IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3805~3817
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3805-3817

Unsupervised voice activity detection based on the envelope's fractal dimension

Nesrine Abajaddi[1], Youssef Elfahm[1], Laila Elmazouzi[2], Ilham Mounir[2], Badia Mounir[2], Abdelmajid Farchi[1]
[1]IMII Laboratory, Department of Electrical and Computer Engineering, Faculty of Sciences and Technologies, Hassan First University, Settat, Morocco
[2]LAPSSII Laboratory, High School of Technology, Cadi Ayyad University, Safi, Morocco

Article Info

Article history:

Received Oct 30, 2025
Revised May 24, 2026
Accepted Jun 19, 2026

Keywords:

Fractal dimension
Signal processing
Single frequency filtering
Speech/nonspeech
Unsupervised learning
Voice activity detection

ABSTRACT

Currently, voice activity detection (VAD) is utilized in many fields, including forensics, healthcare, and medicine, to detect vocal anomalies, as well as in telecommunications and mobile telephony. Due to its importance and the difficulty of distinguishing between speech and nonspeech segments, especially in noisy environments (low signal-to-noise ratio (SNR)), this area remains under continuous development. Most existing VAD algorithms require predefined thresholds or training data, which reduces their compatibility. This study proposes an unsupervised VAD system that utilizes the Katz algorithm to calculate the fractal dimension of envelopes obtained through a single frequency filtering (SFF) approach. This method allows for high temporal and frequency resolution. The proposed VAD algorithm does not require any training data and is suitable for various types of noise and SNRs. To evaluate the effectiveness of the proposed method, two different databases are used: the Texas Instruments Massachusetts Institute of Technology (TIMIT) database and the King Saud University (KSU) Arabic speech database. The experimental results reveal an average detection accuracy of 96.86%, demonstrating its considerable value in various applications.

Corresponding Author:

Nesrine Abajaddi
IMII Laboratory, Department of Electrical and Computer Engineering
Faculty of Sciences and Technologies, Hassan First University
Road of Casablanca B.P: 577, Settat, Morocco
Email: n.abajaddi@uhp.ac.ma

1. INTRODUCTION

Voice activity detection (VAD) is a technique for identifying speech and differentiating it from nonspeech segments. In several fields, such as speech coding [1], digital hearing aids [2], speech recognition [3], [4], and voice-controlled applications [2], VAD is a necessary first step. While humans easily differentiate speech from nonspeech, machines suffer from this task. Furthermore, performance is reduced in low signal-to-noise ratio (SNR) situations, such as whispered or strongly accented speech spoken at a low volume, even when deep learning or advanced VAD algorithms are employed.

Numerous algorithms have been described in the literature that exhibit efficient performance when the SNR is high, but struggle in low-SNR environments. Weak phonemes such as fricatives and nasals, as well as low-amplitude segments at the beginning and end of utterances, are challenging to detect in these conditions, leading to poor performance [5]. Traditional VAD algorithms were based on acoustic characteristics such as energy, zero-crossing rate [6], and autocorrelation function [7]. These features perform

well in speech detection when SNR is high, but suffer when SNR is low. Later research explored more advanced features such as higher-order statistics (HOS) of the error signal in linear predictive coding (LPC) [8] and mel frequency cepstral coefficients (MFCC) [9]. In noisy conditions, these characteristics performed better, but they were still insufficiently reliable for practical uses. Another class of VAD algorithms [10] was inspired by Ephraim and Malah [11], which assumes that clean speech and noise have different statistical properties. They applied a Gaussian distribution model to VAD [12] and improved it with Laplace distributions [13] and mixed distributions [14]. One of the main challenges with these VAD algorithms is calculating the SNR, which requires estimating the noise spectrum by assuming that the first dozen frames contain only noise. Therefore, some researchers have proposed combining these parameters to increase the robustness of VADs in various noisy environments or integrating them into deep learning, which has been adopted in recent years. Neural network models have been applied in speech activity detection [15], [16] to determine whether a frame contains speech or nonspeech. Although supervised VAD systems have a low error rate and high accuracy, they require a significant amount of training data to retrain the network to distinguish speech from nonspeech, which limits their real-time use. Unsupervised VAD is commonly used in digital speech processing. Its primary benefit over supervised VAD techniques is that it can differentiate between speech and nonspeech segments without requiring training data or prior knowledge. Instead, unsupervised VAD methods utilize acoustic features to determine thresholds or other parameters in the time and/or frequency domains.

To extract speech features, the majority of research is based on signal processing using the discrete Fourier transform (DFT) [17] or gammatone filters [18], [19]. Recently, a new approach called single frequency filtering (SFF) [20] has offered an alternative to previous methods. To distinguish speech from nonspeech regions, SFF extracts the signal's temporal envelope at each frequency. It achieves this by determining the parameter contour as a function of time after computing the variance and mean of the signal's weighted energy at each frequency. The method was also used in [21], which proposed replacing the SFF approach with a quasi-quadrature filter bank to obtain envelopes and utilized histograms and estimators of probability distributions to determine the detection threshold. Notably, the SFF approach has also been applied to speech intelligibility enhancement [22], epoch extraction from emotional speech [23], and other similar applications. This suggests that it is a promising approach.

This paper aims to develop a reliable, unsupervised VAD system based on the fractal dimension of envelopes. The envelopes are calculated at each frequency using the SFF approach. The fractal dimension is a mathematical tool widely used in various scientific fields. Previous studies have estimated the fractal dimension of electrocardiogram (ECG) signals in [24], [25] to aid in classification. Numerous techniques have been developed to measure the fractal dimension, including those proposed by Katz [26], Higuchi [27], Petrosian [28], Maragos [29], and the amplitude scale approach [30]. In particular, two algorithms, Higuchi and Katz, have been studied for their effectiveness in VADs [31].

An accurate VAD system is crucial in various scientific fields, such as audio forensics, medical diagnosis, and speech recognition applications. The VAD algorithms must be suitable for various noise types and SNR levels, and preferably should not need any training data or prior knowledge of noise. The novelty of this study is the use of fractal dimension analysis of the envelopes extracted by SFF [20] to develop a new approach for VAD. The fractal dimension's robustness to noise and other distortions is an advantage of applying it to VAD. It captures the overall complexity of the signal, rather than relying on specific spectral or temporal features that may be affected by noise or other distortions. Additionally, using the SFF approach, the envelope is calculated with high temporal and frequency resolution, which improves the accuracy of the fractal dimension and, hence, the performance of the proposed method. Existing VAD algorithms use predefined or adjusted thresholds depending on the type of noise and SNR levels [20], [31]–[33]. In contrast, the method automatically determines the threshold of each envelope, regardless of the SNR level or noise type. Moreover, the algorithm requires fewer envelopes compared to the VAD method based on the SFF proposed in [20], thereby reducing complexity and making it suitable for various applications.

This paper is structured as follows: section 2 details the proposed VAD method, which includes SFF, computation of fractal dimension, and automatic threshold estimation. The results obtained are presented in section 3, where two speech databases are evaluated. Additionally, a discussion is provided to analyze and compare the effectiveness of the method. Finally, section 4 concludes with the findings.

