International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-14 Issue-4, March 2025

Voice Activity Detection Using Weighted K-Means Thresholding Algorithm

Alimi Sheriff, Yussuff I. O. Abayomi

Abstract*: *Voice activity detection (VAD) separates speech segments from silent segments of an audio signal, and it is valuable for many speech-processing applications because it assists in improving performance and system efficiency; such applications include speech recognition and speaker verification. In this study, K-means, a clustering algorithm, was extended to a thresholding algorithm termed K-means weighted thresholding and was utilised for discriminating voiced/speech segments from silent segments in audio or speech signals. The voice signal was fragmented into frames of 2048 samples, and the spectral power of the frames served as input for computing the threshold value by the extended k-means algorithm; hence, any frame whose spectral power is greater than or equal to the threshold value is considered to part of the voice segments; otherwise, it is tagged as a silent frame. The implemented voice activity detection system achieved outstanding performances with an actual acceptance rate (sensitivity), false acceptance rate, actual rejection rate (specificity), false rejection rate (miss rate), and a classification accuracy of 100%, 0.025%, 100%, 0%, and 99.97%, respectively.

Keywords*: *K-Means, Thresholding Algorithm, Voice Activity Detection

I. INTRODUCTION

Voice activity detection separates speech segments from silent segments of an audio signal. It is the front-end module of several speech processing systems, such as speaker and voice recognition, used to enhance system accuracy and performance. Guaranteeing that only the audio signal's voice-active portions are encoded and sent encourages its employment in communication systems to ensure practical usage of transmission bandwidth [1]. Other applications of voice activity detection include wake-up systems. [2].

About 40% of audio or voice signals are made up of speech regions, which contain voice activity, while the other 60% are silent parts [3]. Voice activity detection enhances system performance and efficiency by enabling back-end voice or speech processing applications to focus on the most critical audio components. This study aims to design and implement a reliable voice An Activity detection system that will be used as the front end of a speech-processing application for the detection of schizophrenia and the estimation of the severity of its symptoms. Only a few studies in the literature considered voice activity detection in voice-based schizophrenia diagnostic systems: these are [4], which uses a manual method based on visual examination, and [5], which uses an automated method.

Manuscript received on 15 January 2025 | First Revised Manuscript received on 18 January 2025 | Second Revised Manuscript received on 17 February 2025 | Manuscript Accepted on 15 March 2025 | Manuscript published on 30 March 2025.

*Correspondence Author(s)

Alimi Sheriff*, Department of Computer Science, Babcock University, Ilishan Remo (Ogun State), Nigeria. Email ID: alimi0356@pg.babcock.edu.ng, ORCID ID: 0009-0002-1954-1598

Yussuff I. O. Abayomi, Associate Professor, Department of Electronic and Computer Engineering, Lagos State University, Epe (Lagos), Nigeria. Email ID: abayomi.yussuff@lasu.edu.ng, ORCID ID: 0000-0003-3829-9944

II. REVIEW OF RELATED STUDIES

Many of the reviewed studies on voice activity detection for segregating between voice and silence segments of the speech signal adopted a thresholding approach where specific speech attributes are computed, and any voice segments whose value is greater than this threshold are labelled as "voice frames," otherwise, they are considered "silent frames" [6]. Other research has employed machine learning algorithms such as Support Vector Machine (SVM) [1] Artificial Neural Network (ANN) [7], and Adversarial Neural Network [8] To discriminate between voiced and unvoiced (silent frames). The literature review is structured into two strategies: thresholding and machine learning.

A. Thresholding Strategy for Voice Activity Detection

In the thresholding strategy, specific speech attributes are computed, and any voice segments whose value is greater than or equal to this threshold are labelled as "voice frames," otherwise, they are considered "silent frames" [9]. Methods of studies that adopted this strategy are discussed below in detail.

In the proposed VAD design by [10] Based on adaptive thresholding, the first certain number of frames are considered silent frames, which are used to compute threshold values for discriminating between voice and voice-free frames. The threshold value is updated via adaptive learning. The algorithm achieved a correct classification rate of 84.6%, with a false acceptance rate of 9.7% and an actual rejection rate of 5.7%. The energy level and pitch are used to determine if a segment (a group of consecutive frames) should be considered a voice segment or not in the study conducted by Tan et al [11]. The second-stage decision on each frame is based on the posterior of the signal-to-noise ratio (SNR) weighted energy of the frame. The proposed VAD algorithm yields a significantly lower average FER and superior performance with a substantial margin under all SNR levels compared to some existing VAD methods (both supervised and unsupervised).

