IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 1, 2018, accepted February 10, 2018, date of publication February 13, 2018, date of current version April 4, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2805845

Innovative Method for Unsupervised Voice Activity Detection and Classification of Audio Segments

ZULFIQAR ALI[1] **AND MUHAMMAD TALHA**[2]

[1]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[2]Deanship of Scientific Research, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding authors: Zulfiqar Ali (zuali@ksu.edu.sa)

This work was supported by the Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia, through the Research Group under Project RG-1437-037.

ABSTRACT An accurate and noise-robust voice activity detection (VAD) system can be widely used for emerging speech technologies in the fields of audio forensics, wireless communication, and speech recognition. However, in real-life application, the sufficient amount of data or human-annotated data to train such a system may not be available. Therefore, a supervised system for VAD cannot be used in such situations. In this paper, an unsupervised method for VAD is proposed to label the segments of speech-presence and speech-absence in an audio. To make the proposed method efficient and computationally fast, it is implemented by using long-term features that are computed by using the Katz algorithm of fractal dimension estimation. Two databases of different languages are used to evaluate the performance of the proposed method. The first is Texas Instruments Massachusetts Institute of Technology (TIMIT) database, and the second is the King Saud University (KSU) Arabic speech database. The language of TIMIT is English, while the language of the KSU speech database is Arabic. TIMIT is recorded in only one environment, whereas the KSU speech database is recorded in distinct environments using various recording systems that contain sound cards of different qualities and models. The evaluation of the proposed method suggested that it labels voiced and unvoiced segments reliably in both clean and noisy audio.

INDEX TERMS Voiced and unvoiced segmentation, fractal dimension, Katz algorithm, TIMIT database, KSU speech database.

I. INTRODUCTION

Voice activity detection (VAD) is a process that divides speech signals into at least two types of segments, referred to as speech-presence and speech-absence. When an audio contains low speech-to-signal ratio segments, it creates a negative impact on the performance of the systems using speech-processing techniques [1]. Therefore, VAD plays a significant role in such kinds of systems. Various VAD systems have been designed and developed [1]–[7]. Most studies [1]–[3], [5]–[8] implemented the likelihood-ratio-test-based decision rule for a set of hypotheses. VAD can be identified as a statistical hypothesis problem, where the purpose is to determine the class of the segment in an audio, i.e., the class of the segment in this case is either speech-presence or speech-absence. The decision of a segment depends on the computed feature vectors, which are a vital part of a VAD system. The feature vector extracts the characteristics of a segment and serves as the input to a decision rule that assigns a sample vector to one of the given classes. The drawback of such system is the computational cost of the statistical model implemented to label the segments. On the other hand, unsupervised VAD systems [10], [11] automatically detect the acoustic patterns in an audio by using signal properties. Human-annotated data for acoustic model training are no longer needed. The main goal of the current research is to develop an accurate and reliable unsupervised VAD method based on fractal dimension.

To estimate the fractal dimension, algorithms such as Katz [12], Higuchi [13], Petrosian [14], Maragos [15], and the amplitude scale method [16] have been proposed. These algorithms are used in various scientific areas to compute the fractal dimension of time series and waveforms [17]. In [18] and [19], the fractal dimension of the electrocardiogram signals is calculated to differentiate them into categories. A speech signal can also be seen as a waveform, and to measure its complexity, an algorithm of fractal dimension can be implemented [20], [21]. Some algorithms for measuring the fractal dimension are evaluated and compared in [22]. Two algorithms proposed by Higuchi and Katz are analyzed in [23] to observe their dependency on various factors, such as amplitude and sampling frequency of a signal.

An accurate VAD system with the ability to classify voice segments into speech-presence and speech-absence has many potential applications in different scientific domains, such as speech recognition systems (SRSs), medical diagnosis systems, and audio forensics systems. In an SRS, the detection of speech-presence segments is crucial for generating accurate acoustic models of different words or phonemes depending on the nature of the developed SRS. Inaccurate detection of speech-presence and speech-absence segments by the VAD method may decrease the recognition rate and, ultimately, the performance of the system. The role of the VAD method in the diagnosis of the vocal folds disorder (VFD) system is vital. Most published studies have used sustained vowel for the detection of VFD due to the nonexistence of speech-absence segments. Sustained vowel contains only speech-presence segments, which make the analysis of speech easier. Long-term features such as fundamental frequency, shimmer, and jitter become unreliable in the presence of speech-absence segments. Hence, the VAD method is essential for the computation of long-term features to avoid the speech-absence segments. VAD methods are equally critical in audio forensics, i.e., audio authentication, forgery detection, and localization in audio. Audio forgery can be accomplished by copy-move [24], deletion, insertion, substitution, and splicing [25], [26]. In any case, the detection of speech-presence segments is an important step. In [27], a forged audio database is generated by mixing audio recording in different environments with different recording systems. For forgery, words from the audios are extracted carefully by means of VAD. If words are not extracted properly from the audios, then their mixing will not be flawless. The forged audio is generated sophisticatedly such that nobody may guess its type, neither listening nor by visualizing. In another study [28], a database of forged audio is generated by copying words from one location and moving them to another. Fig. 1 shows that forgery cannot be detected by visual inspection. In [28], a novel method for the blind detection and localization of copy-move forgery is presented. The VAD method is one of the most crucial components for investigating audio recordings to detect and localize the forgery. It was also concluded that copy-move forgery is incorrectly detected in 3.41% of forged recordings. Further investigations revealed that the boundary points of copy-move words are incorrectly determined. When boundary points are inaccurately calculated, the detection and localization of the forgery will be erroneous.

This paper presents an accurate and efficient unsupervised VAD method to classify speech-presence and speech-absence segments in an audio. To the best of our knowledge, an unsupervised VAD method that uses fractal dimension has never been implemented to label segments in audios. The proposed method divides audio samples into shorter frames. Then, fractal dimensions are computed for each frame of an audio to keep the method more efficient. Katz algorithm is employed to calculate the fractal dimension. With the use of the computed fractal dimension, a threshold to detect speech-presence and speech-absence segments is calculated automatically. The threshold varies from one audio to another. Therefore, the method is robust against recording environment and equipment. The method is evaluated by using two speech databases, and it accurately detected speech-presence and speech-absence segments for audio samples from both databases.

The rest of the paper is organized as follows. Section II explains the various steps of the proposed VAD method, including preprocessing, fractal dimension estimation, and automatic adjustment of the threshold. Section III provides the performance evaluation of the proposed method by using two speech databases: an English speech database called TIMIT and the King Saud University (KSU) Arabic speech database. Section IV discusses the proposed method. Finally, Section V draws some conclusions.