2. METHOD

The performance of the suggested approach is assessed using the Texas Instruments Massachusetts Institute of Technology (TIMIT) [34] and King Saud University (KSU) [35], [36] databases, which were recorded at different sampling frequencies (16 kHz and 48 kHz, respectively). To ensure a consistent sampling frequency, the KSU sounds are downsampled to 16 kHz. The SFF technique is used to obtain the signal envelopes at the desired frequency. The steps for calculating the signal envelopes via SFF are

described in the following subsection. To prevent information loss, a zero-padding was implemented at the end of the envelope. To increase the accuracy of segment detection, a shorter window is preferred (10 ms). The fractal dimension is computed for each envelope frame, allowing for automatic speech segment classification. The automatic threshold calculation steps are explained as follows. Figure 1 illustrates the general block diagram of the proposed algorithm.

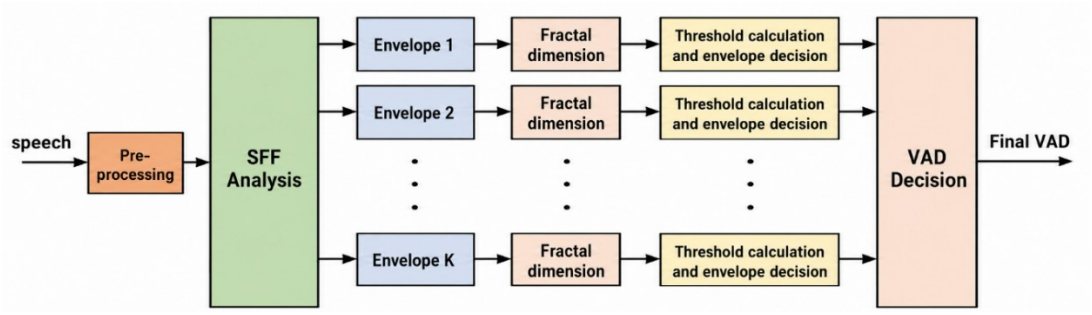


Figure 1. A general block diagram of the proposed method

2.1. Envelopes calculation

To obtain the envelope for a specific frequency, a technique known as SFF is employed [20]. The speech signal is multiplied by a complex sinusoid, and the filtered frequency is shifted using a single-pole filter placed near the unit circle (r > 0.99) at $f_s/2$, where f_s represents the sampling frequency. This ensures that the filter remains stable and consistent across all frequencies, resulting in a more accurate envelope and improved resolution at each frequency. To compute the amplitude envelope of the signal at a desired frequency, the following steps are performed. The shifted signal $x[n]$ at frequency k is as (1).

$$\tilde{x}[k,n] = x[n]e^{-j\frac{2\pi\bar{f}_k}{f_s}n} \quad (1)$$

The desired frequency is denoted by $\bar{f}_k$, where k ranges from 0 to K-1, and n ranges from 1 to N. The total number of samples and frequencies are represented by N and K, respectively. After shifting the signal, it is filtered by a single-pole filter with a transfer function specified as (2).

$$H[z] = \frac{1}{1+rz^{-1}} \quad (2)$$

The output of the filtering process in (3). To ensure stability across all frequencies, the parameter r is chosen to be greater than 0.99. The resulting output, $y[k,n]$, is a complex number and can be represented in polar form using (4). The SFF envelope, denoted by $v[k,n]$, is defined in (5). In contrast, the phase, denoted by $\phi[k,n]$, is defined using (6), where "r" and "i" denote the real and imaginary components of the complex number, respectively.

$$y[k,n] = -ry[k,n-1] + \tilde{x}[k,n] \quad (3)$$

$$y[k,n] = v[k,n]e^{j\phi[k,n]} \quad (4)$$

$$v[k,n] = \sqrt{y_r^2[k,n] + y_i^2[k,n]} \quad (5)$$

$$\phi[k,n] = tan^{-1}\left(\frac{y_i[k,n]}{y_r[k,n]}\right) \quad (6)$$

In this research, the SFF method can obtain the envelope at any desired frequency. However, for practical applications, the envelope was calculated every 100 Hz (Δf =100) within the range of 100 Hz to $f_s//2$, which covers the useful spectral band of speech. This resulted in a total of 80 envelopes that varied over time, compared to a previous study [20] that used 185 envelopes. Using fewer envelopes can reduce the method's implementation time and memory requirements.

2.2. Fractal dimensions and threshold calculation

In this study, the fractal dimensions $FD[k,m]$ were computed for the frame m in the envelope k. The fractal dimensions were obtained by computing the ratio of the normalized length $L[k,m]$ of the

envelope to the maximum distance between the first point and any other point on the envelope. The length $L[k,m]$ (7) was determined by summing up the distances between two successive points using the coordinates X and Y of each point for the frame m of the k^{th} envelope.

$$L[k,m] = \sum_{j=1}^{l-1} \sqrt{(X[k,j+1] - \mathrm{X}[k,j])^2 + (Y[k,j+1] - \mathrm{Y}[k,j])^2} \quad (7)$$

Here, $j = 1.2 \ldots . l-1$, and l represents the sample number for the frame m. As well, the planar extent $D[k,m]$ is defined in (8).

$$D[k,m] = max\left(\sqrt{(X[k,j+1] - \mathrm{X}[k,1])^2 + (Y[k,j+1] - \mathrm{Y}[k,1])^2}\right) \quad (8)$$

To obtain the $FD[k,m]$, the length $L[k,m]$, and planar extent $D[k,m]$ were normalized by the average distance $M[k,m]$, which is the distance between consecutive points on the envelope. The equation for the average distance is given by (9).

$$M[k,m] = mean\left(\sqrt{(X[k,j+1] - \mathrm{X}[k,j])^2 + (Y[k,j+1] - \mathrm{X}[k,j])^2}\right) \quad (9)$$

This approach, defined by the Katz algorithm, determines the fractal dimension $FD[k,m]$ calculated using (10).

$$FD[k,m] = \frac{\log_{10}\left(\frac{L[k,m]}{M[k,m]}\right)}{\log_{10}\left(\frac{D[k,m]}{M[k,m]}\right)} \quad (10)$$

The fractal dimensions are used to differentiate between speech and silence frames. Figure 2 illustrates the first envelope of a speech signal from the KSU database and its corresponding fractal dimension with the VAD decision. Specifically, Figure 2(a) presents the first speech envelope, while Figure 2(b) shows the fractal dimension of the first envelope (blue line) together with the corresponding VAD decision (red line).

