INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Computer Applications in Technology

ISSN online: 1741-5047 - ISSN print: 0952-8091
https://www.inderscience.com/ijcat

Unsupervised VAD method based on short-time energy and spectral centroid in Arabic speech case

Hind Ait Mait, Noureddine Aboutabit

DOI: 10.1504/IJCAT.2023.10061438

Article History:

Received:	13 December 2022
Last revised:	26 February 2023
Accepted:	08 August 2023
Published online:	03 October 2024

Unsupervised VAD method based on short-time energy and spectral centroid in Arabic speech case

Hind Ait Mait* and Noureddine Aboutabit

Laboratory LIPIM,
National School of Applied Sciences Khouribga,
Sultan Moulay Slimane University, Morocco
Email: hind.ait-mait@usms.ac.ma
Email: n.aboutabit@usms.ma
*Corresponding author

Abstract: Voice Activity Detection (VAD) distinguishes speech segments from noise or silence areas. An efficient and noise-robust VAD system can be widely used for emerging speech technologies such as wireless communication and speech recognition. In this paper, we propose two versions of an unsupervised Arabic VAD method based on the combination of the Short-Time Energy (STE) and the Spectral Centroid (SC) features for formulating a typical threshold to detect speech areas. The first version compares only the STE feature to the threshold (STE-VAD). In contrast, the second compares the SC vector and the threshold (SC-VAD). The two versions of our VAD method were tested on 770 sentences of the Arabphone corpus, which were recorded in clean and noisy environments and evaluated under different values of Signal-to-Noise-Ratio. The experiments demonstrated the robustness of the STE-VAD in terms of accuracy and Mean Square Error.

Keywords: unsupervised VAD; short-time energy; spectral centroid; Arabic speech; computer applications.

Reference to this paper should be made as follows: Mait, H.A. and Aboutabit, N. (2024) 'Unsupervised VAD method based on short-time energy and spectral centroid in Arabic speech case', *Int. J. Computer Applications in Technology*, Vol. 74, No. 3, pp.158–170.

Biographical notes: Hind Ait Mait is a PhD student at the National School of Applied Sciences in Khouribga (Sultan Moulay Slimane University, Morocco). She received her Master's degree in Big Data and Help of Decisions in 2020 from the same school. Her main research interest includes speech processing, big data, machine learning and deep learning.

Noureddine Aboutabit is an Associate Professor in Telecommunications and Multimedia Processing at the National School of Applied Sciences, Khouribga since October 2011. In 2007, he received his PhD degree in Signal Image Speech Telecom from Grenoble INP (France). He received his MS degree in 2004 from the same institute. In 2003, he obtained his engineering diploma from Ecole Normale Suprieure d'Ingnieurs Electriciens de Grenoble (ENSIEG). His current research interests include computer vision, machine learning, artificial intelligence, speech processing, Big Data and cloud computing security.

This paper is a revised and expanded version of a paper entitled 'An Unsupervised Voice Activity Detection Using Time-Frequency Features' presented at 'International Conference on Machine Intelligence and Computer Science Applications (ICMICSA'2022)'.

1 Introduction

Speech is the essential human means of communication. It is the articulate vocal sound expression of ideas and thoughts. In the last three decades, the necessity for alternate methods of interacting with computer systems has motivated speech-processing researchers worldwide to extract relevant information from the speech signal efficiently and robustly. Among the speech processing applications, there are Automatic Speech Recognition (ASR), Speech Synthesis, Speech Coding, Speaker Identification and Speech Transmission (Bäckström, 2017). Detecting speech from an audio stream is a critical step that directly influences the performance of these systems. For instance, too many false alarms or non-speech segments misidentified as speech and used in training may taint the acoustic models and lower their accuracy. On the other side, the ASR algorithms will be able to recognise the whole spoken sentence if more speech segments are identified during testing. To deal with this problem, we employ Voice Activity Detection (VAD), which involves solving a binary classification task to distinguish speech segments from ambient silence or noise (Zhang and Wu, 2012). According to the nature of excitation of the vocal cords, speech can be divided into two groups: voiced and unvoiced. In the first type, the airflow from the lungs vibrates the vocal cords, while in the unvoiced speech, there is no use

of the vocal cords (Silva et al., 2017). A typical VAD system comprises two parts: The features extraction and a speech/ non-speech decision mechanism. The first part aims to transform the speech waveform into a parametric representation, which will be used as input to the decision model. Most VAD features proposed in the literature take advantage of the discriminative characteristics of speech in various domains, which can be divided into five categories: energy-based features, spectral-domain features, cepstral-domain features, harmonicity-based features and long-term features. The subsequent component of VAD involves establishing the rule or technique used to assign a class (either speech or non-speech) to the input feature vector. Since background noise interferes with the classifier's performance, the classification problem is frequently more complicated than it first appears. Accuracy, reliability, robustness, latency and memory requirements are essential characteristics of every VAD (Bäckström, 2017). Among these properties, robustness against noisy environments has been the most challenging task. In high Signal-to-Noise Ratio (SNR) conditions, the simplest VAD algorithms can perform satisfactorily, while in low SNR environments, all VAD algorithms degrade to a certain extent. At the same time, the VAD algorithm should be low complexity, which is necessary for real-time systems. VAD approaches can be divided into supervised, semi-supervised and unsupervised methods (Sadjadi and Hansen, 2013). The supervised one's process VAD as a traditional classification problem, which they solve either by directly training a classifier or by separately building statistical models for speech and non-speech and then making VAD judgments. While the semi-supervised methods use the features vector extracted as an input of a classifier for feeding the model and taking the final decision. The last type is the unsupervised method which involves metrics-based methods that depend on continuous observation of a specific metric, such as energy or zero-crossing rate, followed by a simple threshold-based decision stage. In our research, we are motivated by an unsupervised VAD approach as presented in Giannakopoulos (2009). This method utilises two distinct features: Short Time Energy (STE) and Spectral Centroid (SC), which are derived from the time and frequency domains, respectively. The method employs dual thresholds to effectively discriminate between the presence and absence of speech events. Our paper presents a VAD system that combines STE and SC features to determine a single threshold to detect speech segments. This approach has two versions: STE-VAD, which only compares the energy values and the threshold, and STE-SC introduces a VAD system based on comparing the SC vector with the criterion thresholding. The two versions were evaluated using the Arabphone corpus (Frihia and Bahi, 2016), which was recorded in noisy and noiseless environments. The paper is organised as follows: Section 2 introduces state of the art; Section 3 describes the VAD method (Giannakopoulos, 2009), its drawbacks and our proposed method. Section 4 presents the speech corpus used in addition to the accuracy and the Mean Square Error (MSE) as evaluation metrics and discusses the experimental results. Section 5 summarises the paper and presents the future work.

2 Related work

Voice activity detection has many techniques and approaches. This section introduces the supervised and unsupervised methods used in the literature to separate between voiced and unvoiced speech.