II. PROPOSED METHOD

The proposed method for automatic VAD identifies the speech-absence and speech-presence segments in an audio. It performs preprocessing of the audio before extraction of long-term features by using the fractal dimension estimation algorithm. The segments are categorized by using the computed fractal dimensions, the threshold of which is adjusted automatically for each audio.

A. PREPROCESSING OF AUDIO

The first step of the proposed method is the preprocessing of audio samples. The sampling frequencies of all audio samples are down-sampled to 16 KHz because one of the databases is recorded at 16 KHz and the other is at 48 KHz. By doing so, all audios of both databases will have a unique sampling frequency and can be used to evaluate the performance of the proposed method.

To detect the segments of speech-presence and speech-absence, audios are partitioned into frames of shorter durations. The durations of frames are kept shorter for accurate classification of both types of segments. The drawback of long frame duration is that it may contain speech in some parts, and the remaining parts may contain silence or short pauses. Therefore, each audio is divided into frames of 10 milliseconds. Consider an audio D, given in Eq. (1), with its samples d_1, d_2, $d_3, \ldots, d_N$, where N represents a total number of samples in D

$$D = [d_1, d_2, d_3, \ldots, d_N]\,. \tag{1}$$

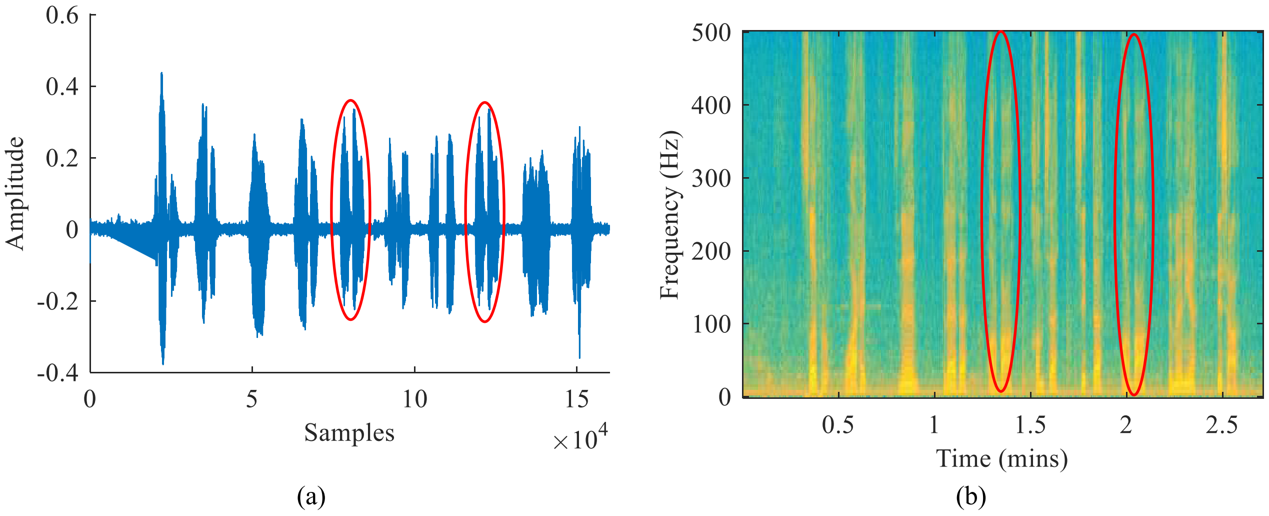

FIGURE 1. Forged audio and its spectrogram: (a) an audio with copy-move forgery. (b) a spectrogram of the forged audio [9].

The i^{th} frame F_i can then be obtained as

$$F_i = \left[d_{(i-1)l+1}, d_{(i-1)l+2}, d_{(i-1)l+3}, \ldots, d_{il-1}, d_{il}\right]$$

where

$$l = s \times r \text{ and } 1 \leq i \leq n \left(= \frac{N}{l}\right) \tag{2}$$

In Eq. (2), s is the sampling frequency of an audio, r stands for the duration of a frame in seconds, l represents the number of samples in a frame, and n provides a total number of frames for an audio D.

After the division of an audio into frames [F_1, F_2, $F_3, \ldots, F_n$], some samples at the end of an audio may not be a part of any frame because of the insufficient amount of samples. For instance, in the case of a 10 millisecond frame ($r = 0.010$ seconds) of an audio recorded at 16 KHz ($s = 16,000$), each frame will contain 160 samples ($l = 160$), and it is possible that 100 samples at the end of an audio are not part of any frame, when n is not an integer. These samples may contain some important information that is vital for the development of speech-related systems. Zero padding is performed at the end of such audios to generate a complete frame by using the unused samples. In this way, no information in an audio will be lost. The last frame F_n of an audio D can be obtained using Eq. (3):

$$F_n = \left[d_{\lfloor \frac{N}{l} \rfloor l+1}, d_{\lfloor \frac{N}{l} \rfloor l+2}, d_{\lfloor \frac{N}{l} \rfloor l+3}, \ldots, 0, 0, 0, \ldots, 0\right] \tag{3}$$

where $\lfloor . \rfloor$ is the floor operator. The number of zeros in the frame F_n is determined by using Eq. (4):

$$\text{Number of zeros} = l - \text{mod}\,(N, l). \tag{4}$$

In the next step, the features for the automatic identification of speech-presence and speech-absence segments are computed. The computed features are fractal dimensions of an audio. The steps to calculate the fractal dimension of each frame is described in the following subsection.

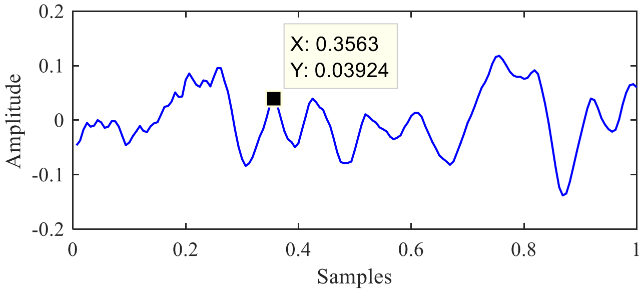

FIGURE 2. Waveform of speech signal in a frame.