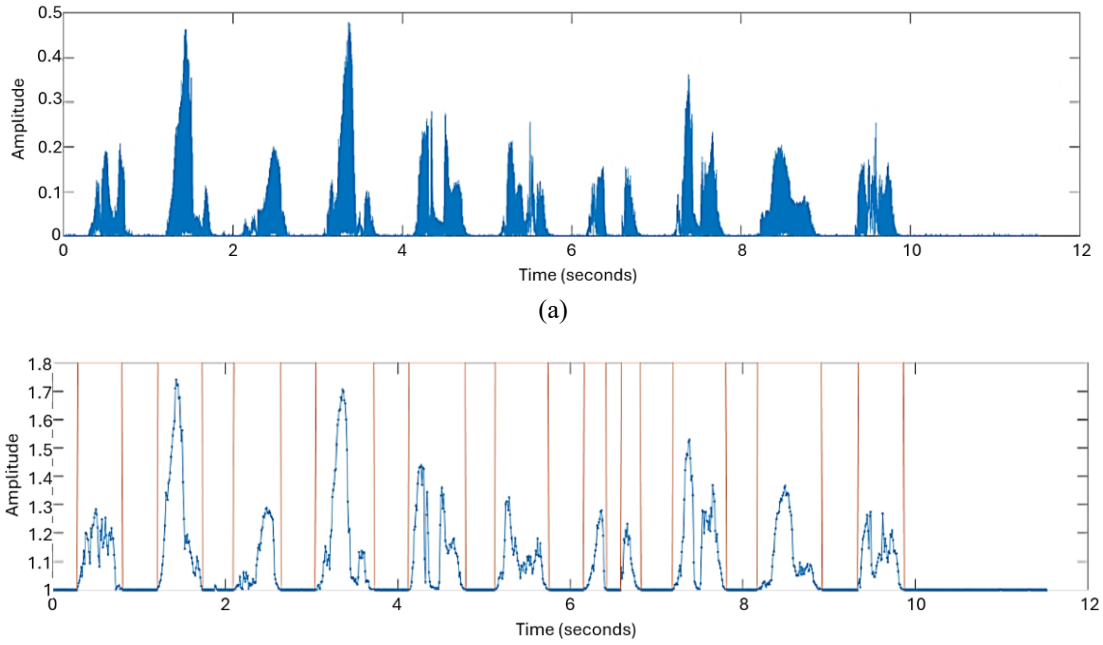


Figure 2. Example of a VAD decision for the first envelope based on its fractal dimension: (a) first envelope of speech of the KSU database and (b) the fractal dimension of the first envelope and its VAD decision

2.3. Logic decision

2.3.1. Threshold determination

To determine the envelope's threshold decision (Figure 2) automatically, was employed the method suggested in [37], [38] was employed to estimate the first significant change (Ind) of the sorted fractal dimension $FD[k,m]$. As expressed in (11), the threshold is calculated by averaging the fractal dimensions for each envelope. The averaging is performed up to the first significant change.

$$Th(k) = mean(sort(FD[k, 1{:}\,\mathrm{Ind}])) \quad (11)$$

2.3.2. Voice activity detection decision

The speech-presence segments for the envelope k are those above the threshold $Th(k)$, and the speech-absence segments are those below the threshold $Th(k)$. The envelope decision $d[k,m]$ is as (12).

$$\begin{aligned} d[k,m] &= 1,\ for\ \ FD[k,m] > Th(k) \\ d[k,m] &= 0,\ for\ \ FD[k,m]\ \leq Th(k) \end{aligned} \quad (12)$$

Where m provides the number of frames for the k^{th} envelope, speech-presence segments are indicated by 1, while speech-absence or noise segments are indicated by 0. On the basis of the envelope's decisions, the Hangover scheme is applied to refine its output and enhance its performance [10], [12], [20]. The hangover threshold is between 20 and 40 ms [20]. For this study, the smoothing threshold employed is 20 ms.

The final VAD decision, $D_f(m)$, is based on the envelope's decision, $d[k,m]$. For clean speech, the algorithm requires a minimum of 40% of the envelopes to exhibit speech activity in order to classify the segment as clean speech [20]. This is due to the fact that clean speech signals typically have a high dynamic range in both time and frequency domains. Conversely, for noisy speech, the algorithm requires a minimum of 60% of the envelopes to exhibit speech activity in order to classify the segment as noisy speech [20]. Consequently, if the percentage is above the threshold percentage η, then the final decision $D_f(m)$ is 1 (i.e., speech is present). Otherwise, the final decision is 0 (i.e., no speech is present), as defined in (13).

$$\begin{cases} D_f(m) = 1,\ for\ \ \sum_{k=1}^{K} d[k,m] \geq \eta \\ D_f(m) = 0,\ for\ \ \sum_{k=1}^{K} d[k,m] < \eta \end{cases} \quad (13)$$

3. RESULTS AND DISCUSSION

In this part, the performance of the developed VAD method is evaluated using two databases, KSU and TIMIT, and compared with other widely recognized VAD algorithms. The performance is tested in both Arabic and English to assess its effectiveness for different languages. Additionally, the robustness of the method is evaluated by introducing different types of noise at varying levels of SNR. Figure 3 illustrates the metrics employed to evaluate the results of the proposed VAD algorithm [39].

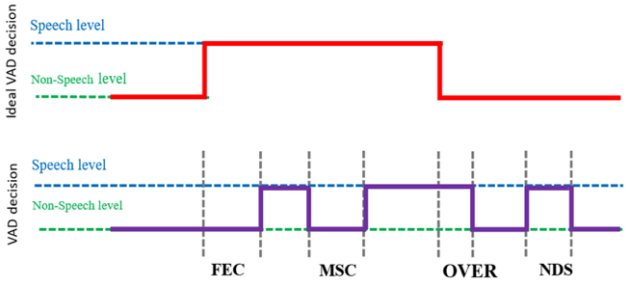


Figure 3. Illustration of the performance metrics

3.1. Performance measures

In evaluating the performance of VAD methods, five parameters proposed by Freeman *et al.* [39] are typically used, as shown in Figure 3 and described as follows:

i) CORRECT: the VAD algorithm's correct decision in detecting speech.
ii) Front-end clipping (FEC): misclassification of speech presence as nonspeech at the start of a speech-presence segment.
iii) Mid speech clipping (MSC): misclassification of speech presence as nonspeech inside a speech-presence segment.
iv) Carry over (OVER): misclassification of nonspeech as speech-presence at the end of a speech-presence segment.
v) Noise detected as speech (NDS): misclassification of nonspeech as speech within a nonspeech segment.

To obtain the percentage value of these parameters, each is divided by the total frame number and then multiplied by 100. The sum of FEC and MSC is referred to as true rejection (TR), representing the percentage of speech-presence rejected. The sum of OVER and NDS is referred to as false acceptance (FA),

representing nonspeech that is detected as speech presence. For better performance, the CORRECT metric should be higher, while TR and FA should be more miniature.

3.2. Databases

In this work, two databases are used: the TIMIT database, which is widely used in voice VAD applications and speech systems, developed by Texas Instruments and the Massachusetts Institute of Technology [34], and the KSU speech Arabic database, created by the speech group (SG) at KSU [35], [36]. Fifty audio samples of the first sentence from the TIMIT database of each speaker are selected to assess the effectiveness of the proposed algorithm. Although the TIMIT sentences contain more speech than nonspeech segments, algorithms generally perform well in terms of speech acceptance but may have poorer performance in nonspeech rejection. To solve this issue, a 2-second silence period was added at the beginning and end of each TIMIT sentence, as described in [20], [37]. For the KSU database, 50 recordings of isolated digits were collected from various microphones used in the office, as they provided clear pauses between spoken words, allowing for an accurate assessment of the algorithm's performance in detecting speech and silence segments. The TIMIT dataset was recorded at a 16 kHz sampling rate, whereas the KSU database was recorded at a 48 kHz sampling rate. Both databases are downsampled to 16 kHz.