The speech signal is separated into two frequency bands, the Low-Frequency Band (LFB) and the High-Frequency Band (HFB). The total spectrum energy in the LFB and HFB For each frame, the EnL and EnH values are calculated,

respectively. A five-point moving average filter was applied to EnH, which was considered the noise signal. The mean EnH is subtracted from the 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. The results show that the adopted strategy is robust at SNR values of 1 dB, 2 dB, 10 dB, and 25 dB [12].

In the thresholding voice segment detection system of Yang et al. [9] The long-term pitch divergence (LTPD) was computed for each frame using filter banks. If the LTPD of a frame is greater than the threshold calculated from the SNR, it is labelled as a speech frame. The work utilised computed spectral energy per frame for the frequency range associated with humans, made possible by the short-time Fourier transform (STFT). The spectral energy threshold was the basis for classifying frames as either voice frames or silent frames.

In the case of [6], the envelopes of narrow frequency bands were computed using a complex filter; the average of the weighted squares of envelopes, which is calculated over a set of frequencies, and the variance of the square of weighted envelopes are used in deciding if a fragment has voice activity or not based on the threshold determined by the Gaussian mixture model. In [2]Hardware implementation of voice activity detection was realised based on a thresholding algorithm. The number of token pulses was generated by a level-crossing analogue-to-digital converter (LC-ADC) for each time window; if this number exceeds a certain threshold, it is considered to contain voice activity. The study achieved hit rates of 91.02% and 82.64% for speech and non-speech, respectively.

The unsupervised voice detection algorithm of [13] It is based on the thresholding of fractal dimensions derived using the Katz algorithm, which was used for detecting voiced frames and yields an average classification accuracy of 90.45% with audio signals across three types of noise (white noise, car noise, and babble noise). According to [14] The zero-crossing threshold and mean square energy threshold derived from all the frames were combined to form a global threshold value used for segregating voiced and unvoiced frames concerning the signal coverage for each frame.

B. Machine Learning Strategy for Voice Activity Detection

The research studies that adopted this strategy employed machine learning algorithms such as SVM [15], and Adversarial Neural Network [8] Trained with features extracted manually or automatically to discriminate between voiced and unvoiced (silent frames). The review of this strategy is divided into two sections: Automatic Feature Extraction and Manual Feature Extraction.

i.Automatic Extracted Features-Based Strategy

An Adversarial Neural Net that consists of a SincNet for feature extraction from the raw waveform (speech recording), LSTMs with feed-forward layers for voice activity detection, and an LSTM with temporal pooling and feed-forward for domain classification was implemented by [8] As the VAD system, which reported a detection error rate of 9.3%, a false alarm rate of 5.7%, and a missed detection rate of 4.2%.

The raw waveform form (both noisy and clean recordings) is applied to Convolutional Long Short-Term Memory Deep Neural Networks (CLDNN) for training, which also automatically extracts the necessary features for the classification exercise. The CLDNN classifier recorded its best performance in terms of false alarms, at 4.1%, and fixed false rejection at 2% on noisy data [16].

Sehgal and Kehtarnavaz [17] Converted each voice frame to log-mel-filter bank energy images, which were used to train a Convolutional Neural Network (CNN) directly; this CNN-based VAD achieved a speech hit rate (SHR) and noise hit rate (NHR) of 91.3% and 99% for real-time application. The voice activity detection algorithm implemented by [18] An ensemble of five SVMs, utilising MFCC features from voice frames, achieved a prediction accuracy of 87.4%.

ii.Manual Extracted Features-Based Strategy

In this subdivision, the features are handcrafted and then used to train traditional machine learning algorithms and deep learning models. This section highlights various methods in this respect.

In the VAD system design of [19], source-related and filter-based features were combined and used in training an ANN, and the two features were separately used to train two different ANNs. The final decision was based on the geometric mean of the predictions of the two sets of classifiers. The feature fusion and decision fusion-based ensemble classifiers yield F1-scores of 93.6% and 94.8%, respectively. In their voice segment detection algorithm, Elton et al [15]. Extracted fuzzy entropy features from voice frames and, with an SVM classifier, were able to discriminate between voiced and unvoiced frames from low-noise speech with an accuracy of 93%.