2.1 Supervised methods

Supervised learning approaches have recently been used more because they offer the potential to overcome the constraints of statistical model-based methods. In this section, we will introduce numerous techniques that tackle the challenge of VAD through the lens of machine and deep learning. The primary objective is to categorise segments into classes of speech and non-speech. The MFCC features and the Support Vector Machine (SVM) were proposed to detect the speech segments in Kinnunen et al. (2007). According to the experiments, the proposed method works excellently, and the SVM is more straightforward to adapt to new data sets than the traditional approach. Convolution Neural Network (CNN) based model along with a Denoising Autoencoder (DAE) was presented in Shin et al. (2010). The test was done against acoustic features and their delta at noise levels ranging from SNR 35 to 0 dB. The results demonstrated that adding more expressive audio features with DAE improves accuracy, especially at noise levels. The suggested model has achieved good accuracy. In the work (Eyben et al., 2013), the approach suggested for realising a VAD system has based on Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM). The method could pattern the long-range dependencies between the inputs. The results demonstrated that LSTM-RNN outperforms the statistical VAD baselines on real-life noisy speech data from Hollywood movies. A powerful hierarchical generative model called a Deep Belief Network (DBN) was suggested in Zhang and Wu (2012) to combine the benefits of several acoustic features in a linear way for extracting a new feature. The experiments carried out on the AURORA2 corpus demonstrated that DBN-based VAD outperforms 11 referenced algorithms with low-detection complexity. Hughes and Mierle (2013) introduced a Recurrent Neural Network (RNN) architecture for detecting speech. It had nodes that computed the quadratics polynomials. The chosen model exceeds the Gaussian Mixture Models (GMMs) and a hand-tuned State Machine (SM) by a 26% reduction in False Alarms Rate (FAR). A VAD algorithm based on CNN was presented in Silva et al. (2017). It detected voice frames using the audio spectrogram raw image in a specific audio source. The method was compared with five baseline systems evaluated in Dean et al. (2010): ETSI, G729B, Sohn, LTSD, GMM-MFCC-1. The results demonstrated that CNN outperformed the introduced algorithm regarding Half-Total Error Rate (HTER). Bai et al. (2017) presented a new VAD algorithm based on Deep Neural Networks (DNN) and Viterbi. The miss rate and FAR were used to evaluate the performance of the proposed method. The obtained values showed its effectiveness and its flexibility in real-time. Sehgal and Kehtarnavaz (2018) described a smartphone application that uses a CNN for detecting speech in real time. The acquired results exhibited that CNN outperforms the previously

developed random forest applications. Deep architecture based on an RNN was proposed Ariav et al. (2018), and it has been trained to make a VAD system. According to the experimental results, the suggested architecture exceeded state-of-the-art detectors in terms of accuracy, even in low SNR conditions and complicated types of transients. Arslan and Engin (2019), a VAD method has been proposed using various features of the time and the spectral domains. The first one includes STE and ZCR. The second has entropy, centroid, roll-off and flux of speech signals as a feature. Multi-Layer Feed-Forward Neural Network was chosen as a classifier to separate between speech and non-speech segments. The algorithm was tested for six different noises with four levels of SNR. The suggested technique was compared with G.729B (Benyassine et al., 1997) and Long-Term Spectral Flatness Measure (Ma and Nishihara, 2013) in terms of Correct Speech Rate, FAR and Overall Accuracy Rate. This evaluation exhibited its efficiency. Zhang and Xu (2022) presented a VAD method based on a DNN for maximising Area Under the Curve (AUC) to increase the performance of DNN-based VAD at various threshold settings. The test was applied on different SNR levels in babbling and factoring noise scenarios. The experiments revealed that using DNN to optimise AUC outperforms the typical methods of using DNN to optimize the Minimum Squared error. A lightweight CNN architecture for real-time voice activity detection has been nominated in Alam and Khan (2020). The trained model was evaluated in a noisy environment, and the experiments showed that the model was potent. Furthermore, the data augmentation and regularisation techniques provided good results. Rho et al. (2022) proposed a Neural Architecture Search (NAS) with search space and macro-structure optimised for the VAD problem, which might be applied to build a network structure that automatically enhances detection precision. The outcomes of the research showcased the superiority of the proposed NAS framework over manually designed state-of-the-art VAD models across diverse real-world data sets augmented with noise.

2.2 Unsupervised methods

In this section, we will present the methods which separate between active and inactive speech founded only on the internal characteristics of the signal and which do not require any prior knowledge. Those are the unsupervised methods. In the literature, only some researchers are interested in detecting Arabic speech. A noise-robust Voice Activity Detection system was suggested in Ali and Talha (2018) to label the speech presence and absence segments in the signal. The use of long-term features implements it. The Texas Instruments Massachusetts Institute of Technology (experiments were carried) and the King Saud University (KSU) Arabic speech databases were used to evaluate the method performance. The results revealed that it accurately classified the voiced and unvoiced segments in clean and noisy environments. The Wavelet Packet Transform (WPT) method presented in Ghanbari and Karami-Mollaei (2006) begins by applying a wavelet transform to the signal, resulting in sub-band decomposition using WPT coefficients. The voice inside the signal is then determined by comparing the sub-band energy of components between detail and approximation coefficients. Moattar and Homayounpour (2009) proposed a Voice Activity Detection (VAD) method that was designed to be robust in noisy environments. The proposed method relied on short-term features such as Short-term Energy (STE), Spectral Flatness Measure (SFM), and the most dominant frequency component of the speech frame spectrum, represented by F. The authors tested their approach on four different data sets in various noise settings with different Signal-to-Noise Ratio (SNR) levels. The first data set used was experiments were carried, followed by Farsdat, a microphone voice corpus, TPersianDat, a Farsi telephony speech corpus and Aurora2 Speech Corpora. They compared their proposed algorithm to several previously stated methods, and the results showed that their method outperformed the others in terms of VAD performance. The Long-term Spectral Flatness Measure (LSFM) is used in Ma and Nishihara (2013). Twelve (12) different types of noise were used in the experiments on the TIMIT corpus, with five different SNRs ranging from –10 to 10 dB. The Accuracy and the Error Rate (ER) were the two metrics utilised to assess the performance. Yoo et al. (2015) presented a robust formant-based VAD approach to handle the problem of detecting formants in noisy situations. It outperformed typical VAD algorithms under various noise conditions and had a far faster processing time. At SNR of 0, 5, 10 and 15 db. A VAD approach based on the Power Spectral Deviation of Teager Energy was proposed in Kim et al. (2016) to discriminate between speech and non-speech in various noisy situations (babble, office, automobile). The evaluation is performed by ER, False Acceptance Rate (FAR) and False Rejection Rate (FRR). It showed better accuracy than the traditional methods. In Zaw and War (2017), a combination of ZCR, Spectral Entropy, STE and Linear Prediction Error (LPE) was used. The results revealed that the approach could more precisely recognise the endpoints of voice signals. To detect speech regions, a novel thresholding approach based on a modified global threshold was proposed in Elton et al. (2022). It greatly enhanced the overall VAD performance. Several experiments showed that the introduced method could detect active human speech in low-SNR and diverse noisy conditions. Moreover, it handled signals with non-stationary noises, which can include a variety of complicated occurrences that are a mix of different noises. An effective VAD algorithm was introduced in Çolak and Akdenniz (2021). It is based on three short-time features: Short Time Energy, Periodicity and Spectral Flatness. The data set used in testing was created by combining several types of noise (white, vehicle, airport) with various SNR levels. This VAD approach produced the most outstanding results with white noise and can be used in adaptive filter applications.