3.3. Performance benchmarking

3.3.1. Experimental setup

This section aims to evaluate the performance of both existing and proposed VAD algorithms, using the performance metrics outlined earlier. For this purpose, five VAD algorithms were selected, namely the VAD for adaptive multi-rate option 2 (AMR2), which was normalized for European Telecommunications Standards Institute (ETSI) [32], the SFF approach for discriminating speech and nonspeech (VAD-SFF) [20], the unsupervised VAD and classification of audio segments using fractal dimensions (VAD-FD) [31], and the VAD algorithm using modified global and fuzzy entropy (FuzzyEn) [33]. To assess the performance of the developed algorithm, noise-added speech signals with various SNR levels, ranging from -5 dB to 25 dB in increments of 10 dB, were utilized. Their performance was evaluated on the TIMIT and KSU databases.

3.3.2. Texas Instruments Massachusetts Institute of Technology benchmarking

The evaluation results indicate that the proposed VAD method performs well on clean TIMIT sentences, with an average correct score of 96% for 50 audio files. The ratios of FEC, MSC, NDS, and OVER are 0.04%, 2.32%, 1.27%, and 0.35%, respectively. The analysis shows that the proposed method has a low FEC and OVER ratio, while MSC and NDS are below 2.5%. However, even with manual segmentation of the TIMIT corpus, an accuracy below 100% was obtained, which could explain the values of the ratios. Figure 4 illustrates an example of the VAD decision made by the proposed method (green line) and manual labeling (red line) for a TIMIT speech, demonstrating that the method accurately detects the start (FEC) and end (OVER) of the speech, as well as the short silence between speech segments (MSC), which justifies the MSC value. Overall, the results illustrate the effectiveness of the suggested VAD algorithm.

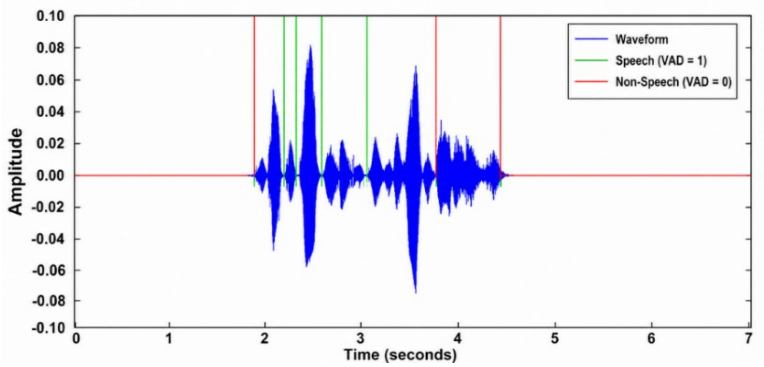


Figure 4. Example of the VAD proposed method and the manual labeling segmentation of TIMIT database

Figure 5 compare the performance of several VAD methods, including AMR2 [32], VAD-SFF [20], FuzzyEn [33], VAD-FD [31], and the proposed method. The evaluation was applied to TIMIT speech signals corrupted by additive white Gaussian noise (Figure 5(a)) and five non-stationary noise types, including pink (Figure 5(b)), leopard Figure 5(c), restaurant (Figure 5(d)), train (Figure 5(e)) and car (Figure 5(f)), at SNR levels ranging from -5 dB to 25 dB. The maximum correct values obtained for 25, 15, 5, and -5 dB are

95.71%, 94.99%, 93.32%, and 88.19%, respectively. Table 1 presents the average detection accuracy for all noise types.

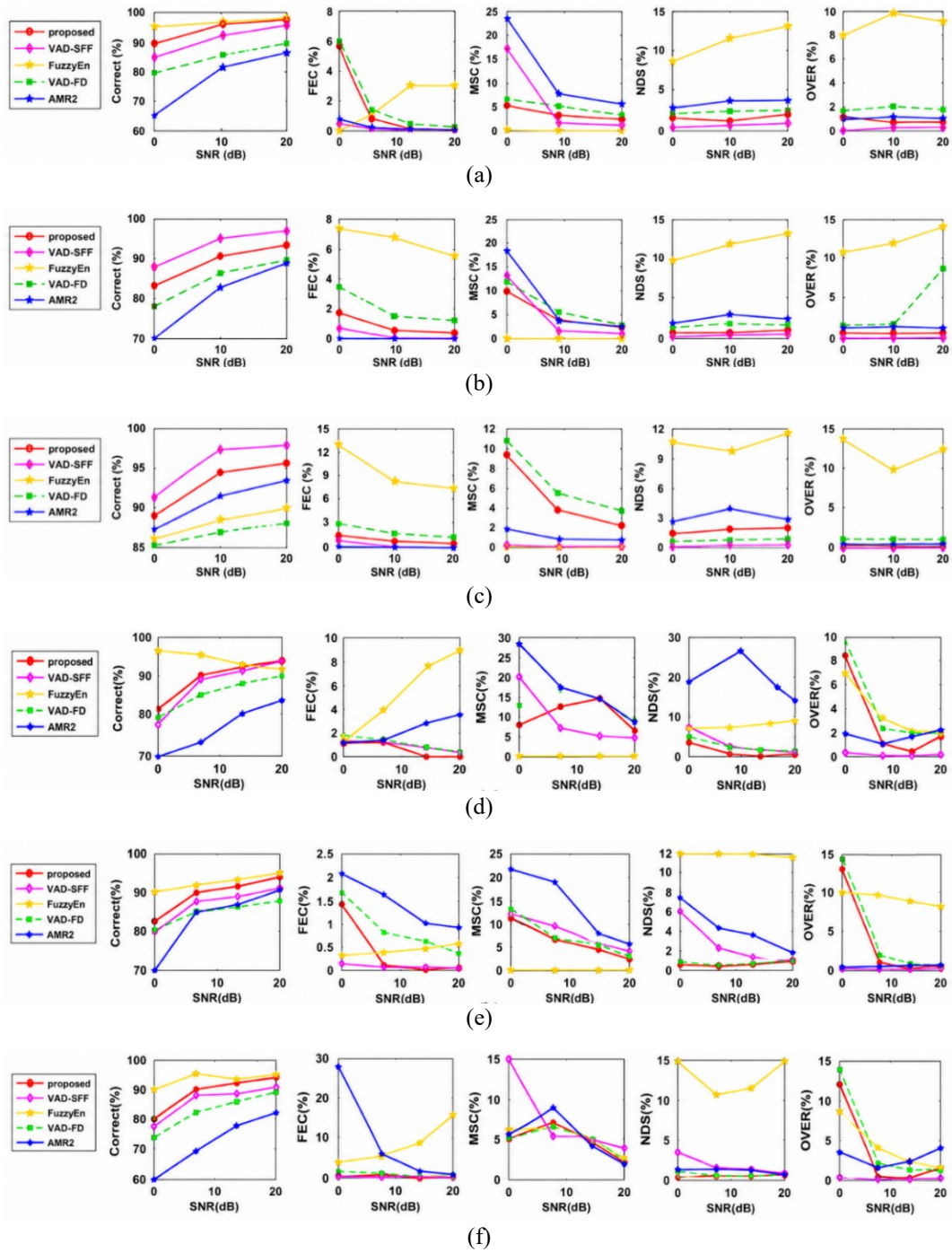


Figure 5. Performance comparison of the TIMIT database as a percentage of CORRECT, FEC, MSC, NDS, and OVER for various noise types at different SNR: (a) white noise, (b) pink noise, (c) leopard noise, (d) restaurant noise, (e) train noise, and (f) car noise