From voice frames, time and frequency domain features were extracted, including zero-crossing, standard deviation, normalised envelope, kurtosis, skewness, and 13 MFCC coefficients. Recursive feature elimination reduced the number of features to seven (7), which was used to train an SVM to be able to distinguish between voiced and unvoiced frames. The proposed VAD system reported an accuracy of 100%, a recall of 100%, a precision of 100%, and an F1 score of 100% [1].

C. Comparison of Thresholding and Machine Learning-Based Strategies

Threshold techniques make assumptions about how the threshold value is computed [10] And this value is sometimes updated [2] Via adaptive mechanisms. The performance of the VAD system might be unpredictable if the assumption is inappropriate for a particular voice signal. The advantage of this strategy is that the intrinsic property of the speech in question is taken into consideration to arrive at the assumed threshold value.

For the machine learning strategy, a universal model is built based on samples of speech signals that have been seen, and it serves as a basis for making decisions for unseen samples. The intrinsic properties of each unseen sample are not considered, and the model's performance degrades significantly if the sample's speech intrinsic behaviour is an outlier.

This study attempts to combine the two strategies, using a machine learning algorithm to compute the threshold

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

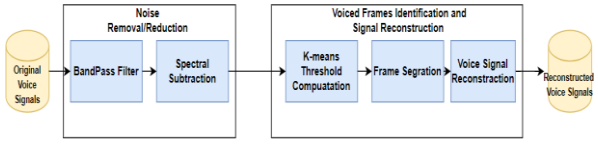
value, and the decision on whether a frame is voiced or unvoiced is then based on whether its value is less than or greater than the calculated threshold value.

III. METHODOLOGY

The design strategy adopted a combination of machine learning and thresholding techniques to realize a voice activity detection engine, and the methodology comprises two primary stages, which are:

- Noise removal/reduction and
- Voice frames identification and signal reconstruction

as represented in Figure 1. The details of each stage are discussed in subsequent sections.



[Fig.1: Methodology]

This study is the pre-processing phase of a larger research project that focuses on detecting and estimating the severity of symptoms of schizophrenia from the recorded speech of the chosen, consenting subjects. The database, as mentioned earlier, provided the voice files used in the development of this VAD system.

A. Noise Removal/Reduction Strategy

The noise removal strategy combines signal frequency band limiting and spectral noise subtraction. The Bandpass filter and spectral noise reduction techniques are elaborated upon in sections A1 and A2, respectively

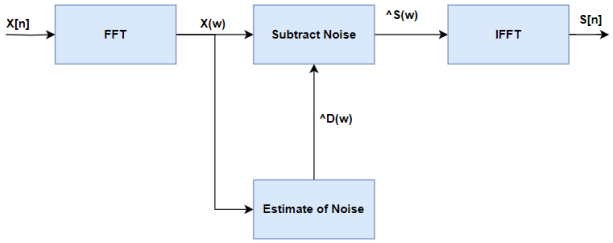
i.Bandpass Filter

Band limiting is one of the preprocessing techniques in signal processing for removing signals outside the range of interest, as such signals are considered noise and interference (especially high-frequency signals). The frequency range of human speech is between 80 Hz and 8 kHz [20] The introduction of a digital bandpass filter with a bandpass of 80 Hz to 8000 kHz will ensure that components of the signal outside this range are removed or significantly attenuated. Butterworth filter approximation technique was used in designing digital filters from analogue filter specifications because of their maximally flat passband and stop (no ripple) characteristics [21].

The Butterworth function, available in the SciPy library of Python, is utilised to implement this research bandpass filter, with a low-pass frequency set at 80 Hz, an upper-pass frequency set at 8000 Hz, and a filter order set to 5.

ii.Spectral Subtraction

Many of the speech signals contain background noise, which corrupts the signal, and the noise interferes with the understanding and intelligibility of the message [22]. To reduce the noise content in the speech signal, a noise reduction algorithm is needed, and one of the techniques is spectral subtraction. Spectral subtraction involves estimating the noise power spectrum during periods of noisy or no-speech activity, which is then subtracted from the power spectra of all frames to obtain the clean speech power spectrum. The spectral noise subtraction process is represented in Figure 2.