3 VAD based on spectral centroid and short time energy

The VAD method (Giannakopoulos, 2009) is based on the Short-Term Energy (STE) and the Spectral Centroid (SC). A speech signal is a non-stationary signal that varies over time. Therefore, it was processed by dividing it into frames of 50 ms.

Then, the values of these features are calculated from each frame. After that, their histograms were computed to identify the first and local maximal for formulating the two thresholds (T_e, T_{sc}). The main idea is to compare the feature values with T_e and T_{sc}. The regions which contain values that surpassed both thresholds were determined as speech segments in equation (3). The two thresholds were calculated using the formula below:

$$T_e = \frac{W.M_{1e} + M_{2e}}{W+1} \tag{1}$$

$$T_{sc} = \frac{W.M_{1sc} + M_{2sc}}{W+1} \tag{2}$$

$$STE > T_e SC > T_{sc} \tag{3}$$

where W is a user-defined parameter.

M_{1e}, M_{2e}: the first and second local maxima of the STE histogram, respectively.

M_{1sc}, M_{2sc}: are the positions of the first and second local maxima of the SC histogram.

3.1 Spectral centroid

The spectral centroid (Schubert and Wolfe, 2006) is linked to a sound's brightness measurement. The frequency and magnitude data obtained from the Fourier transform were employed to compute the 'centre of gravity'. This includes computing the average frequency, which is weighted by amplitudes and dividing it by the sum of the amplitudes, as shown below:

$$SC = \frac{\sum_{k=1}^{N} KF[K]}{\sum_{k=1}^{N} F[K]}. \tag{4}$$

$F[k]$ is the amplitude in the Discrete Fourier Transform range corresponding to bin k, and N is the frame length.

3.2 Short-time energy

The short-time energy function (Giannakopoulos, 2009) is the energy of the short-speech segment. It is a simple and effective classifying parameter for the voiced and unvoiced parts; its definition is as follows:

$$E(i) = \frac{1}{N}\sum_{n=1}^{N} |x_i(n)|^2 \tag{5}$$

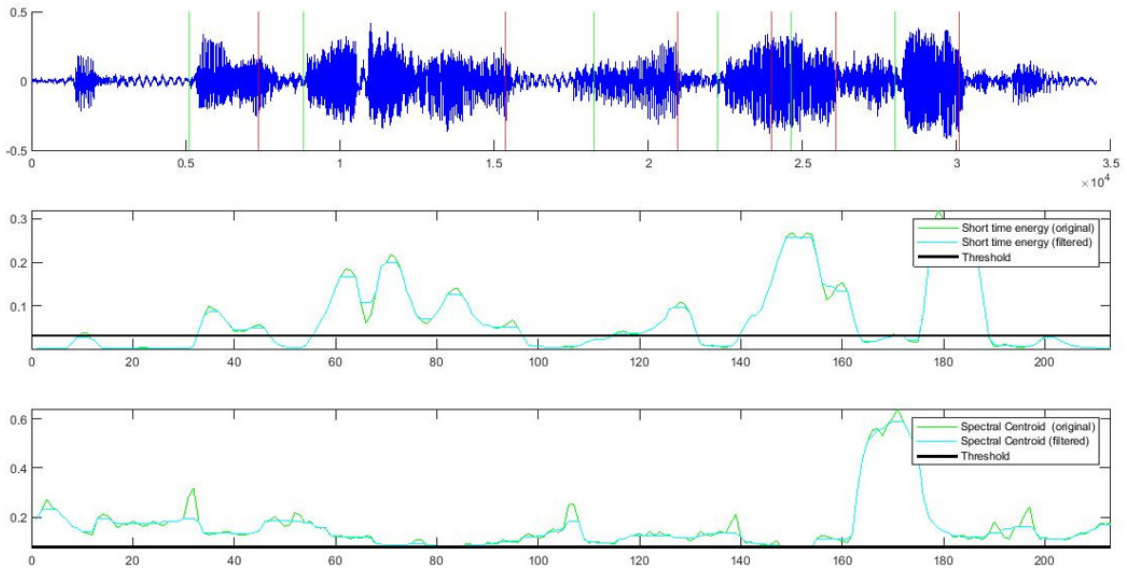
where i is the short-term frame and N is the frame length.

3.3 Method's drawbacks

While implementing the method (Giannakopoulos, 2009), we encountered several errors. This part will discuss two major problems identified in many database sequences. The Figures 1 and 2 present an illustrative example of the method limitations.

Issue 1: Masking effect: As we have previously stated, the threshold was established using the first and the second local maxima of the feature histogram. However, upon analysing the histograms of the Short Time Energy (STE) and Spectral Centroid (SC) features (depicted in Figure 1), it became apparent that the second, third, and fourth peaks of maximum values seemed to dominate over the initial peak. This was particularly evident in the first histogram representing the STE feature. This masking effect was a concern, as it meant that the first local maximum was not always the most accurate representation of the speech or noise/silence threshold. Instead, it was often influenced by other features that had higher peak values in the histogram. This could potentially lead to incorrect classifications of speech or noise/silence areas within the audio data.

Figure 1 Illustrative example of the neighbouring peaks dominance problem (see online version for colours)



To address this problem, one potential solution was to use a combination of the features, rather than relying solely on the first and second local maximum of a single feature. This would allow us to take into account the impact of particular features on the speech threshold, which could obscure our results. By acknowledging this concern and investigating alternative approaches, we were able to enhance the precision of our methodology and effectively differentiate between speech and non-speech regions in audio recordings.

Issue 2: Criterion thresholding: The second issue with the speech segmentation method pertains to the thresholding criterion, where the detection of speech segments is dependent on certain threshold values. This algorithm uses two feature vectors to determine whether a given time interval contains speech or non-speech. In order to classify an interval as containing speech, both feature vectors must have values greater than their respective thresholds (as specified in equation (3)).

However, the algorithm can encounter problems when the values of the two feature vectors do not meet the threshold criteria at the same time. Figure 2 illustrates such a scenario, where the Short-Time Energy (STE) values are higher than T_e in a certain interval (e.g. [100–150]), but the Spectral Centroid (SC) values are lower than T_{sc} in the same interval. As a result, the algorithm misclassifies this region as non-speech.

This issue highlights the importance of careful selection of threshold values for the feature vectors, as well as consideration of the relative importance of each feature in speech classification.

4 Proposed method

To address the problems of the VAD method (Giannakopoulos, 2009), we suggested our method, which is based on the combination of the Short-Time Energy (STE) and the Spectral Centroid (SC), to determine a thresholding criterion. The Figure 3 presents the algorithm architecture.