Table 1. Average accuracy of detection (%) at different SNRs

SNR (dB)	VAD methods				
	AMR2 [32]	VAD-SFF [20]	FuzzyEn [33]	VAD-FD [31]	Proposed method
-5	70.17	82.15	92.05	81.31	84.68
5	83.85	94.55	92.37	86.47	92.74
15	88.44	94.66	92.53	88.43	94.26
25	91.45	95.94	93.13	89.92	95.37

3.3.3. KSU benchmarking

For better accuracy, the robustness of the algorithm was evaluated using the Arabic KSU database, recordings in an office room, and with different recording tools. The average correct score for 50 clean speech samples in the KSU database was 98.48%, with negligible NDS and OVER values. The FEC and MSC errors are near zero, 0.88% and 0.65%, respectively. These results demonstrate that the approach performs excellently in non-isolated rooms. To evaluate the robustness of the approach on the KSU database, various types of noise were added, including white, restaurant, train, car, pink, and leopard noise, at different SNR levels (25, 15, 5, and -5 dB). Figure 6 illustrates the average performance of the KSU database with various noise levels at different SNRs. Figure 7 shows the accuracy of the suggested approach for several types of noise at different SNRs. The best results are obtained with accuracy rates of 96.77%, 93.79%, 91.08%, and 81.42% at 25, 15, 5, and -5 dB, respectively. The results indicate that the proposed method also performs well for Arabic noisy speech.

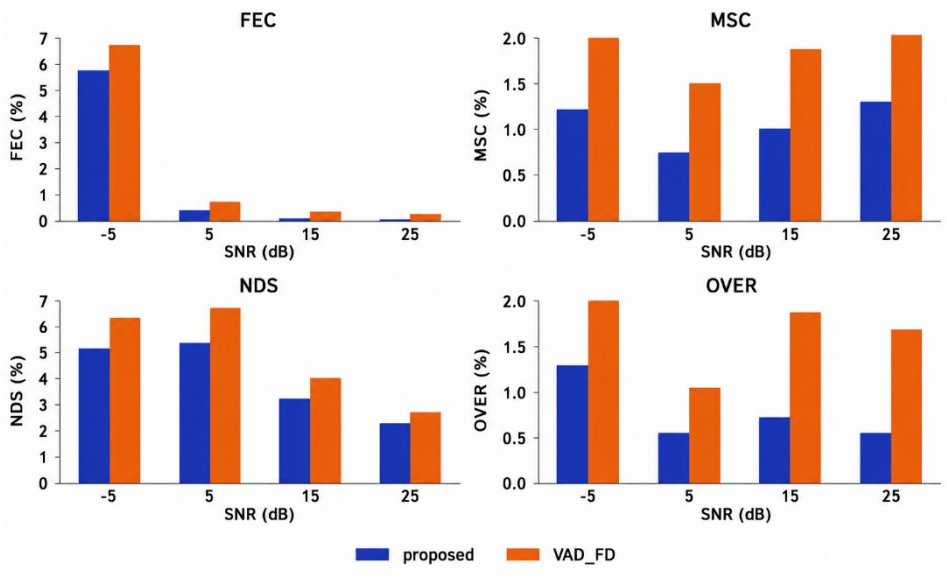


Figure 6. Performance comparison of the KSU database as average for different noise types at various SNR

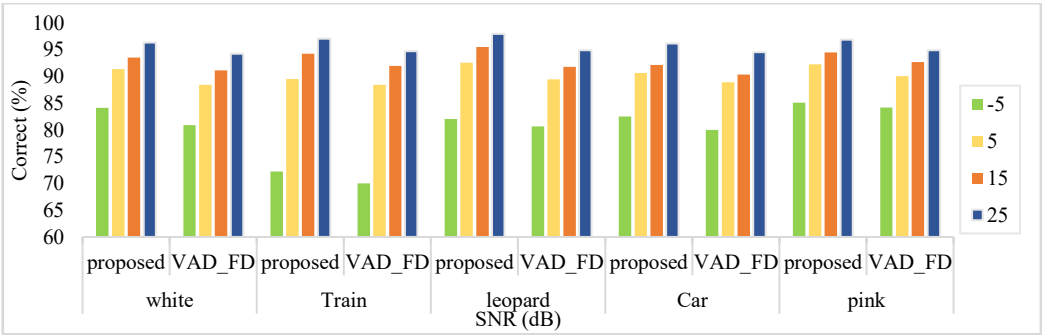


Figure 7. Average accuracy of detection (%) at various SNRs of VAD-FD [31] and the proposed method

3.4. Discussion

The proposed method leverages the fractal dimension of envelopes, derived using the SFF method, to differentiate between speech and nonspeech segments. The SFF approach provides high temporal and spectral resolution for computing the amplitude envelope at each frequency of the speech signal. The Katz algorithm is utilized to determine the fractal dimension, which captures amplitude variation, and the Katz algorithm demonstrates superior performance in detecting speech activity [31].

The results presented in Table 1 from the TIMIT database show that the proposed approach and the VAD-SFF [20] under positive SNR conditions. At the same time, the FuzzyEn [33] method outperforms

them in adverse SNR conditions. However, the FuzzyEn method requires selecting and adjusting the global threshold based on the type of noise and the SNR levels. Notably, the approach achieves an accuracy of 84.64% without any predefined or adjusted threshold, demonstrating its potential for real-world applications. In addition, the algorithm reduces the number of processed envelopes compared to VAD-SFF, which employs 182 envelopes and relies on the statistics feature δ(n) to make a decision based on standard deviation, means, smoothing parameters, and other parameters. In contrast, the proposed method uses only 80 envelopes, more than 50% reduction in the number of envelopes, and automatically calculates the threshold based only on the fractal dimension of the envelopes. The fractal dimension accurately detects amplitude variation, without the need for other parameters, as in VAD-SFF. These advantages make the developed method highly reliable and suitable for favorable SNR conditions, demonstrating good performance for a wide range of applications.