B. ESTIMATION OF FRACTAL DIMENSIONS

Fractal dimensions for each divided frame of an audio are calculated by using the Katz algorithm. To calculate the fractal dimension of a frame, the length of the waveform is computed by adding the Euclidean distances between all consecutive points on the waveform. Each point on the waveform is represented by an ordered pair (X, Y). In Fig. 2 the point highlighted by a black rectangle is represented by the coordinates (0.3563, 0.03924). The length M of the waveform is computed by using Eq. (5):

$$M = \sum_{j=1}^{l-1} \sqrt{\left(X_{j+1} - X_j\right)^2 + \left(Y_{j+1} - Y_j\right)^2}. \tag{5}$$

To calculate the fractal dimension, the next step is to calculate the planar extent P, which is the maximum distance between the first and any point on the waveform. The planar extent is given by Eq. 6.

$$P = \max\left(\sqrt{\left(X_{j+1} - X_1\right)^2 + \left(Y_{j+1} - Y_1\right)^2}\right) \tag{6}$$

where $j = 1, 2, 3, \ldots, l-1$

Then, length M and planar extent P are normalized by dividing them with the average distance between consecutive points on the waveform. The average distance V is computed by using Eq. (7).

$$V = \text{mean}\left(\sqrt{\left(X_{j+1} - X_j\right)^2 + \left(Y_{j+1} - Y_j\right)^2}\right) \tag{7}$$

where $j = 1, 2, 3, \ldots, l - 1$

The required fractal dimension T is a ratio between normalized length and planar extent of the waveform, and it is obtained by using Eq. (8),

$$T = \frac{\log_{10}\left(\frac{M}{V}\right)}{\log_{10}\left(\frac{P}{V}\right)}. \tag{8}$$

The fractal dimension for all audio samples is computed by following the procedure. On the basis of the fractal dimensions of each frame, it will be decided whether a certain frame contains speech or silence.

C. THRESHOLD FOR DECISION

The automatic decision for VAD in the proposed method is made by computing a threshold. The fractal dimension above the threshold will determine if it is a speech-presence or a speech-absence segment. To determine the threshold, the computed fractal dimensions are sorted in ascending order, as shown in Fig. 3(a). The first most significant change in the sorted fractal dimension is estimated by using the process proposed by Killick *et al.* [29] and Lavielle [30]. The vertical line in Fig. 3(b) signifies the position of the first most significant change in the curve. The threshold, given in Eq. (9), is computed by taking the average of sorted fractal dimensions up to the occurrence of the first most significant change.

$$\mathit{Threshold} = \text{mean}\left(\text{sort}\left(\mathit{fractal}\ (1 : \mathit{Ind})\right)\right) \tag{9}$$

where *fractal* stands for the fractal dimensions of all segments of an audio, and *Ind* represents the intersection point of the vertical line with x-axis in Fig. 3(b). All segments above the threshold will be categorized as speech-presence segments, while those below the threshold will be labeled as speech-absence segments.

III. PERFORMANCE EVALUATION

The effectiveness of the proposed method is evaluated by using two speech databases recorded in different languages, English and Arabic. The method is evaluated by using clean and noisy audio to observe its robustness against noise.

A. EVALUATION BY USING THE TIMIT DATABASE

The Texas Instruments Massachusetts Institute of Technology (TIMIT) database [31] is used to evaluate the proposed method. The database is developed to provide speech data for obtaining acoustic-phonetic knowledge. The TIMIT database is recorded at Texas Instruments and then transcribed at the Massachusetts Institute of Technology. The database is distributed by the National Institute of Standards and Technology.

The database has been widely used in VAD systems [1], [32] and various speech-related applications [33]–[37]. The TIMIT corpus is recorded by 630 male and female speakers from eight dialect regions in the United States. The language of the TIMIT database is English, and each speaker recorded 10 sentences at 16 KHz sampling frequency and 16-bit rate. For each speaker, the first two sentences are fixed, while the remaining sentences vary from one speaker to another. All sentences are read by the speakers, and they are recorded in one session in a soundproof room.

Audio samples of the TIMIT database are partitioned into frames, and the fractal dimension is calculated for each frame. A clear difference between the fractal dimension of speech-presence and speech-absence segments can be observed in Fig. 4. The fractal dimensions are approximately equal to 1 for the silence parts and greater than 1 for the speech-presence parts.

A threshold is automatically calculated for each audio to determine the voice (V) and unvoiced (U) segments. The frames of audio above the threshold are referred to as speech-presence (voiced), and the segments lying below the threshold are considered as speech-absence (unvoiced) segments. The parts of an audio are automatically labeled as voiced and unvoiced segments in Fig. 5. In the proposed method, the classification of segments is done through unsupervised VAD; therefore, no training data are required, which is the most discriminant aspect of the method.

The performance of the proposed method is evaluated by using the metric defined in Eq. (10). The metric provides frame-level accuracy, and it is also used in [38]. In Eq. (10), TP stands for true positive, meaning the method detected a voice frame as a voice frame. FN represents the false negative, which means the method detected the voice frame as an unvoiced frame. TN stands for the true negative, which means the method detected an unvoiced frame as an unvoiced frame. Finally, FP represents the false positive, which means the method detected an unvoiced frame as a voiced frame.

The accuracy of the method for some clean audio of the TIMIT database is shown in Fig. 6. The first sentence of the TIMIT database is used. The average accuracy of the randomly selected 50 audio is 90.45%. The method is also evaluated by adding different types of noise in audios. Three types of noise with different signal-to-noise ratios (SNRs) are added in the audios to observe the robustness of the proposed system. The accuracy of the method with noisy audio is listed in Table 1.

$$\begin{aligned}&\mathit{Accuracy}(\%)\\ &= \frac{TP^*TN - FP^*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \times 100\end{aligned} \tag{10}$$

The method is evaluated by using SNRs of 5, 15, and 25 dB. The obtained maximum accuracy values are 85.50%, 88.05%, and 89.53% for 5, 15, and 25 dB, respectively. Results show

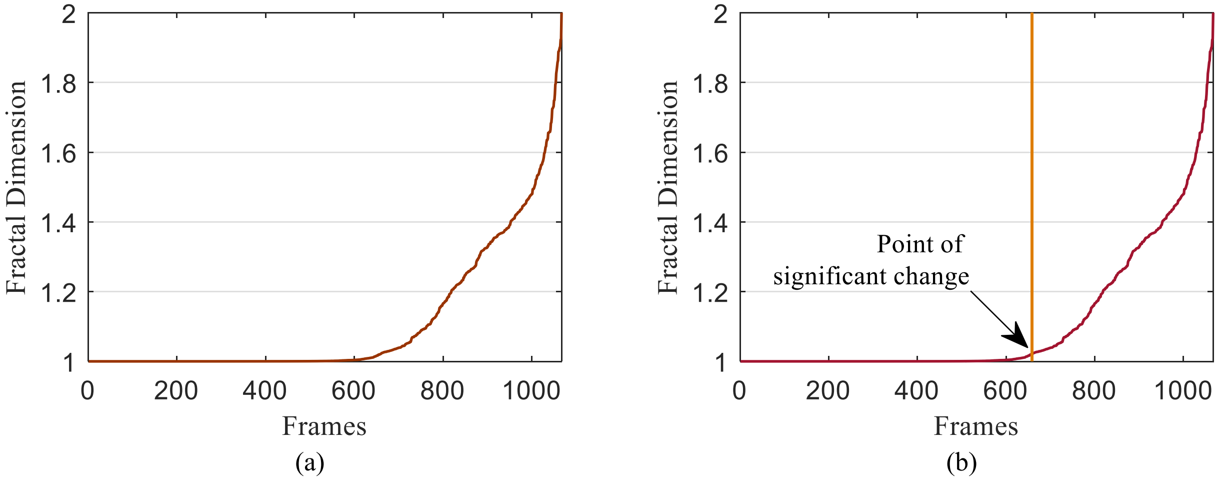

FIGURE 3. (a) Fractal dimension of all frames in an audio (ascending order). (b) The vertical line provides the value of the fractal dimension where the first most significant change appeared.