[Fig.2: Noise Reduction by Spectral Subtraction]

Let the noisy speech signal be $x(n)$Clean speech signal and noise signal are represented by $s(n)$ and $d(n)$ Respectively. The relationship between these three variables is denoted by equation (1).

$$x(n) = s(n) + d(n) \ldots \quad (1)$$

With Fast Fourier Transform (FFT),

$$x(n) \Rightarrow X(w), s(n) \Rightarrow S(w) \text{ and } d(n) \Rightarrow D(w).$$

The noise estimate $D(w)$ In the frequency domain, it is computed as the average of the summation of the absolute value of the spectral level of each frame $X_i(\omega)$ as expressed in equation (2)

$$\hat{D}(w) = \frac{1}{N}\sum_i^N |S_i(w)|_{average} \ldots \quad (2)$$

Then, an estimate of the clean signal in the frequency domain was computed using equation (3)

$$\hat{S}(w) = (|X(w)| - |\hat{D}(w)|e^{j\theta_x(w)} \ldots \quad (3)$$

$\theta_x(w)$ Is the phase of the noisy speech signal, which is computed with equation (4)

$$\theta_x(w) = \tan^{-1}\frac{\mathrm{Im}(X(w))}{\mathrm{Re}(X(w))} \quad \ldots \quad (4)$$

Equation (5) is then extended to equation (6) to take care of scenarios where $D(w)$ is greater than $S(w)$.

$$|\hat{S}(w)| = \begin{Bmatrix} X(w) - D(w) & if\ X(w) \geq D(w) \\ 0 & if\ X(w) < D(w) \end{Bmatrix} \quad \ldots \quad (5)$$

Finally, to obtain a clean speech signal $s(n)$Inverse Fast Fourier Transform (IFFT) of $S(w)$ Was computed.

B. Voiced Frames Identification and Signal

This section consists of three sub-sections

- K-Means Thresholding
- Voiced Frames Identification
- Reconstruction of the Signal without Silent Frames

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

i.Voiced Frames Identification and Signal

The voice signal is segmented into frames consisting of 2048 samples each and a Short Time Fourier transform (STFT) with Hann window computed for each of the frames using equations (6) and (7) respectively.

$$S_i(k) = \sum_{n=1}^{N} S_i(n)W(n)e^{\frac{-j2\pi nk}{N}} \quad \ldots \quad (6)$$

$$W(n) = 0.5(1 - \cos(\frac{2\pi}{N})) \quad \ldots \quad (7)$$

$n = 1, 2, 3, \ldots . N$, where N=2048, the number of samples per frame, i Is the position of the frame in the signal?

The minimum spectral power for each frame is defined as the spectral power of that frame and is calculated using the formula in equation (8).

$$P_i(k) = \text{min}imum|S_i(k)|^2 \quad \ldots \quad (8)$$

Next, the K-means clustering algorithm was used to segment the frames into two clusters, with the spectral power of each frame, computed earlier, serving as the input feature. The centroids of the two clusters are $v1$ and $v2$, and the respective number of data points for each of the clusters is $N1$ and $N2$, respectively.

The essence of the K-means introduction is to utilise its output to compute a threshold value that represents the intrinsic nature of the signal. The formula for computing the threshold value is expressed in equation (9), which is a weighted average.

Algorithm: K-Means (Weighted) Thresholding Algorithm

(1) Randomly select 2 points or centroids υ_1 and υ_2 from the input dataset (list of voice frames, minimum spectral power)
(2) Assign each frame of the frame spectra's mean to the closest centroid, to form 2 clusters

(3) Compute the mean of each cluster to form new centroids, υ_1 and υ_2

While the number of iterations is less than the predefined times, repeat steps 2 and 3. End
(4) Obtain the number of data points, cap N sub 1and cap N sub 2, assign troidsupsilon sub 1ne and troidsupsilon sub 2, $d\ \upsilon_2$ respectively, and compute the weighted average as expressed in equation (9).

$$\mathrm{T_{value}} = \frac{\mathrm{N_1 v_1 + N_2 v_2}}{\mathrm{N_1 + N_2}} \quad \ldots \quad (9)$$

ii.Frame Segregation

The computed threshold value, T_{value} , was used to segregate between voiced frames and silence frames. Frames with spectral power greater than or equal to the computed threshold value are considered voiced silence; they are regarded as silence frames, as depicted in equation (10).