Figure 8 illustrates examples of the VAD decision related to the proposed algorithm for the TIMIT database: Figure 8(a) clean speech, Figure 8(b) speech degraded by white noise at 25 dB, Figure 8(c) at 15 dB, Figure 8(d) at 5 dB, and Figure 8(e) at –5 dB. The VAD labeling across all SNR levels indicates good performance, which can be attributed to the high resolution provided by the SFF and the automatic threshold computation by fractal dimension. The proposed method performs well for positive SNR values, while for negative SNR values (Figure 8(e)), it can accurately detect nonspeech segments without missing speech. In this case, adjusting the percentage to smooth the final decision improves the results.

Similarly, for the KSU database, the proposed method significantly outperforms VAD-FD, as shown in Figure 9. The main difference between the developed algorithm and VAD-FD is that the proposed method calculates the fractal dimension in each frame of each envelope. The SFF approach enables the calculation of the envelope with high temporal and frequency resolution, which in turn enhances the accuracy of the VAD algorithms by improving the fractal dimension. Figure 9 illustrates the VAD decisions of the proposed algorithm for the KSU database: Figure 9(a) clean speech, Figure 9(b) speech degraded by white noise at 25 dB, Figure 9(c) at 15 dB, Figure 9(d) at 5 dB, and Figure 9(e) at –5 dB. The proposed algorithm successfully detects speech for positive SNR values (Figure 9(b) to (d)). For negative SNR values (Figure 9(e)), it accurately detects nonspeech segments without missing speech, and the percentage of final decisions remains acceptable.

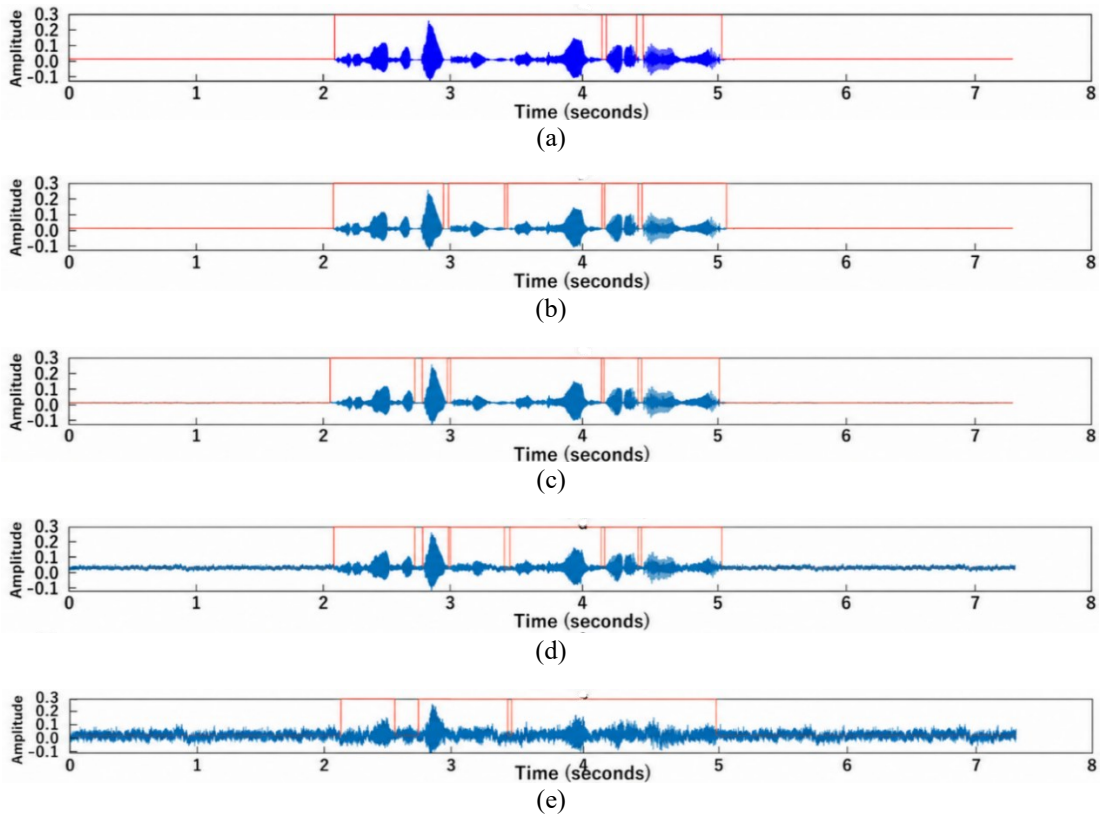


Figure 8. VAD performance of the proposed method for TIMIT database: (a) clean speech, (b) speech degraded by white noise at 25 dB, (c) at 15 dB, (d) at 5 dB, and (e) at -5 dB