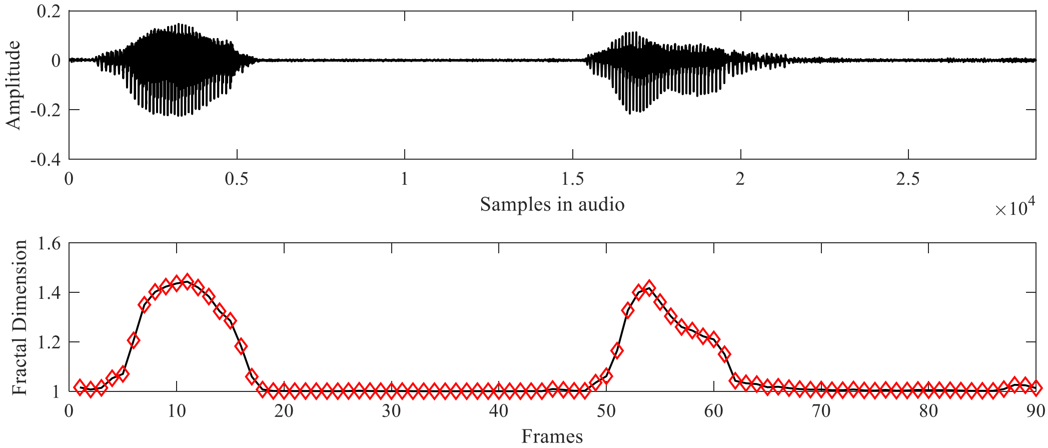

FIGURE 4. Segments of an audio and their estimated fractal dimensions.

TABLE 1. Accuracy for noisy audios in the case of the TIMIT database.

Noise	SNR		
	5 dB	15 dB	25 dB
White	84.91%	87.10%	89.53%
Car	85.15%	88.05%	89.46%
Babble	85.50%	87.95%	89.23%

that the performance of the proposed method in the case of noisy audio is also good.

B. EVALUATION BY USING THE KSU DATABASE

The KSU Arabic Speech Database [39], [40] is also used to evaluate the proposed VAD method because of its diversity in many aspects. The KSU database is recorded in three sessions with a gap of six weeks. In this database, 328 male and female speakers recorded both the pre-written text and the spontaneous text [39]. Conducting the recordings in different environments, office, cafeteria, and soundproof room, is one of the significant aspects of the KSU database. Various recording systems are used to record isolated digits, sentences, paragraphs, and answers to questions. The KSU database is publicly available through the Linguistic Data Consortium, which is hosted by the University of Pennsylvania, Philadelphia, USA [41].

In the TIMIT database, all audios are recorded in the soundproof room, and no significant noise is present in the audio samples. The proposed method perfectly labeled the segments of the audio in the TIMIT database. The audio samples of the KSU database recorded in the soundproof room are also considered to evaluate the performance of the proposed method. Fig. 7 depicts the automatic segmentation of voiced and unvoiced parts of the audio. All segments of the audio are labeled correctly without any error. Although, the language of the KSU database is different from that of the TIMIT database, the performance of the proposed method remains high. This clearly indicates

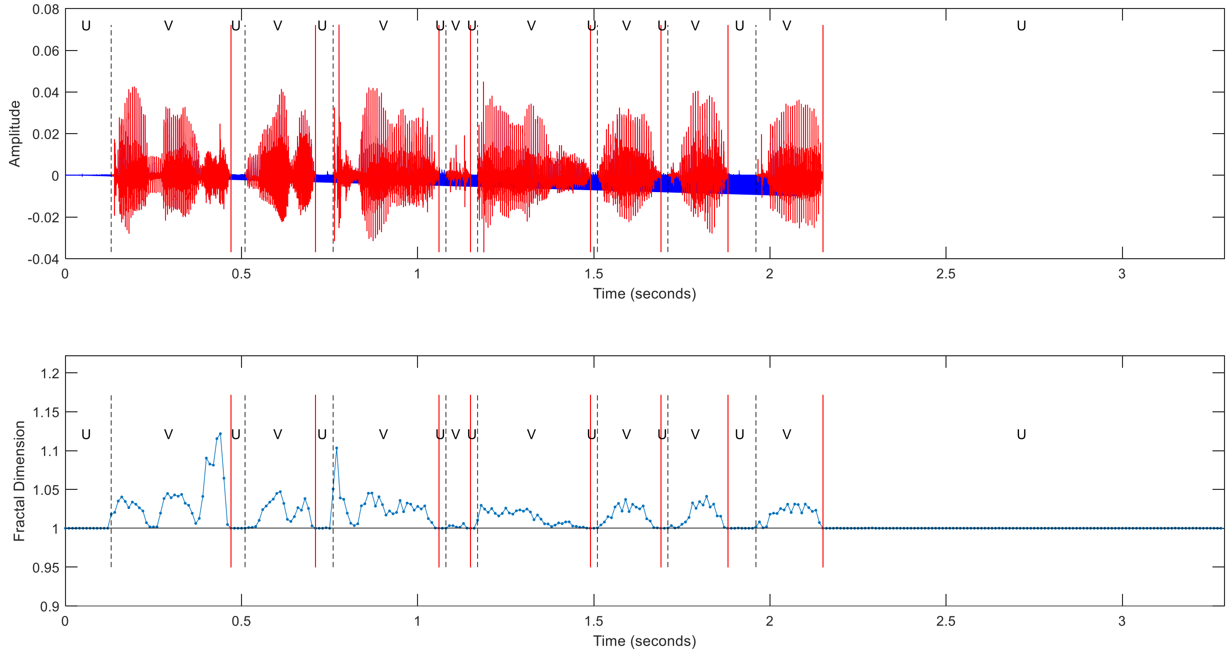

FIGURE 5. Automatic segmentation and fractal dimension of all frames of audio recorded in a soundproof room.