$$Decision = \begin{Bmatrix} Voice\ frame & if P(w) \geq T \\ Silence\ frame & if\ P(w) < T \end{Bmatrix} \quad \ldots \quad (10)$$

iii.Reconstructed Voice Signal

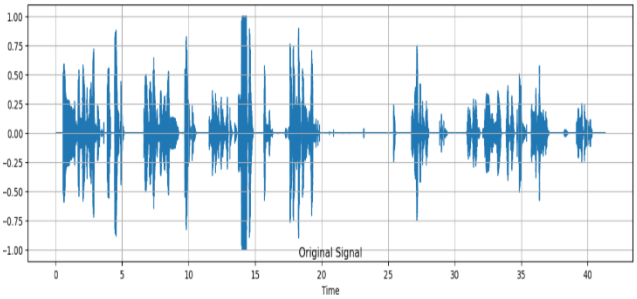
The voice signal is reconstructed by iterating through the frames in sequential order per their respective position in the original signal, and any frame tagged as a silence frame is discarded. The newly reconstructed voice signal is expected to be free of silence frames, with noise signals removed or reduced. It is likely to be of shorter duration compared to the original voice signal.

IV. RESULTS

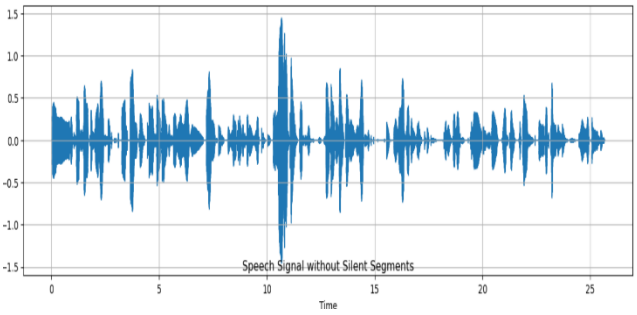
The results of the visual examination of the newly created voice activity detection system, which is based on the k-means thresholding algorithm, applied to a collection of speech signals, are presented in this section. Standard performance metrics, such as classification accuracy, were also used to evaluate the system's performance.

A. Visual Examination and Analysis

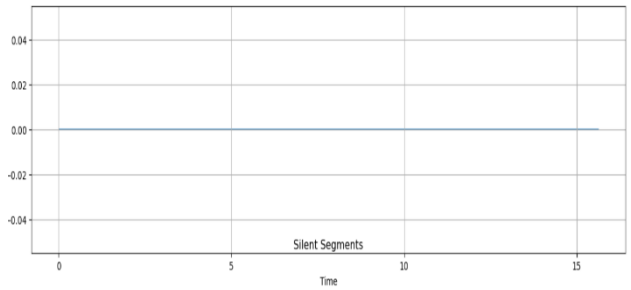
Figure 5(a) is the original voice signal before the voice activity detection system was applied to remove the silent frames, and the duration is 41.32 seconds. Figures 5(b) and 5(c) represent the reconstructed speech with the silent frame and the extracted silent segment of the original speech, respectively.



[Fig.3(a): Original Speech Signal]



[Fig.3(b): Voiced Segment]



[Fig.3(c): Silent Segment]

The voiced segment's duration is 25.68 seconds, while that of the silent segment is 15.62 seconds. By visual inspection, the amplitude of the samples in the silent segment is all zero, which is an indication of the absence of a voiced frame in this segment.

Retrieval Number: 100.1/ijitee.D105114040325
DOI: 10.35940/ijitee.D1051.14040325
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

B. Performance Metrics

With spectral subtraction, noise within the voice signal is removed, and what is left are the silent and voiced frames. The amplitude of samples in silent frames is expected to be zero, while that of voiced samples is greater than zero. Analysis at this stage is performed based on frame classification, and the number of frames in any particular segment is computed by dividing the total number of samples by the number of samples per frame (2048 samples per frame in this study).

Beyond visual assessment, specific performance metrics are also available to evaluate the performance of the k-means thresholding-based voice activity detection algorithm, providing a basis for comparing the results of this study with those of related studies. The metrics are True Acceptance Rate (TAR), False Acceptance Rate (FAR), True Rejection Rate (TRR), False Rejection Rate (FRR), and Accuracy (ACC). Formulas for calculating these metrics are expressed with equations (11) to (15).