Figure 2 Illustration of the threshold issue (see online version for colours)

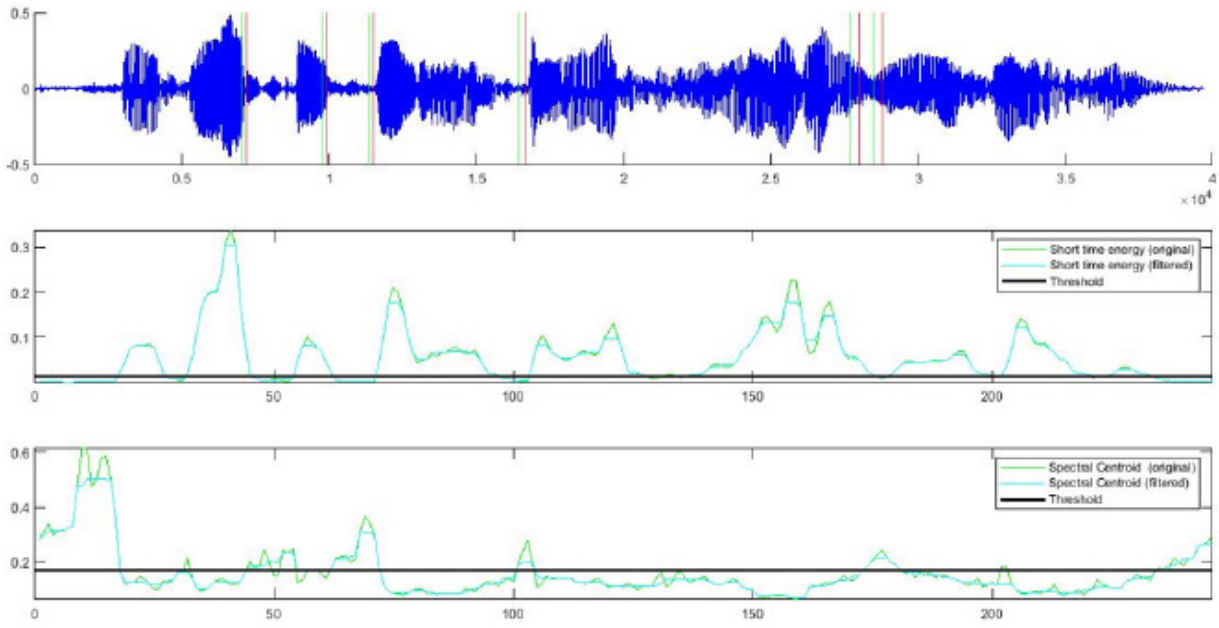
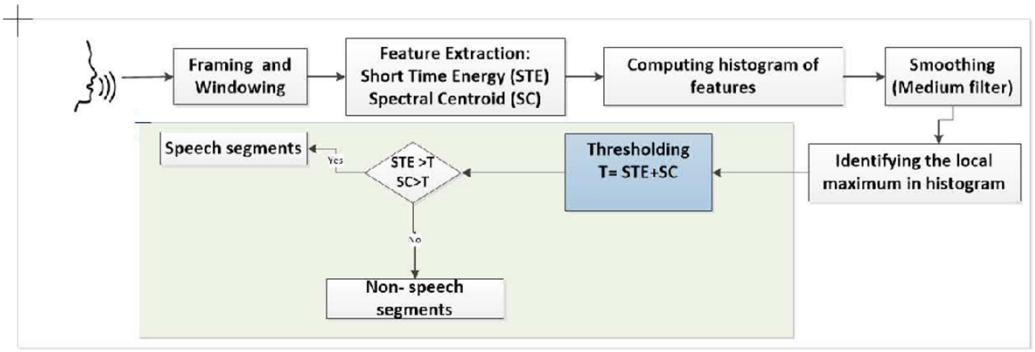


Figure 3 The algorithm architecture



- *Framing and windowing*: Speech signal preprocessing is crucial because it transforms the speech waveform into a parametric representation. It describes the acoustic events in a voice signal using several speech characteristics. Although it is known that the preprocessing step of speech signal contains multiple techniques that could be used, in our work, we applied the two most powerful concepts in this step: the framing and the windowing techniques. Speech is a non-stationary signal; its statistical features do not remain constant across time. As a result, spectral characteristics should be retrieved from tiny signal segments, predicated on the assumption that the signal in this short frame is stationary. For this reason, we divided the acquired signal into short frames of 10 ms, known as the framing process.

In the next step, we applied the windowing process, which is presented as the technique of multiplying the speech signal segment's waveform by a time window to emphasise the signal's predefined characteristics and smooth out the discontinuity at the beginning and the end of the sampled signal. The function chosen for applying this process was the Hamming window, where mathematical formulation can be defined as in equation (6)

$$w[n] = \left\{ cc0.54 - 0.46\cos\left(\frac{2\pi n}{L}\right) 0 <= n <= L0 \qquad \text{otherwise} \right\} \quad (6)$$

where

L is the number of samples in each frame.

n is the number of the frame under processing.

$w[n]$ is the result of the windowing procedure.

- *Features extraction*: The process of converting raw data into numerical features that may be analysed while keeping the information in the original data is known as feature extraction. In this stage, we extracted the characteristics of the signal from the temporal and frequency domains, which are presented as the STE and the SC values. Afterward, we combined these two features into one vector:

$$\left(STE = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, SC = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \Rightarrow STE \oplus SC = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right) \quad (7)$$

- *Histogram computing and smoothing*: In this step, we used the Matlab function: *Histogram (x)*, which chose an adequate number of bins to span the range of values in x and displayed the shape of the fundamental distribution. Its use is to compute the histogram of the vector's values. Afterward, the medium filter, a non-linear filter, was applied to the histogram to remove the high-frequency fluctuations from a signal; this is known as the smoothing technique.

- *Thresholding*: This stage introduced the final phase of our work when we defined a decision threshold to apply a binary classification (speech and non-speech classes). The computed threshold was based on the first and second local maximums of the histogram, as shown in Figure 4. The mathematical representation of the threshold can be described in equation (8).

$$T_{e+sc} = \frac{W.M_{1e+sc} + M_{2e+sc}}{W} \quad (8)$$

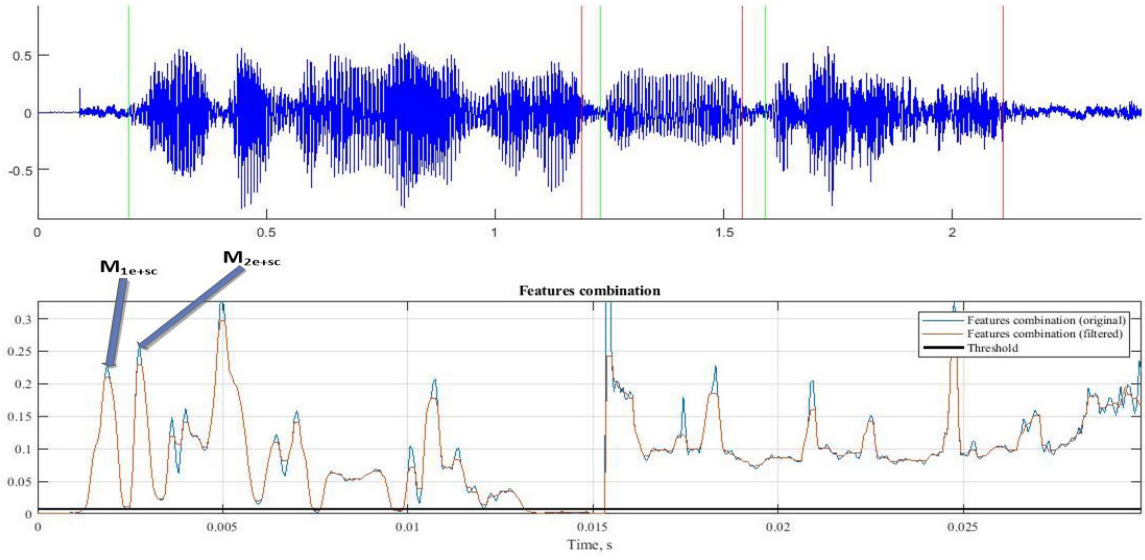
where

W is a user-defined parameter.

