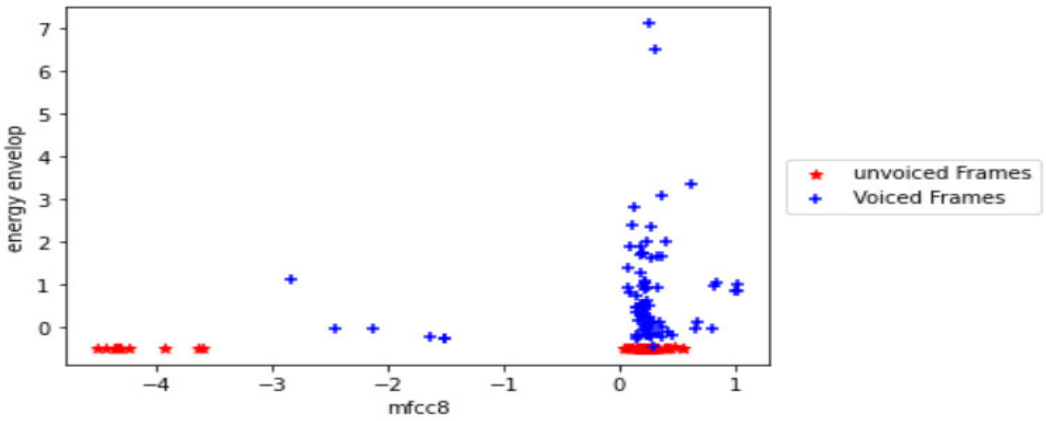

Figure 7 Mfcc8 vs Energy envelop feature Separation Space

The SVM classifier was trained with 70% of the dataset, and the remaining 30% (95 records) was used for evaluation. Table 5 is the confusion matrix of the output of the validation exercise captured for the best outcome and table 6 captures various performance metrics for the same instance.

Table 3 Confusion Matrix of the validation outcome

		PREDICTION	
		FALSE	TRUE
ACTUAL	FALSE	56	0
	TRUE	0	39

From table 3, the True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN) are 39,0,0 and 59 respectively.

Table 4 Performance Metrics

Class	**Precision**	**Recall**	**Support**
0	100%	100%	56
1	100%	100%	39
Weighted Avg Accuracy	100%	100%	95

Regarding Table 4, the classification report shows the weighted average Precision, Recall, and F1-score of the SVM based VAD discriminator as 100%,100% and 100% respectively while the accuracy is reported as 100%.

5. Discussion

The VAD algorithm of this work makes use of the features fusion strategy by combining both the time and frequency domain features of the speech signal.

The set objectives were achieved as frequency components outside the human auditory range were removed using digital filters, and features were extracted via visual inspection of the speech signal to segregate between sections with and without voice activity. Recursive feature elimination was used to reduce the number of features from nineteen (19) to seven (7) and a Support Vector Machine using RBF kernel was trained with the reduced features. Finally, the VAD built on the SVM was validated using test data set.

6. Conclusion

The performance of this VAD yielded a state-of-art performance by recording accuracy of 100%, recall of 100%, precision of 100% and F1 score of 100% despite the simplicity of the design and implementation.

The computational demand of this voice activity detection system is very low as the architecture is simple, an SVM classifier. Likewise, the training time is insignificant as very few features were considered which is the benefit derived from the introduction of feature selection. Overall, our implementation of VAD recorded performance that is at par with similar systems that use complex algorithms and architectures of Deep Neural Networks which naturally demand high computation power.

Based on the highlighted advantages, this VAD algorithm will be suitable as a front-end for speech processing applications meant to run devices with low computational resources.

References

Bäckström, T. (2017). Voice Activity Detection (pp. 185–203). https://doi.org/10.1007/978-3-319-50204-5_13

Chelloug, C. E., & Farrouki, A. (2019). Robust Voice Activity Detection Against Non Homogeneous Noisy Environments. 2018 International Conference on Signal, Image, Vision and Their Applications, SIVA 2018. https://doi.org/10.1109/SIVA.2018.8661045

Dey, J., Hossain, M. S. Bin, & Haque, M. A. (2019). An ensemble SVM-based approach for voice activity detection. ICECE 2018 - 10th International Conference on Electrical and Computer Engineering. https://doi.org/10.1109/ICECE.2018.8636745

Drugman, T., Stylianou, Y., Kida, Y., & Akamine, M. (2016). Voice Activity Detection: Merging Source and Filter-based Information. IEEE Signal Processing Letters, 23(2). https://doi.org/10.1109/LSP.2015.2495219

Dwijayanti, S., Yamamori, K., & Miyoshi, M. (2018). Enhancement of speech dynamics for voice activity detection using DNN. Eurasip Journal on Audio, Speech, and Music Processing, 2018(1). https://doi.org/10.1186/s13636-018-0135-7

Elton, R. J., Vasuki, P., & Mohanalin, J. (2016). Voice activity detection using fuzzy entropy and support vector machine. Entropy, 18(8). https://doi.org/10.3390/e18080298

Krishnakumar, H., & Williamson, D. S. (2019). A comparison of boosted deep neural networks for voice activity detection. GlobalSIP 2019 - 7th IEEE Global Conference on Signal and Information Processing, Proceedings. https://doi.org/10.1109/GlobalSIP45357.2019.8969258

Lavechin, M., Gill, M. P., Bousbib, R., Bredin, H., & Garcia-Perera, L. P. (2020). End-to-end domain-adversarial voice activity detection. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2020-Octob. https://doi.org/10.21437/Interspeech.2020-2285

Lee, K., & Ellis, D. P. W. (2006). Voice activity detection in personal audio recordings using autocorrelogram compensation. INTERSPEECH 2006 and 9th International Conference on Spoken Language Processing, INTERSPEECH 2006 - ICSLP, 4. https://doi.org/10.21437/interspeech.2006-540

Makowski, R., & Hossa, R. (2020). Voice activity detection with quasi-quadrature filters and GMM decomposition for speech and noise. Applied Acoustics, 166. https://doi.org/10.1016/j.apacoust.2020.107344

Mohammed, S. N., & Hassan, A. K. (2020). Automatic voice activity detection using fuzzy-neuro classifier. Journal of Engineering Science and Technology, 15(5).

Pang, J. (2017). Spectrum energy based voice activity detection. 2017 IEEE 7th Annual Computing and Communication Workshop and Conference, CCWC 2017. https://doi.org/10.1109/CCWC.2017.7868454

Park, J. S., Yoon, J. S., Seo, Y. H., & Jang, G. J. (2017). Spectral energy based voice activity detection for real-time voice interface. Journal of Theoretical and Applied Information Technology, 95(17).

Sehgal, A., & Kehtarnavaz, N. (2018). A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection. IEEE Access, 6. https://doi.org/10.1109/ACCESS.2018.2800728

Tan, Z. H., Sarkar, A. kr, & Dehak, N. (2020). rVAD: An unsupervised segment-based robust voice activity detection method. Computer Speech and Language, 59. https://doi.org/10.1016/j.csl.2019.06.005

Yang, X.-K., He, L., Qu, D., & Zhang, W.-Q. (2016). Voice activity detection algorithm based on long-term pitch information. https://doi.org/10.1186/s13636-016-0092-y

Zazo, R., Sainath, T. N., Simko, G., & Parada, C. (2016). Feature learning with raw-waveform CLDNNs for Voice Activity Detection. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 08-12-September-2016. https://doi.org/10.21437/Interspeech.2016-268