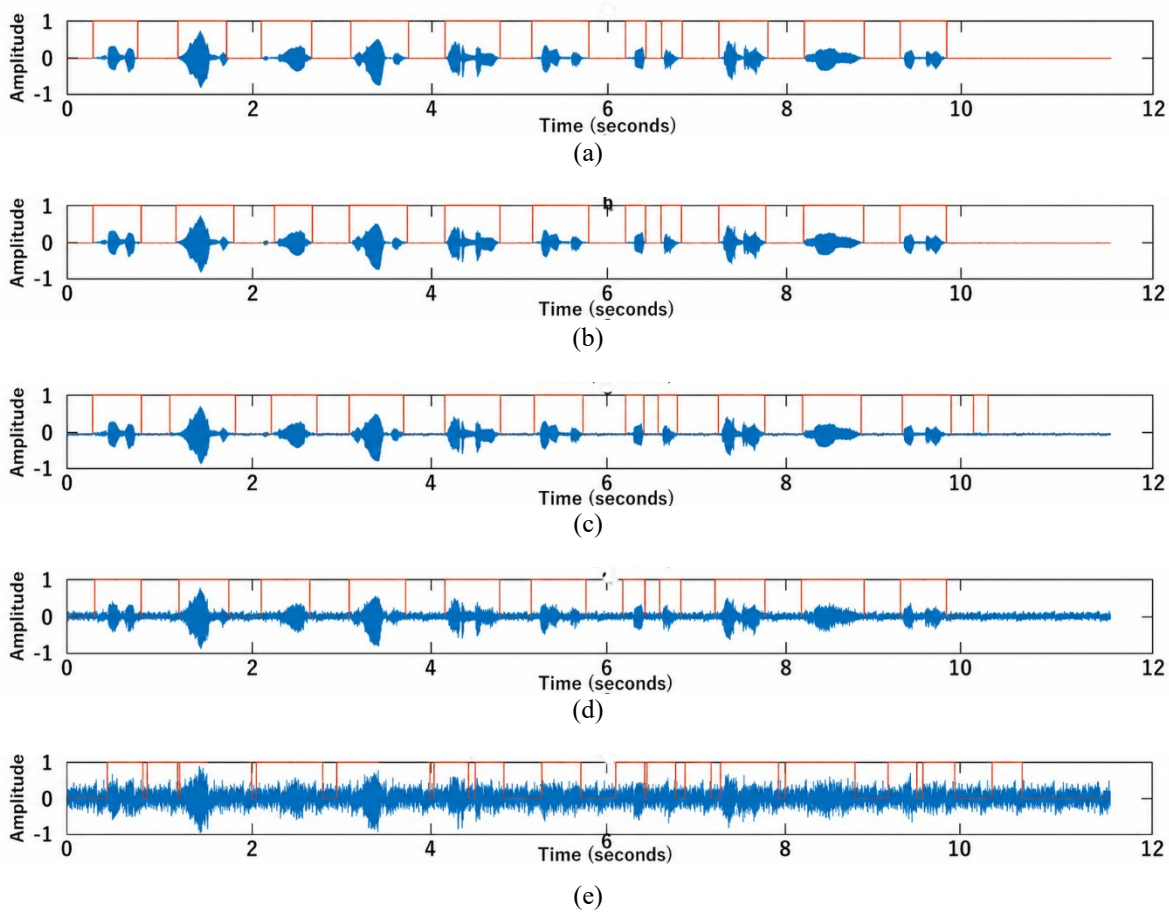


Figure 9. VAD performance of the proposed method for the KSU database: (a) clean speech, (b) speech degraded by white noise at 25 dB, (c) at 15 dB, (d) at 5 dB, and (e) at -5 dB

4. CONCLUSION

In this study, a reliable method for VAD was developed that effectively discriminates between speech and nonspeech segments. The methodology uses a combination of SFF and fractal dimension analysis. The envelopes are calculated using the SFF approach, and the fractal dimension of these envelopes is obtained using the Katz algorithm. Thresholds are then determined automatically for each envelope to make the final VAD decision. The proposed method's performance is evaluated using two different databases, namely, the TIMIT and KSU databases. The results show that the VAD decision is robust for noisy and clean speech. The approach significantly surpassed or matched the performance of other methods such as AMR2, VAD-SFF, FuzzyEn, and VAD-FD. Moreover, this method reduces memory and time consumption and does not require any training data or prior knowledge of speech or noise characteristics. The developed algorithm offers an effective and efficient solution for VAD, enhancing the performance of speech processing applications. Finally, in short term, focus will be placed on hybrid fractal dimension–machine learning models, and in long term for real time deployment in speech recognition systems.

ACKNOWLEDGEMENTS

The authors express their gratitude to the Researchers Supporting Project (RSP2023R34), King Saud University, Riyadh, Saudi Arabia, for the support received.

FUNDING INFORMATION

The authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nesrine Abajaddi	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Youssef Elfahm	✓	✓	✓			✓		✓	✓	✓	✓	✓		
Laila Elmazouzi			✓	✓	✓		✓			✓	✓	✓	✓	
Ilham Mounir		✓		✓					✓	✓				
Badia Mounir	✓			✓	✓	✓				✓				✓
Abdelmajid Farchi					✓		✓			✓		✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, [NA]. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

REFERENCES

[1] S. M. Joseph and A. P. Babu, "Wavelet energy-based voice activity detection and adaptive thresholding for efficient speech coding," *International Journal of Speech Technology*, vol. 19, no. 3, pp. 537–550, Sep. 2016, doi: 10.1007/s10772-014-9240-x.

[2] J. G.-Gómez, R. G. -Pita, M. A.-Ortega, M. U.-Manso, M. R.-Zurera, and I. M.-Herranz, "Linear detector and neural networks in cascade for voice activity detection in hearing aids," *Applied Acoustics*, vol. 175, Apr. 2021, doi: 10.1016/j.apacoust.2020.107832.

[3] T. Yoshimura, T. Hayashi, K. Takeda, and S. Watanabe, "End-to-end automatic speech recognition integrated with CTC-based voice activity detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain: IEEE, May 2020, pp. 6999–7003, doi: 10.1109/ICASSP40776.2020.9054358.

[4] M. F. Alghifari, T. S. Gunawan, M. A. W. Nordin, S. A. A. Qadri, M. Kartiwi, and Z. Janin, "On the use of voice activity detection in speech emotion recognition," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, pp. 1324–332, Dec. 2019, doi: 10.11591/eei.v8i4.1646.

[5] J.-C. Junqua and J.-P. Haton, "Dealing with noisy speech and channel distortions," in *Robustness in Automatic Speech Recognition*, Boston, MA: Springer, 1996, pp. 155–189, doi: 10.1007/978-1-4613-1297-0_5.

[6] R. G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *2008 ASEE Zone 1 Conference*, 2008, pp. 1-7, doi: 10.18260/1-2-1153-53698.

[7] B. F. Wu and K. -C. Wang, "Voice activity detection based on auto-correlation function using wavelet transform and Teager energy operator," *Computational Linguistics and Chinese Language Processing*, vol. 11, no. 1, pp. 87–100, 2006.

[8] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 217–231, Mar. 2001, doi: 10.1109/89.905996.

[9] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," in *Interspeech 2005*, ISCA, Sep. 2005, pp. 369–372, doi: 10.21437/Interspeech.2005-186.

[10] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 412–424, Mar. 2006, doi: 10.1109/TSA.2005.855842.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984, doi: 10.1109/TASSP.1984.1164453.

[12] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan. 1999, doi: 10.1109/97.736233.

[13] J.-H. Chang and N. S. Kim, "Voice activity detection based on complex Laplacian model," *Electronics Letters*, vol. 39, no. 7, pp. 632–634, Apr. 2003, doi: 10.1049/el:20030392.

[14] J.-H. Chang, J. W. Shin, and N. S. Kim, "Voice activity detector employing generalised Gaussian distribution," *Electronics Letters*, vol. 40, no. 24, pp. 1561–1563, Nov. 2004, doi: 10.1049/el:20047090.

[15] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7378–7382.