$$TAR(sensitivity) = \frac{TP}{TP+FN} \quad \ldots \quad (11)$$

$$FAR(Fallout) = \frac{FP}{FP+TN} \quad \ldots \quad (12)$$

$$TRR(Specificity) = \frac{TN}{TN+FP} \quad \ldots \quad (13)$$

$$FRR(Miss\ Rate) = \frac{FN}{FN+TP} \quad \ldots \quad (14)$$

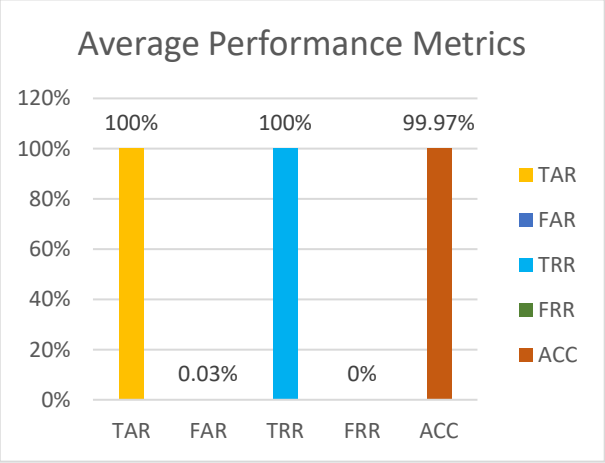
$$ACC = \frac{TP+TN}{FN+FP+TN+TP} \quad \ldots \quad (15)$$

TP: True Positives (correctly accepted), TN: True Negatives (correctly rejected), FP: False Positives (incorrectly accepted), and FN: False Negatives (incorrectly dismissed).

Eight voice files were used to validate the VAD system, and the performance metrics for each file are presented in Table 1. The TRR and FRR for all the files are 100%, while the FAR ranges from 0% to 24%. The FRR is 0% for all the voice files, while classification accuracy ranges from 95.15% to 99.92%.

Table 1: Performance Metrics

File No	True Positive	False Positive	True Negative	False Negative	Actual Acceptance Rate (TAR)	False Acceptance Rate (FAR)	Actual Rejection Rate (TRR)	False Rejection Rate (FRR)	Classification Accuracy (ACC)
1	277	18	337	0	100%	5.15%	100%	0%	97.108%
2	276	44	1,024,000	0	100%	0.00%	100%	0%	99.996%
3	768	86	921	0	100%	8.55%	100%	0%	95.152%
4	871	1	11	0	100%	6.39%	100%	0%	99.915%
5	649	1	11	0	100%	4.35%	100%	0%	99.924%
6	560	33	214	0	100.%	13.27%	100%	0%	95.938%
8	1,480	77	238	0	100%	24.42%	100%	0%	95.720%



[Fig.4: Overall Performance of the Eight (8) Voice Files]

Figure 4 shows the VAD system's overall performance for the eight (8) voice files, the True Acceptance Rate (Sensitivity), False Acceptance Rate, True Rejection Rate (Specificity), False Rejection Rate (Miss Rate), and classification Accuracy of 100%, 0.025%,100%, 0%, and 99.97%, respectively, based on aggregate frames summation.

V. DISCUSSION

The VAD system designed and implemented in this study combined two major strategies utilised in the review studies: thresholding and machine learning. Vital information generated by the K-means clustering algorithm was utilised to compute the intrinsic threshold value of the speech signal, with the minimum spectral power of each frame serving as the input feature. The centroid values and the number of data points for each cluster were used to compute a weighted average, which serves as the threshold value. Any frame whose minimum power spectral value was greater than the threshold value was considered a voiced frame, forming an integral part of the expressed segment; otherwise, they were tagged as silent frames and part of the silent segment.

Before computing the threshold value, the Butterworth bandpass filter allowed only components of the signal with frequencies between 80 Hz and 8000 Hz to pass through, and spectral subtraction was used to remove noise within this frequency range.

The performance of the developed VAD system is outstanding, with an actual

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

acceptance rate (sensitivity), false acceptance rate, actual rejection rate (specificity), false rejection rate (miss rate), and classification accuracy of 100%, 0.025%, 100%, 0%, and 99.97%, respectively. Table 2 is used to compare the outcomes of this study with closely related work for benchmarking purposes.