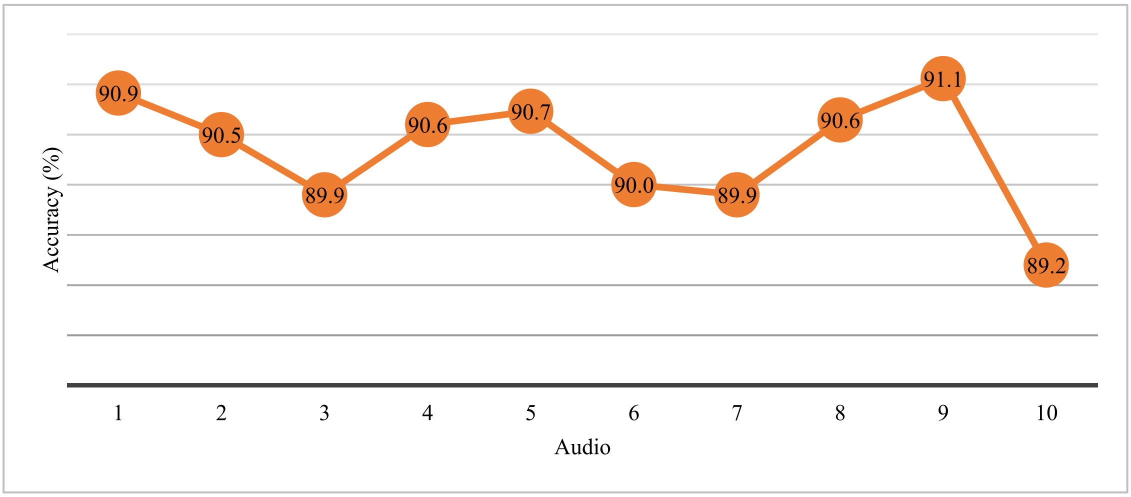

FIGURE 6. Accuracy of the clean audio samples of the TIMIT database.

that the proposed system works equally well for different languages.

While the performance of the proposed method is good for the clear audio, it may not be practical in most application scenarios. Therefore, investigating the robustness of the proposed method against noise is crucial. Robustness can be observed by adding noise of different SNRs in the audios. Then, artificially generated noisy audio can be used for segment detection. In this study, instead of using artificial noisy audio, speech samples of the KSU database recorded in a cafeteria are used to check the robustness of the proposed method. The audios are recorded in the cafeteria in the presence of significant noise. One of the reasons for using the KSU database is to record in different environments. In Fig. 8, the audio recorded in the cafeteria is labeled by using the proposed method. The text of the audio is the same in Fig. 7; the only difference is in the recording environment.

The labeling of voiced and unvoiced segments of the noisy audio is also carried out with perfection. This outcome suggests that the presence of noise in an audio does not affect the performance of the proposed method. In Fig. 8, the recording system consists of a professional Yamaha mixture, which suppresses the noise and avoids audio degradation. Further investigation of the method is conducted by considering the audios in the cafeteria, but the recording system contains normal- and medium-quality sound cards.

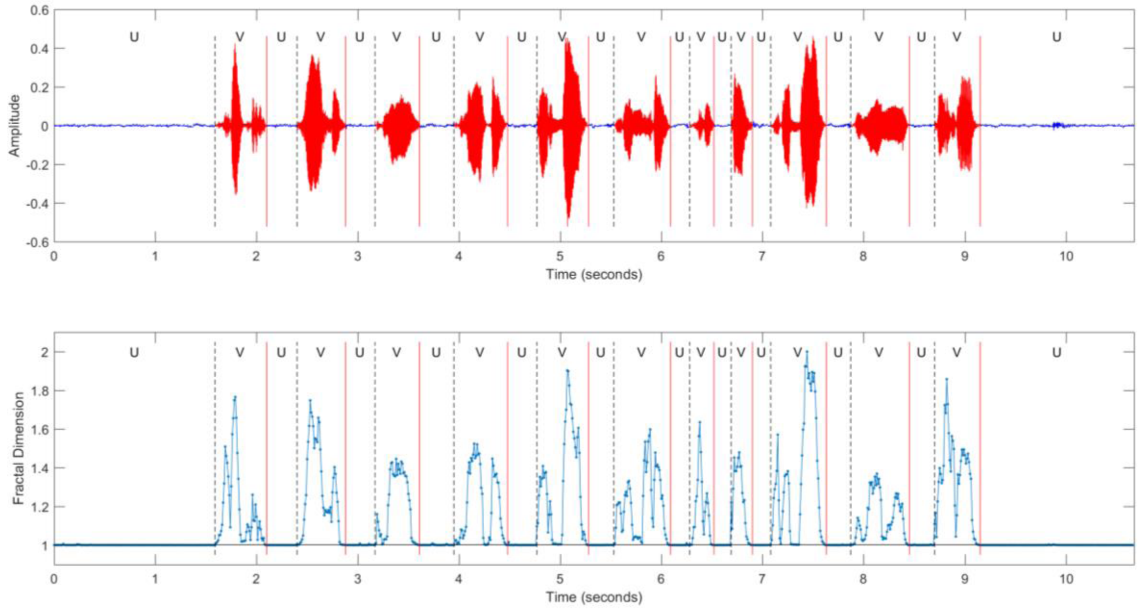

FIGURE 7. Automatic segmentation and fractal dimension of all frames of a clean audio recorded in a soundproof room (quiet environment) by using a professional sound mixer (Yamaha MW-12CX).

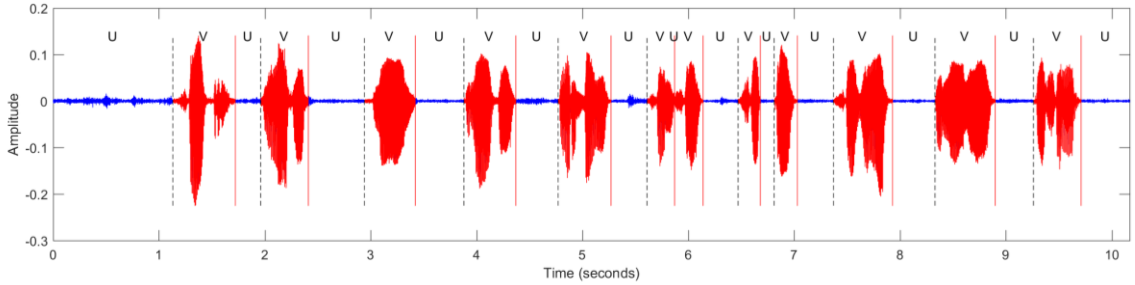

FIGURE 8. Automatic segmentation of the audio recorded in the cafeteria (noisy environment) by using a professional sound mixer (Yamaha MW-12CX).