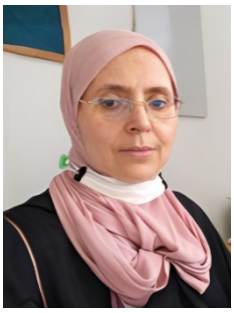
[16] Y. Xiong, V. Berisha, and C. Chakrabarti, "Computationally-efficient voice activity detection based on deep neural networks," in *2021 IEEE Workshop on Signal Processing Systems (SiPS)*, Coimbra, Portugal, Oct. 2021, pp. 64–69, doi: 10.1109/SiPS52927.2021.00020.
[17] X. Huang, Z. Liu, W. Lu, H. Liu, and S. Xiang, "Fast and effective copy-move detection of digital audio based on auto segment," *International Journal of Digital Crime and Forensics*, vol. 11, no. 2, pp. 47–62, Apr. 2019, doi: 10.4018/IJDCF.2019040104.
[18] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acustica united with acta acustica*, vol. 88, no. 3, pp. 433–442, 2002.
[19] W. Q. Ong and A. W. C. Tan, "Robust voice activity detection using gammatone filtering and entropy," in *2016 International Conference on Robotics, Automation and Sciences (ICORAS)*, Melaka, Malaysia, Nov. 2016, pp. 1–5, doi: 10.1109/ICORAS.2016.7872630.
[20] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 705–717, Apr. 2015, doi: 10.1109/TASLP.2015.2404035.
[21] R. Makowski and R. Hossa, "Voice activity detection with quasi-quadrature filters and GMM decomposition for speech and noise," *Applied Acoustics*, vol. 166, Sep. 2020, doi: 10.1016/j.apacoust.2020.107344.
[22] N. Chennupati, S. R. Kadiri, and Y. B., "Spectral and temporal manipulations of SFF envelopes for enhancement of speech intelligibility in noise," *Computer Speech & Language*, vol. 54, pp. 86–105, Mar. 2019, doi: 10.1016/j.csl.2018.09.002.
[23] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52–63, Feb. 2017, doi: 10.1016/j.specom.2016.11.005.
[24] Y. W. Kim, K. K. Krieble, C. B. Kim, J. Reed, and A. D. R. -Grant, "Differentiation of alpha coma from awake alpha by nonlinear dynamics of electroencephalography," *Electroencephalography and Clinical Neurophysiology*, vol. 98, no. 1, pp. 35–41, Jan. 1996, doi: 10.1016/0013-4694(95)00186-7.
[25] A. K. Mishra and S. Raghav, "Local fractal dimension based ECG arrhythmia classification," *Biomedical Signal Processing and Control*, vol. 5, no. 2, pp. 114–123, Apr. 2010, doi: 10.1016/j.bspc.2010.01.002.
[26] M. J. Katz, "Fractals and the analysis of waveforms," *Computers in Biology and Medicine*, vol. 18, no. 3, pp. 145–156, Jan. 1988, doi: 10.1016/0010-4825(88)90041-8.
[27] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D: Nonlinear Phenomena*, vol. 31, no. 2, pp. 277–283, Jun. 1988, doi: 10.1016/0167-2789(88)90081-4.
[28] A. Petrosian, "Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns," in *Proceedings Eighth IEEE Symposium on Computer-Based Medical Systems*, Lubbock, United States, 1995, pp. 212–217, doi: 10.1109/CBMS.1995.465426.
[29] P. Maragos, "Fractal aspects of speech signals: Dimension and interpolation," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 1991, pp. 417–420, doi: 10.1109/icassp.1991.150365.
[30] T. R. Senevirathne, E. L. J. Bohez, and J. A. Van Winden, "Amplitude scale method: new and efficient approach to measure fractal dimension of speech waveforms," *Electronics Letters*, vol. 28, no. 4, pp. 420–422, Feb. 1992, doi: 10.1049/el:19920264.
[31] Z. Ali and M. Talha, "Innovative method for unsupervised voice activity detection and classification of audio segments," *IEEE Access*, vol. 6, pp. 15494–15504, 2018, doi: 10.1109/ACCESS.2018.2805845.
[32] 3GPP, "ANSI-C code for the floating-point adaptive multi-rate (AMR) speech codec," portal.3gpp.org. Accessed: Jun. 29, 2025. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=1400
[33] R. J. Elton, J. Mohanalin, and P. Vasuki, "A novel voice activity detection algorithm using modified global thresholding," *International Journal of Speech Technology*, vol. 24, no. 1, pp. 127–142, Mar. 2021, doi: 10.1007/s10772-020-09777-w.
[34] J. S. Garofolo, L. F. Lamel, W. M. Fischer, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *The DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM*, Gaithersburg, United States: Department of Commerce USA, 1986.
[35] M. Alsulaiman, Z. Ali, G. Muhammed, M. Bencherif, and A. Mahmood, "KSU speech database: text selection, recording and verification," in *2013 European Modelling Symposium*, Nov. 2013, pp. 237–242, doi: 10.1109/EMS.2013.41.
[36] M. Alsulaiman, G. Muhammad, M. A. Bencherif, A. Mahmood, and Z. Ali, "KSU rich Arabic speech database," *Information*, vol. 16, no. 6 B, pp. 4231–4253, 2013.
[37] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, Mar. 2011, doi: 10.1109/TASL.2010.2052803.
[38] M. Masud, G. S. Gaba, K. Choudhary, M. S. Hossain, M. F. Alhamid, and G. Muhammad, "Lightweight and anonymity-preserving user authentication scheme for IoT-based healthcare," *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 2649–2656, Feb. 2022, doi: 10.1109/JIOT.2021.3080461.
[39] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the Pan-European digital cellular mobile telephone service," in *International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, United Kingdom: IEEE, 1989, pp. 369–372, doi: 10.1109/ICASSP.1989.266442.

BIOGRAPHIES OF AUTHORS

Nesrine Abajaddi received a master's degree specializing in automatic control, signal processing, and industrial computing from Hassan First University in Settat, Morocco, in 2017. She is currently a Ph.D. in Engineering, Mechanical, Industrial Management, and Innovation (LIMII), Faculty of Sciences and Techniques, Hassan First University. She can be contacted at email: n.abajaddi@uhp.ac.ma.

Youssef Elfahm received a master's degree specializing in automatic control, signal processing and industrial computing from the Hassan First University in Settat, Morocco, in 2017. Currently, he is a professor in the Department of Electrical Engineering at Alkhaouarizmi Technical High School. His research interests include speech recognition systems, speech production, and artificial intelligence. He can be contacted at email: y.elfahm@uhp.ac.ma.

Laila Elmazouzi received the Ph.D. degree in Telecommunication and Networks. She is habilitated to supervise research (HDR) at the High School of Technology, Cadi Ayyad University. She is a member of the LAPSSII Laboratory (Laboratory of Process, Signals, Industrial Systems, Informatics). Her research interests include telecommunications, signal processing, emotion recognition, and machine learning. She can be contacted at email: Elmazouzi2001@yahoo.fr.



Ilham Mounir received the Ph.D. in Applied Mathematics. She is habilitaded to supervise research (HDR) at the High School of Technology, Cadi Ayyad University. She is a member of the LAPSSII Laboratory (Laboratory of Process, Signals, Industrial Systems, Informatics). Her research interests include applied mathematics, signal processing, emotion recognition, speech recognition, and energy: optimization and modeling. She can be contacted at email: ilhamounir@gmail.com.

Badia Mounir is engineer degree (1992) in Automatic and Industrial Computing, The Mohammadia School of Engineering, Rabat, Morocco. She is assistant professor at the Graduate School of Technology, Cadi Ayyad University since 1992. She is habilitaded to supervise research (HDR) since 2007 and professor of higher education (PES) since 2017. She is member of the Laboratory of Process, Signals, Industrial Systems, Informatics (LAPSSII). Her research interests include speech recognition, signal processing, energy optimization, and modeling. She can be contacted at email: Mounirbadia2014@gmail.com.

Abdelmajid Farchi received the Ph.D. degree in Electrical Engineering and Telecommunications. He is the chief of the research team signals and systems in the Laboratory of Engineering, Industrial Management, and Innovation. He is the educational person responsible for the cycle engineer, electrical systems, and embedded systems of the Faculty of Sciences and Technology of Settat, Hassan First University, Morocco. He can be contacted at email: abdelmajid.farchi1@gmail.com.