Table 2: Comparative Table of Closely Related Studies

SN	Authors	Strategy and Technique	Results
1	[10]	Adaptive Thresholding	TRR=5.7%, FAR=9.7%, Acc=84.6%
2	[8]	Machine Learning: Adversarial Neural Net	False Alarm Rate=5.7%, Missed Detection Rate=4.2%, Detection Error Rate=9.3%
3	[16]	Machine Learning: CLDNN	False Alarm Rate=4.1%, FRR=2%
4	[17]	Machine Learning: CNN	SHR=91.3%, NHR=99%
5	[18]	Machine Learning: Assembling of SVMs	Acc=87.4%
6	[19]	Machine Learning: NN	F1-Score 93.6%, 94.8%
7	[15]	Machine Learning: SVM	Acc=93%
8	[2]	Thresholding	91.02% and 82.64%
9	[1]	Machine Learning: SVM	Acc=100%, Recall=100%, Precision=100% and F1-Score=100%
10	Current study	Machine Learning and thresholding: K-means weighted thresholding	TRR=100%, FRR=0%, FAR=0.025%, TAR=100%, Acc=99.97%

According to Table 2, the K-means thresholding-based voice activity detection system exhibits the best performance among the reported studies that employ the thresholding strategy, ranking second overall, behind the study that reported a classification accuracy of 100%.

VI. CONCLUSION

This study's findings demonstrate that the developed voice activity detection algorithm, which uses K-means to compute a weighted threshold value, is very effective at differentiating between voiced and silent (unvoiced) segments of speech signals. This is evidenced by its 99.97% classification accuracy, which is almost perfect, and its excellent actual rejection and accurate acceptance rates. At the core of this developed voice activity detection system is the extended K-means clustering method, which serves as a thresholding algorithm and can also be applied to other clustering or segregation tasks.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/Competing Interests: Based on my understanding, this article does not have any** conflicts of interest.
- **Funding Support:** This article has not been sponsored or funded by any organization or agency. The independence of this research is a crucial factor in affirming its impartiality, as it was conducted without any external influence.
- **Ethical Approval and Consent to Participate:** The data provided in this article is exempt from the requirement for ethical approval or participant consent.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. S. Alimi and A. Oludele, "Voice Activity Detection: Fusion of Time and Frequency Domain Features with an SVM Classifier," Comput. Eng. Intell. Syst., vol. 13, no. 3, pp. 20–29, 2022, DOI: https://doi.org/10.7176/CEIS/13-3-03
2. M. Faghani, H. Rezaee-Dehsorkh, N. Ravanshad, and H. Aminzadeh, "Ultra-Low-Power Voice Activity Detection System Using Level-Crossing Sampling," Electron., vol. 12, no. 4, 2023, DOI: https://doi.org/10.3390/electronics12040795
3. H. Krishnakumar and D. S. Williamson, "A comparison of boosted deep neural networks for voice activity detection," in GlobalSIP 2019 - 7th IEEE Global Conference on Signal and Information Processing, Proceedings, 2019. DOI: https://doi.org/10.1109/GlobalSIP45357.2019.8969258
4. J. N. de Boer et al., "Acoustic speech markers for schizophrenia-spectrum disorders: A diagnostic and symptom-recognition tool," Psychol. Med.., 2021, DOI: https://doi.org/10.1017/S0033291721002804
5. V. Rapcan, S. D'Arcy, S. Yeap, N. Afzal, J. Thakore, and R. B. Reilly, "Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia," Med. Eng. Phys., vol. 32, no. 9, pp. 1074–1079, Nov. 2010, DOI: https://doi.org/10.1016/j.medengphy.2010.07.013
6. R. Makowski and R. Hossa, "Voice activity detection with quasi-quadrature filters and GMM decomposition for speech and noise," Appl. Acoust., vol. 166, 2020, DOI: https://doi.org/10.1016/j.apacoust.2020.107344
7. S. Dwijayanti, K. Yamamori, and M. Miyoshi, "Enhancement of speech dynamics for voice activity detection using DNN," Eurasip J. Audio, Speech, Music Process., vol. 2018, no. 1, 2018, DOI: https://doi.org/10.1186/s13636-018-0135-7
8. M. Lavechin, M. P. Gill, R. Bousbib, H. Bredin, and L. P. Garcia-Perera, "End-to-end domain-adversarial voice activity detection," in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2020. DOI: https://doi.org/10.21437/Interspeech.2020-2285
9. X.-K. Yang, L. He, D. Qu, and W.-Q. Zhang, "Voice activity detection algorithm based on long-term pitch information," 2016, DOI: https://doi.org/10.1186/s13636-016-0092-y
10. C. E. Chelloug and A. Farrouki, "Robust Voice Activity Detection Against Non-Homogeneous Noisy Environments," in 2018 International Conference on Signal, Image, Vision and their Applications, SIVA 2018, 2019. DOI: https://doi.org/10.1109/SIVA.2018.8661045
11. Z. H. Tan, A. K. Sarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," Comput. Speech Lang., vol. 59, 2020, DOI: https://doi.org/10.1016/j.csl.2019.06.005
12. J. Pang, "Spectrum energy based voice activity detection," in 2017 IEEE 7th Annual Computing and Communication Workshop and Conference, CCWC 2017, 2017. DOI: https://doi.org/10.1109/CCWC.2017.7868454
13. Z. Ali and M. Talha, Innovative Method for Unsupervised Voice Activity Detection and Classification of Audio Segments, vol. 6, 2018. DOI: https://doi.org/10.1109/ACCESS.2018.2805845
14. P. D. Ortiz, L. F. Villa, C. Salazar, and O. L. Quintero, "A simple but efficient voice activity detection algorithm through Hilbert transform and dynamic threshold for speech pathologies," J. Phys. Conf. Ser.., vol. 705, no. 1, 2016, DOI: https://doi.org/10.1088/1742-6596/705/1/012037
15. R. J. Elton, P. Vasuki, and J. Mohanalin, "Voice activity detection using fuzzy entropy and support vector machine," Entropy, vol. 18, no. 8, 2016, DOI: https://doi.org/10.3390/e18080298
16. R. Zazo, T. N. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform CLDNNs for Voice Activity Detection," in Proceedings of the Annual Conference of the International Speech Communication