M_{1e+sc} : is the position of the first local maxima of the histogram of the combined features.

M_{2e+sc} : is the position of the second local maxima of the histogram of the combined characteristics.

Figure 4 Histogram of the combined vector (see online version for colours)



In order to accurately distinguish between speech and non-speech areas within audio recordings, we implemented two different versions of our method: the STE-VAD and SC-VAD. These two methods were designed to identify non-speech regions as those with values below the T_{e+sc}, while speech regions were defined as those with values above it (equations (9) and (10)). Our method stood in contrast to the approach used in a previous study (Giannakopoulos, 2009), which relied on a double threshold endpoint detection method (as described in equation (3)). By using our method, we were able to achieve more accurate results and reduce the potential for false positives or false negatives. This was particularly important in situations where the audio data was complex or contained background noise that could interfere with the accurate identification of speech and non-speech areas.

$$STE > T_{e+sc} \tag{9}$$

$$SC > T_{e+sc} \tag{10}$$

4.1 Signal-to-noise ratio estimation

The Signal-to-Noise Ratio (SNR) is among the most basic signal-processing metrics. It is defined as the ratio of signal power to noise power expressed in decibels (dB) and indicates the amount of background noise in a speech signal (see equation (11)).

$$SNR = \frac{P_s}{P_n} \tag{11}$$

where P_s is the power of signal and P_n is the background noise.

However, assessing a signal in practice can be challenging due to the variety of forms and ways that it could be corrupted. Furthermore, the inherent fluctuation in the signal provides another level of difficulty to SNR computation. As a result, it is critical to investigate and estimate the impact of noise on the original signal in relevant ways.

In our experiment, we applied our VAD method on the Arabphone corpus (Frihia and Bahi, 2016), which does not contain any information on the SNR. Therefore, we used the WADA-SNR method (Waveform Amplitude Distribution Analysis) (Kim and Stern, 2008) for estimating the SNR of speech signals, which is based on statistical information obtained from the amplitude distribution of a speech waveform. The approach assumes that an additive noise signal is Gaussian and that the Gamma distribution with a shaping value of 0.4 can approximate the amplitude distribution of clean speech.

5 Experimental results

This section describes the database used to evaluate the VAD approach (Giannakopoulos, 2009) and the proposed method. Then, we present the implementation details and compare the two methods in terms of accuracy, Mean Square Error and Error Rate.

5.1 Work environment

The method was implemented in Matlab, and the technical characteristics of the computer used during the implementation phase are:

- *Central processing unit*: Intel (R) Core (TM) i5-6300U CPU @ 2.40 GHz 2.50 GHz.
- *Random access memory*: 8 GB.
- *Operating system*: Windows 10 Pro 64-bit.

5.2 Database description

The study was carried out using a data set of 770 Arabic sentences sourced from the Arabphone database (Frihia and Bahi, 2016). This database comprises spoken Modern Standard Arabic and features recordings of 30 Algerian adults, encompassing both genders, from the regions of Annaba, Jiel and Tarif. The data set consists of 2520 words and 12,000 phonemes recorded across diverse environments at a 16 kHz sampling rate. Each phoneme within a phrase is positioned at one of three available locations within the word: beginning, middle or end. Table 1 provides a selection of Arabic sequences utilised in this data set.

Table 1 Sequences of Arabphone corpus

Consonant	*Sequence*
ث	ثياب ثلاثة ثابت ورث
ج	جاء نجيب مع الحجاج
ح	ن الحيوا حافر الحجر جرح
خ	خديجة خاتم خالد أخذ
ذ	ذرى الفلاح القمح بالمذراة
ز	زار عزام جزيرة الكرز
س	السمع والبصر من الحواس
ش	لا أشرب الشاي بعد العشاء
ص	سرق اللصوص صندوق الصياد
ظ	نظيفة محفوظ أظافر
ع	لعماد العربة عادل عجبا

5.3 Assessment metric

The performance of the two methods was measured by using two criteria, namely Accuracy (ACC) and Mean Square Error (MSE):

- *Accuracy (ACC)*: For a comprehensive evaluation of the detection outcomes, it is essential to have parameters that characterise the accuracy rate. In our investigation, we outline four parameters pertaining to accuracy rate, as detailed below:

- *True positive (TP)*: Speech segments correctly identified as the accurate class.
- *True negative (TN)*: Silence or noise regions correctly categorised as non-speech segments.
- *False positive (FP)*: Speech segments incorrectly classified as silence.
- *False negative (FN)*: Silence or noise sections inaccurately labelled as speech.
- The ACC is computed by the following formula:

$$ACC(\%) = \frac{TP + TN}{TP + TN + FP + FN}. \quad (12)$$

- *Mean square error (MSE)*: The MSE (Sammut and Webb, 2011) refers to the sum of the squares of the errors, that is, the average squared difference between the predicted values and what is estimated. It is a risk function that displays the anticipated value of the squared error loss. It is defined as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}\left(X_i - \bar{X}_i\right)^2 \quad (13)$$

where X_i : The observed value, $\bar{X}_i$: The predicted value.

5.4 Results and discussion

In this section, the outcomes achieved by implementing both the method described in Giannakopoulos (2009) and our proposed approach to distinguish between speech and non-speech segments will be presented.

Table 2 lists the comparison results in terms of accuracy. It has four columns: One for speakers of various genders, another for the accuracy of the original approach, the third for the STE-VAD method and the last column describes the accuracy based on the SC-VAD. The last row displays the overall accuracy of each method: the original approach has a 78.54%, the STE-VAD has 90.79% while the STE-SC provides only 62.39%.

Table 2 Comparison of the accuracy

Speaker	*Sequence*	*STE-VAD*	*SC-VAD*
1	97.32	94.87	47.93
2	90.1	94.56	75.71
3	89.82	**98.61**	63.78
4	74.10	95.01	70.53
5	37.76	92.61	50.91
6	82.09	88.43	45.83
7	83.25	92.78	64.50
8	86.84	92.25	69.51
9	80.86	**98.58**	60.72
10	66.94	78.30	40.48
11	54.78	92.60	46.64
12	81.90	95.21	80.22
13	72.52	89.16	65.56
14	83.48	88.43	70.17
15	65.27	88.71	54.87
16	70.55	91.46	65.05
17	77.16	87.39	87.34
18	81.14	**96.56**	45.54
19	87.27	82.21	48.01
20	92.68	91.98	79.76
21	86.25	94.31	83.61
22	81.30	89.04	78.37
23	96.02	94.86	73.36
24	93.83	95.21	56.17
25	83.40	95.28	61.38
26	82.64	92.52	53.85
27	82.31	85.62	60.93
Total accuracy	78.54%	**90.79%**	62.39%

Based on this table, we can see that the STE-VAD method improves the accuracy of the original method, while the SC-VAD method reduces the accuracy. In our experiments, we observed that the values of the STE were significantly higher in the clean environment, which explains why the accuracy of speakers 3 and 9 (98.61%, 98.58%) who recorded their utterances in a clean environment improved.