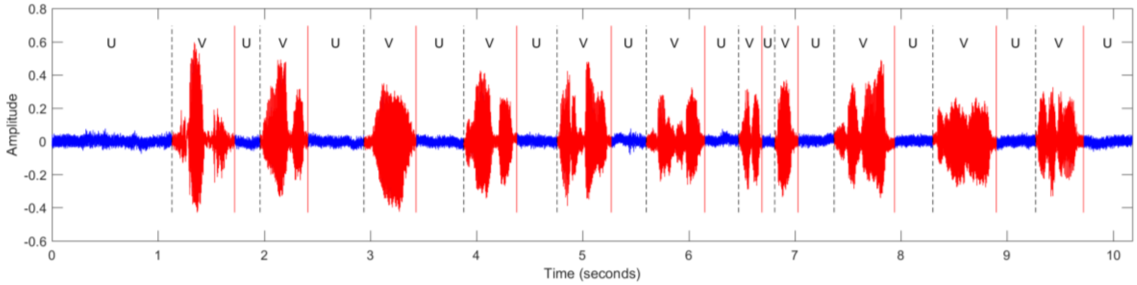

FIGURE 9. Automatic segmentation of very noisy audio recorded in the cafeteria by using a normal quality sound card (built-in sound card of Dell OptiPlex 760).

In Fig. 9, the built-in sound card of a desktop computer is used for recording. Therefore, significant noise can be noted compared to the audio shown in Fig. 8. Despite this substantial noise, the proposed system performs well and labels the voiced and unvoiced segment accurately. The audio shown in Fig. 10 is recorded by using a medium-quality sound card. The proposed system also performed well.

The accuracy of the proposed method for some audio of the KSU database is shown in Fig. 11. The audios recorded in the soundproof room and contained digits are used to label

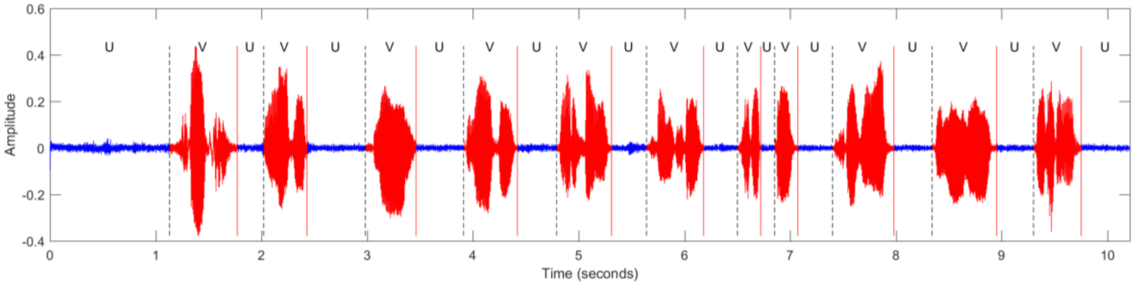

FIGURE 10. Automatic segmentation of noisy audio recorded in the cafeteria by using a medium-quality external sound card (Sound Blaster X-Fi Surround 5.1 Pro).

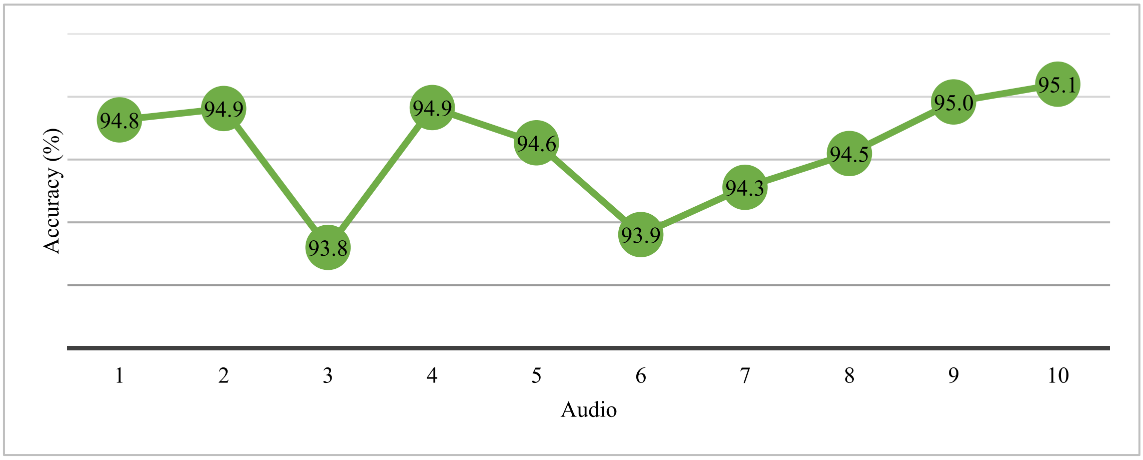

FIGURE 11. Accuracy of the clean audio samples of the KSU database.

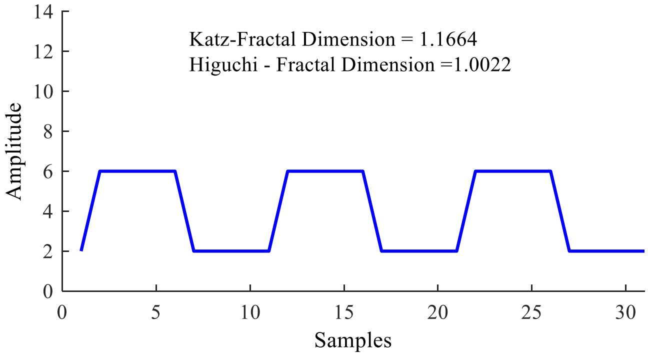

FIGURE 12. Fractal dimensions of a waveform with maximum amplitude equal to 6.

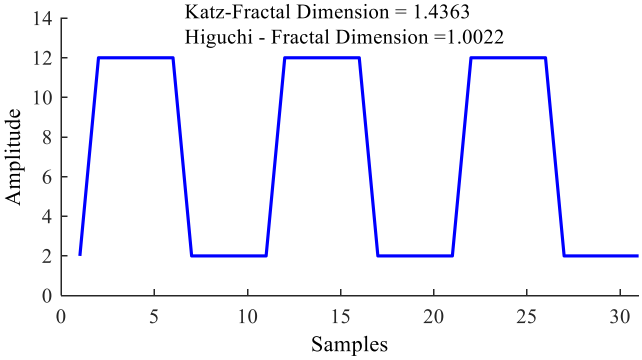

FIGURE 13. Fractal dimensions of a waveform with maximum amplitude equal to 12.

the speech. The average accuracy of 50 randomly selected clean audio is 95.4%. Similar to the TIMIT database, different types of noise are added in the audios of the KSU database to observe the robustness of the proposed method against noise. The results of the proposed method for the noisy audios are provided in Table 2.

The noisy audios are generated by adding white, vehicle, and babble noise in the clean audios. The noise of different SNRs is added. The values of maximum obtained accuracy for 5, 15, and 25dB are 88.91%, 91.12%, and 94.72%, respectively. The obtained results show the robustness of the proposed method against noise.