Association, INTERSPEECH, 2016. DOI: https://doi.org/10.21437/Interspeech.2016-268
17. A. Sehgal and N. Kehtarnavaz, "A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection," IEEE Access, vol. 6, 2018, DOI: https://doi.org/10.1109/ACCESS.2018.2800728
18. J. Dey, M. S. Bin Hossain, and M. A. Haque, "An ensemble SVM-based approach for voice activity detection," in ICECE 2018 - 10th International Conference on Electrical and Computer Engineering, 2019. DOI: https://doi.org/10.1109/ICECE.2018.8636745
19. T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice Activity Detection: Merging Source and Filter-based Information," IEEE Signal Process. Lett., vol. 23, no. 2, 2016, DOI: https://doi.org/10.1109/LSP.2015.2495219
20. G. Fant, Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations. Berlin: Gruyter Mouton, 1971. [Online]. Available: DOI: https://doi.org/10.1515/9783110873429
21. H. Uhrmann, R. Kolm, and H. Zimmermann, "Analogue Filters," Springer Ser. Adv. Microelectron., vol. 45, pp. 3–11, 2014, DOI: https://doi.org/10.1007/978-3-642-38013-6_2
22. P. Händel, "Power spectral density error analysis of spectral subtraction type of speech enhancement methods," EURASIP J. Adv. Signal Process., vol. 2007, 2007, DOI: https://doi.org/10.1155/2007/96384

AUTHOR'S PROFILE

Alimi Sheriff, holds a PhD in Computer Science with specialization in Artificial Intelligence from Babcock University, Nigeria; an M.Sc. in Electronic and Computer Engineering from Lagos State University, Nigeria; and an M.Sc. in Information Technology. He was Director of IT Operations and Service Management at 9Mobile, formerly Etisalat Nigeria. He is currently the Head of Technology of an MVNO in Nigeria, where he leads the implementation of the Business Support System and the Network Core Elements. He is a certified Enterprise Architect and Project Manager. He is an adjunct lecturer in the Department of Electronic and Computer Engineering at Lagos State University.

Yussuff Abayomi Isiaka O. (PhD) is an Associate Professor of Communication Engineering. He obtained his bachelor's and master's degrees in Electronic and Computer Engineering from Lagos State University, Nigeria, in 1994 and 2003, and a PhD in Electrical Engineering from Universiti Teknologi Malaysia (UTM), Skudai, Malaysia, in 2014. He is currently a researcher and lecturer in the Department of Electronic and Computer Engineering, Lagos State University, Epe campus, Nigeria. His research areas of interest include radio propagation, rain attenuation, satellite communication, network security, and emerging technologies. He has published a few papers in international journals related to satellite rain attenuation issues in tropical regions. He is a member of the Nigerian Society of Engineers (NSE) and the IEEE, and holds a license issued by the Council for the Regulation of Engineering in Nigeria (COREN).