Additionally, it was observed that the energy levels of the speech segments surpass those of the non-speech areas, which is the reason which justifies why the low energy sounds like ځ[_H] is defined as a silence area. This sound presents one of the fricative consonants in the Arabic language, which result from a narrowing or very narrow constriction of the vocal tract at the airflow meeting.

Moreover, several sounds in Arabic are not found in other languages. An example is the uvular plosive ق[*q*], which is considered a noise since the air releases an explosive noise during its production.

Table 3 compares the outcomes in terms of Mean Square Error (MSE), with MSE values by each speaker for the original approach, STE-VAD and SC-VAD. The SC-VAD does have a larger MSE than the other methods, with a value of 0.1706, this is to be expected considering its low accuracy of 62.39%. The original method yielded an MSE of 0.0028, whereas the STE-VAD provided the lowest value of 0.0013, indicating the efficiency of this approach.

Table 3 Comparison of the mean square error

Speaker	*Sequence*	*STE-VAD*	*SC-VAD*
1	0.0052	0.0009	2.4983
2	0.0009	0.0022	0.0596
3	0.0051	0.0002	0.0085
4	0.0014	0.0007	0.0140
5	0.0010	0.0078	0.0458
6	0.0059	0.0013	1.0486
7	0.0011	0.0009	0.0173
8	0.0011	0.0007	0.0108
9	0.0023	0.0009	0.0411
10	0.0038	0.0021	0.9759
11	0.0027	0.0018	0.1733
12	0.0007	0.0036	0.0062
13	0.0047	0.0022	0.0234
14	0.0037	0.0004	0.1313
15	0.0060	0.0001	0.0130
16	0.0014	0.00005	0.0125
17	0.0038	0.0003	0.0333
18	0.0015	0.0097	0.0130
19	0.0012	0.0014	0.0151
20	0.0036	0.0026	0.0033
21	0.0005	0.0004	0.0093
22	0.0049	0.0055	0.0244
23	0.0022	0.0009	0.0059
24	0.0028	0.0005	0.2526
25	0.0044	0.0009	0.0134
26	0.0012	0.0006	0.0063
27	0.0016	0.0008	0.0165
Mean square error	0.0028	**0.0013**	0.1706

Figure 6 states an example of a sentence recorded from the speaker 27, the green line indicates the beginning of the speech segment and the red line shows the end of the speech. The areas between these selected regions present the noise in the signal. From the reference (see Figure 5), this sentence has 6 speech segments, but the original VAD method detected only 3 active segments correctly. Figure 7 illustrates the results obtained using the STE-VAD where all speech segments are accurately classified. Figure 8 presents the output of the SC-VAD which considered the whole sentence as one speech segment, the existing noise in the signal causes this occurred error. Therefore, we can conclude that the STE-VAD was very performant.

Figure 5 Result of the VAD reference using Praat tool (see online version for colours)

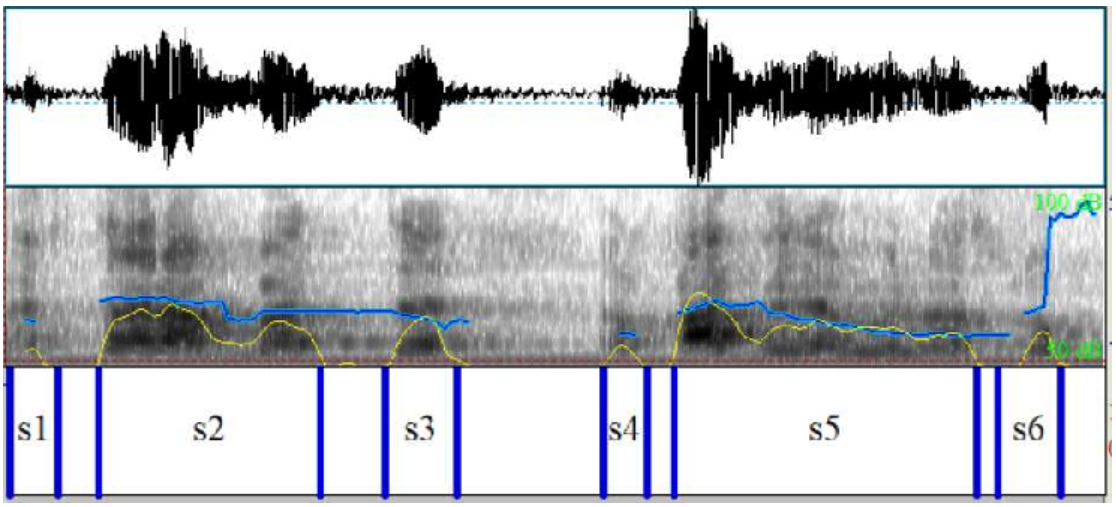


Figure 6 Result of the method (Giannakopoulos, 2009) (see online version for colours)

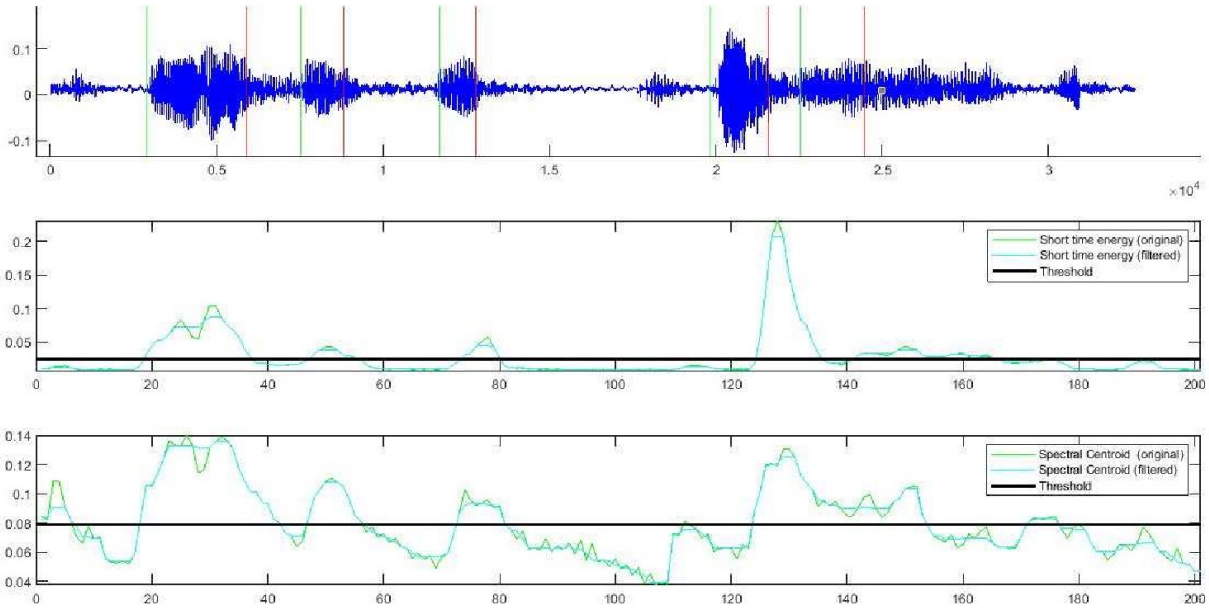


Figure 7 Result of STE-VAD method (see online version for colours)