TABLE 2. Accuracy for noisy audios in the case of the KSU database.

Noise	SNR		
	5 dB	15 dB	25 dB%
White	88.46%	91.12%	94.27%
Car	88.91%	90.38%	94.53%
Babble	88.21%	90.16%	94.72%

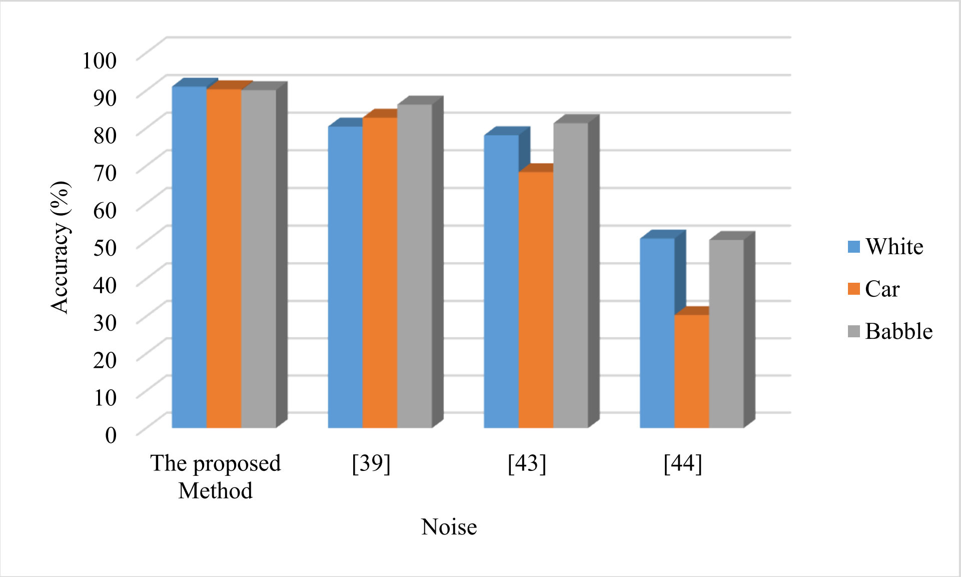

FIGURE 14. Comparison of the proposed methods with existing methods for noisy audios (15 dB).

IV. DISCUSSION

The proposed method classifies speech-presence and speech-absence segments of an audio based on the computed features. The features are the fractal dimension of segments, which determine the category of the segments. Fractal dimension measures the complexity of a waveform. The waveform with the higher amplitude has a greater fractal dimension compared to a waveform exhibiting the lower amplitude. Many algorithms exist to measure the fractal dimension of a waveform [12]–[16]. The success of a developed method depends on the right choice of the fractal estimation algorithm.

The purpose of the developed method is to detect speech-presence and speech-absence segments so they can be used in various speech-related applications. The speech-presence segments have higher amplitude compared to speech-absence segments. Therefore, the difference in the amplitude can be captured to label segments. Two algorithms, Katz algorithm [12] and Higuchi algorithm [13], are analyzed to capture the amplitude difference. For this purpose, two synthetic waveforms are generated with different amplitudes but the same duration, as shown in Figs. 12 and 13.

In Fig. 12, the maximum amplitude of the generated waveform is six and the computed fractal dimensions by using the Katz and Higuchi algorithms are 1.664 and 1.0022, respectively. The fractal dimension of the 2-D waveform lies between 1 and 2, where the fractal dimension of 1 represents a straight line and no variation occurs in the amplitude. In Fig. 12, the fractal dimension by using Higuchi algorithm is 1.0022, which is very close to 1. The reason is that Higuchi algorithm generates new waveforms from a given waveform with different delay factors and starting points. The delay factor makes the waveform smoother. Therefore, Higuchi algorithm is not a good option to observe the variations in amplitude.

Fig. 13 illustrates that after an increase in amplitude from 6 to 12, a significant change occurs in the fractal dimension of the Katz algorithm from 1.1664 to 1.4363. This finding indicates that the response of the Katz algorithm for variation in amplitude is good. On the other hand, the fractal dimension of the Highuchi algorithm is unchanged despite a clear variation in the amplitude. This indicates that the Katz algorithm is more sensitive to the amplitude and is the reason it is employed in the proposed methods to differentiate between the speech-presence and speech-absence segments of an audio. A comparison of the proposed method with existing studies is depicted in Fig. 14.

The results of the proposed method are taken from Table 2. All accuracies in Fig. 14 are for the moderate noise (15 dB). The accuracies of the methods proposed in [42] and [43] are taken from [38]. Fig. 14 shows that the proposed method outperforms the existing methods of VAD.

V. CONCLUSION

A method to detect the speech-presence and speech-absence segments of an audio is presented. The proposed method is unsupervised and does not need any training data to differentiate between voiced and unvoiced segments, a feature that is a positive aspect of the method. Two databases are used to evaluate the performance of the proposed method. The performance evaluation of the method suggests that it labels the segments accurately even for different languages. The presence of significant noise in an audio does not affect the performance of the proposed VAD method. This method can be used reliably to automatically generate forged audio where the detection of boundary points is very crucial and vital. Furthermore, the method can be implemented in various applications related to continuous speech recognition.

REFERENCES

[1] T.-J. Park and J.-H. Chang, "Dempster-Shafer theory for enhanced statistical model-based voice activity detection," *Comput. Speech Lang.*, vol. 47, pp. 47–58, Jan. 2018.

[2] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[3] J. Ramirez, J. C. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, no. 10, pp. 689–692, Oct. 2005.

[4] J. Wu and X.-L. Zhang, "Maximum margin clustering based statistical VAD with multiple observation compound feature," *IEEE Signal Process. Lett.*, vol. 18, no. 5, pp. 283–286, May 2011.

[5] Y. Suh and H. Kim, "Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection," *IEEE Signal Process. Lett.*, vol. 19, no. 8, pp. 507–510, Aug. 2012.

[6] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

[7] S. Deng and J. Han, "Likelihood ratio sign test for voice activity detection," *IET Signal Process.*, vol. 6, no. 4, pp. 306–312, Jun. 2012.

[8] I. Hwang and J. H. Chang, "Voice activity detection based on statistical model employing deep neural network," in *Proc. 20th Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, 2014, pp. 582–585.

[9] M. Imran, Z. Ali, S. T. Bakhsh, and S. Akram, "Blind detection of copy-move forgery in digital audio forensics," *IEEE Access*, vol. 5, pp. 12843–12855, 2017.

[10] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 197–200, Mar. 2013.

[11] D. Cournapeau and K. Tatsuya, "Using variational bayes free energy for unsupervised voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 4429–4432.