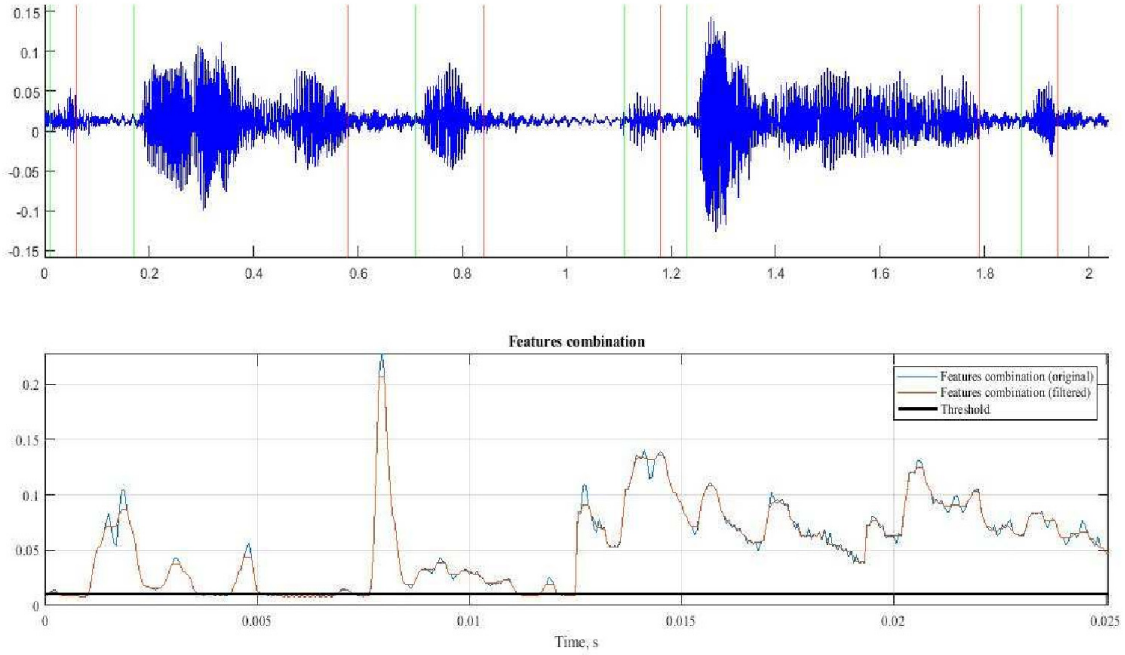
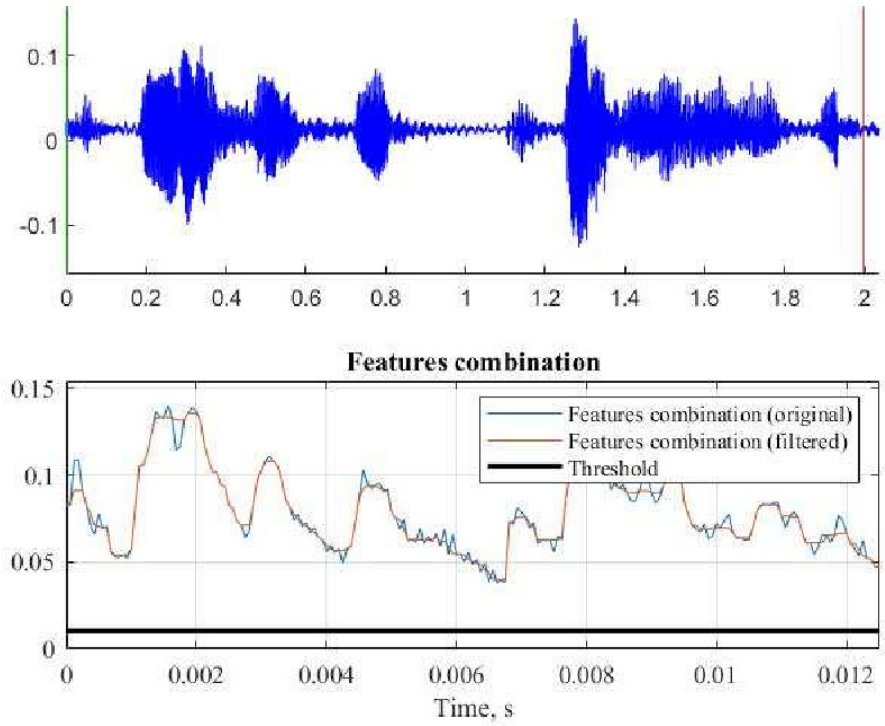


Figure 8 Result of SC-VAD method (see online version for colours)



5.4.1 SNR comparison

Table 4 summarises the evaluation of the two methods by using the same test data (Frihia and Bahi, 2016) and under different values of SNR: <5, 5, 20 and >20. Our STE-VAD method has a maximum and promising value of accuracy compared to the method (Giannakopoulos, 2009), in particular in the interval exceeding 20 dB with a value of 92.16%, which confirms its performance.

Table 4 VAD accuracy in different SNR intervals

SNR (db)	*<5*	[5, 20]	*>20*
The original method	76.24%	77.14%	86.52%
STE-VAD	**88.66%**	**91.41%**	**92.16%**
SC-VAD	52.56%	60.18%	69.28%

Table 5 shows the accuracy results after utilising the unsupervised VAD approach (Ali and Talha, 2018), which

used long-term features such as fundamental frequency, shimmer and jitter. The performance of this method was evaluated using the King Saud University (KSU) Arabic speech database in three different noisy environments (white, automobile and babble) at different SNR values. The highest accuracy for 5, 15 and 25 db is 88.91%, 91.12% and 94.72%, respectively.

Table 5 Accuracy for noisy sequences using the KSU database (Ali and Talha, 2018)

Noise / SNR (db)	*5*	*15*	*25*
White	88.46%	**91.12%**	94.27%
Car	**88.91%**	90.38%	9 4.53%
Babble	88.21%	90.16%	**94.72%**

According to the two Tables 4 and 5, we can see that our STE-VAD method yields promising results compared with the method (Ali and Talha, 2018), especially in the interval where SNR <5 with an accuracy value of 88.66%, this confirms that our method can identify the speech areas properly under low SNRs, while the other method has 88.46% and 88.21% at the 5 db, respectively for white and babbling environments.

The fricatives are the most numerous consonants in the Standard Arabic language, with 13 different fricative consonants (see Table 6). They are produced in the vocal cavity by a narrow constriction that makes the air circulation turbulent. There are two types of fricatives: Voiced and voiceless fricatives. Acoustically, the voiced fricatives have weak resonance structures that appear as shadows of feeble formants with a slight noise, while the voiceless possess a random high noise.