[12] M. J. Katz, "Fractals and the analysis of waveforms," *Comput. Biol. Med.*, vol. 18, no. 3, pp. 145–156, 1988.

[13] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Phys. D, Nonlinear Phenom.*, vol. 31, no. 2, pp. 277–283, 1988.

[14] A. Petrosian, "Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns," in *Proc. 8th IEEE Symp. Comput.-Based Med. Syst.*, Jun. 1995, pp. 212–217.

[15] P. Maragos, "Fractal aspects of speech signals: Dimension and interpolation," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1. 1991, pp. 417–420.

[16] 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," *Electron. Lett.*, vol. 28, no. 4, pp. 420–422, Feb. 1992.

[17] R. Lopes and N. Betrouni, "Fractal and multifractal analysis: A review," *Med. Image Anal.*, vol. 13, no. 4, pp. 634–649, 2009.

[18] Y. W. Kim, K. K. Krieble, C. B. Kim, J. Reed, and A. D. Rae-Grant, "Differentiation of alpha coma from awake alpha by nonlinear dynamics of electroencephalography," *Electroencephalogr. Clinical Neurophysiol.*, vol. 98, no. 1, pp. 35–41, 1996.

[19] A. K. Mishra and S. Raghav, "Local fractal dimension based ECG arrhythmia classification," *Biomed. Signal Process. Control*, vol. 5, no. 2, pp. 114–123, 2010.

[20] P. N. Baljekar and H. A. Patil, "A comparison of waveform fractal dimension techniques for voice pathology classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 4461–4464.

[21] A. Accardo, M. Affinito, M. Carrozzi, and F. Bouquet, "Use of the fractal dimension for the analysis of electroencephalographic time series," *Biological*, vol. 77, no. 5, pp. 339–350, 1997.

[22] R. Esteller, G. Vachtsevanos, J. Echauz, and B. Litt, "A comparison of waveform fractal dimension algorithms," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 48, no. 2, pp. 177–183, Feb. 2001.

[23] B. S. Raghavendra and D. N. Dutt, "A note on fractal dimensions of biomedical waveforms," *Comput. Biol. Med.*, vol. 39, no. 11, pp. 1006–1012, 2009.

[24] Q. Yan, R. Yang, and J. Huang, "Copy-move detection of audio recording with pitch similarity," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 1782–1786.

[25] X. Pan, X. Zhang, and S. Lyu, "Detecting splicing in digital audios using local noise level estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2012, pp. 1841–1844.

[26] A. J. Cooper, "Detecting butt-spliced edits in forensic digital audio recordings," in *Proc. 39th Int. Conf., Audio Forensics, Pract. Challenges*, 2010, p. 1.

[27] Z. Ali, M. Imran, and M. Alsulaiman, "An automatic digital audio authentication/forensics system," *IEEE Access*, vol. 5, pp. 2994–3007, 2017.

[28] M. Imran, Z. Ali, S. T. Bakhsh, and S. Akram, "Blind Detection of Copy-Move Forgery in Digital Audio Forensics," *IEEE Access*, vol. 5, pp. 12843–12855, 2017.

[29] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *J. Amer. Stat. Assoc.*, vol. 107, no. 500, pp. 1590–1598, 2012. [Online]. Available: http://arxiv.org/abs/1101.1438

[30] M. Lavielle, "Using penalized contrasts for the change-point problem," *Signal Process.*, vol. 85, no. 8, pp. 1501–1510, 2005.

[31] J. S. Garofolo *et al.*, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: The Linguistic Data Consortium, 1993.

[32] Y. Liang, X. Liu, Y. Lou, and B. Shan, "An improved noise-robust voice activity detector based on hidden semi-Markov models," *Pattern Recognit. Lett.*, vol. 32, no. 7, pp. 1044–1053, 2011.

[33] G. Friedland, O. Vinyals, Y. Huang, and C. Muller, "Prosodic and other long-term features for speaker diarization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 5, pp. 985–993, Jul. 2009.

[34] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 468–477, Mar. 2006.

[35] Y. Wang and L. Lee, "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 3, pp. 564–579, Mar. 2015.

[36] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Comput. Speech Lang.*, vol. 14, no. 4, pp. 333–353, 2000.

[37] M. Alhussein, Z. Ali, M. Imran, and W. Abdul, "Automatic gender detection based on characteristics of vocal folds for mobile healthcare system," *Mobile Inf. Syst.*, vol. 2016, 2016, Art. no. 7805217. [Online]. Available: https://www.hindawi.com/journals/misy/2016/7805217/

[38] B. Xulei, Z. Jie, and C. Ning, "A robust voice activity detection method based on speech enhancement," in *Proc. IET Intell. Signal Process. Conf. (ISP)*, 2013, pp. 1–4.

[39] M. Alsulaiman, Z. Ali, G. Muhammed, M. Bencherif, and A. Mahmood, "KSU speech database: Text selection, recording and verification," in *Proc. Eur. Modelling Symp.*, 2013, pp. 237–242.

[40] M. M. Alsulaiman, G. Muhammd, M. A. Bencherif, A. Mahmood, and Z. Ali, "KSU rich Arabic speech database," *J. Inf.*, vol. 16, no. 6, pp. 4231–4253, 2013.

[41] M. Alsulaiman, G. Muhammad, B. Abdelkader, A. Mahmood, and Z. Ali, *King Saud University Arabic Speech Database*. Philadelphia, PA, USA: The Linguistic Data Consortium, 2014.

[42] L. N. Tan, B. J. Borgstrom, and A. Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4466–4469.

[43] Q. H. Jo, J. H. Chang, J. W. Shin, and N. S. Kim, "Statistical model-based voice activity detection using support vector machine," *IET Signal Process.*, vol. 3, no. 3, pp. 205–210, May 2009.

ZULFIQAR ALI received the M.Sc. degree in computational mathematics from the University of the Punjab, Lahore, in 2001, the M.Sc. and M.S. degrees in computer science from the University of Engineering and Technology, Lahore, in 2007 and 2010, respectively, and the Ph.D. degree in electrical and electronic engineering from the Center for Intelligent Signal and Imaging Research, Universiti Teknologi PETRONAS, Malaysia, in 2017. He has been a Researcher with the Department of Computer Engineering, King Saud University, Saudi Arabia, since 2010. His research and teaching career span over 16 years. He has over 50 publications in journals and conferences of international repute, and has a very active research profile. He has the leading roles in many funded projects. His research interests include speech and language processing, cloud and multimedia for healthcare, privacy and security, watermarking, and multimedia forensics.

MUHAMMAD TALHA received the Ph.D. degree in computer science from the Faculty of Computing, University of Technology, Malaysia. He is currently involved in the Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia. His research interests include image processing, medical imaging, features extraction, classification, and machine learning techniques.