Table 6 Classification of fricative consonants in the Standard Arabic language

Fricatives	ث	ح	خ	ذ	ز	س	ش	ص	ظ	ع	غ	ف	ه
Transliteration	th	H	x	dh	z	s	sh	S	Z	‘	R	f	h
Voiced		+		+	+			+	+	+	+		+
Voiceless	+		+			+	+					+	

In our experiments, we treated the problem of the Arabic voiceless fricatives. Table 7 presents the values of the two parameters, FP and FN, used to calculate the Error Rate (ER), which is described in equation (14). We have found that the ER of the proposed method is 50.89%, while the method (Giannakopoulos, 2009) has 29.15%. This result demonstrates the robustness of the suggested method since solving the problem of voiceless fricatives will lead to a 50.89% drop in the ER, representing half of the committed errors. On the other hand, accuracy will improve, whereas the ER of the original method will only drop by 29.15%.

$$Error_{Rate}(\%) = \frac{FP_f + FN_f}{FP_t + FN_t} \tag{14}$$

where

FP_f: The areas of voiceless fricatives are detected as silence or noise.

FN_f: The non-speech segments in the sequences of voiceless fricatives are classified as speech.

FP_t: The total segments of speech detected as non-speech areas.

FN_t: The total segments of non-speech detected as speech.

Table 7 Number of voiceless fricatives misclassified

Voiceless fricatives	ث		خ		س		ش		ف	
	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
Original method (Giannakopoulos (2009)	103	4	147	11	128	7	130	13	111	8
STE-VAD	50	14	116	7	69	9	98	5	54	8

6 Conclusion and future work

An unsupervised Arabic VAD algorithm was proposed in this paper. It is based on the combination of the two features: Short Time Energy (STE) and Spectral Centroid (SC). The objective of combining these characteristics was the determination of criterion thresholding: T_{e+sc}. We use two versions of this approach: STE-VAD and SC-VAD. To evaluate our performance method and the efficiency of method (Giannakopoulos, 2009), we use the Arabphone database (Frihia and Bahi, 2016), which was recorded in noisy and clean environments and tested at various SNRs levels. The experimental results clearly show that the STE-VAD attained an excellent accuracy of 90.79% and a value of 0.0013 in terms of the MSE. While the SC-VAD reduces the accuracy to 62.39% and increases the MSE value to 0.1706, the method (Giannakopoulos, 2009) produces an accuracy of 78.54% and an MSE of 0.0028. Furthermore, we show that voiceless fricatives caused many problems in the Arabic VAD system. Solving these problems will develop a path-breaking method that can be used in real-time voice processing applications. As a perspective, our approach holds the potential to serve as a dependable input for a phonetic segmentation method for Arabic and Moroccan dialect speech. This future integration aims to streamline processing time and enhance the overall accuracy of the system.

References

Alam, T. and Khan, A. (2020) ‘Lightweight CNN for robust voice activity detection’, *Proceedings of the International Conference on Speech and Computer*, Springer, Cham, pp.1–12.

Ali, Z. and Talha, M. (2018) ‘Innovative method for unsupervised voice activity detection and classification of audio segments’, *IEEE Access*, Vol. 6, pp.15494–15504.

Ariav, I., Dov, D. and Cohen, I. (2018) ‘A deep architecture for audio-visual voice activity detection in the presence of transients’, *Signal Processing*, Vol. 142, pp.69–74.

Arslan, A. and Engin, E.Z. (2019) ‘Noise robust voice activity detection based on multi-layer feed-forward neural network’, *Electrica*, Vol. 19, No. 2, pp.91–100.

Bäckström, T. (2017) *Speech coding: with Code-Excited Linear Prediction*, Springer.

Bai, L., Zhang, Z. and Hu, J. (2017) 'Voice activity detection based on deep neural networks and Viterbi', *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, Vol. 231, No. 1. Doi: 10.1088/1757-899X/231/1/012042.

Benyassine, A., Shlomot, E., Su, H.Y., Massaloux, D., Lamblin, C. and Petit, J.P. (1997) 'ITU-T Recommendation G. 729 Annex B: a silence compression scheme for use with G. 729 optimized for V. 70 digital simultaneous voice and data applications', *IEEE Communications Magazine*, Vol. 35, No. 9, pp.64–73.

Çolak, R. and Akdenniz, R. (2021) 'A novel voice activity detection for multi-channel noise reduction', *IEEE Access*, Vol. 9, pp.91017–91026.

Dean, D., Sridharan, S., Vogt, R. and Mason, M. (2010) 'The QUT-NOISE-TIMIT corpus for evaluation of voice activity detection algorithms', *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pp.3110–3113.

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

Eyben, F., Weninger, F., Squartini, S. and Schuller, B. (2013) 'Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies', *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp.483–487.

Frihia, H. and Bahi, H. (2016) 'Embedded learning segmentation approach for Arabic speech recognition', *Proceedings of the International Conference on Text, Speech, and Dialogue*, Springer, Cham, pp.383–390.

Ghanbari, Y. and Karami-Mollaei, M.R. (2006) 'A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets', *Speech Communication*, Vol. 48, No. 8, pp.927–940.

Giannakopoulos, T. (2009) *A method for Silence Removal and Segmentation of Speech Signals, Implemented in Matlab*, University of Athens, Athens.

Hughes, T. and Mierle, K. (2013) 'Recurrent neural networks for voice activity detection', *IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp.7378–7382.

Kim, C. and Stern, R.M. (2008) 'Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis', *Proceedings of the 9th Annual Conference of the International Speech Communication Association*, pp.2598–2601.

Kim, S.K., Kang, S.I., Park, Y.J., Lee, S. and Lee, S. (2016) 'Power spectral deviation-based voice activity detection incorporating teager energy for speech enhancement', *Symmetry*, Vol. 8, No. 7.

Kinnunen, T., Chernenko, E., Tuononen, M., Fränti, P. and Li, H. (2007) 'Voice activity detection using MFCC features and support vector machine', *Proceedings of the International Conference on Speech and Computer (SPECOM'7)*, Moscow, Russia, Vol. 2, pp.556–561.

Ma, Y. and Nishihara, A. (2013) 'Efficient voice activity detection algorithm using long-term spectral flatness measure', *EURASIP Journal on Audio, Speech, and Music Processing*, No. 1, pp.1–18.

Moattar, M.H. and Homayounpour, M.M. (2009) 'A simple but efficient real-time voice activity detection algorithm', *Proceedings of the 17th European Signal Processing Conference*, IEEE, pp.2549–2553.

Rho, D., Park, J. and Ko, J.H. (2022) 'NAS-VAD: neural architecture search for voice activity detection', *arXiv preprint arXiv:2201.09032*.

Sadjadi, S.O. and Hansen, J.H. (2013) 'Unsupervised speech activity detection using voicing measures and perceptual spectral flux', *IEEE Signal Processing Letters*, Vol. 20, No. 3, pp.197–200.

Sammut, C. and Webb, G.I. (Eds) (2011) *Encyclopedia of Machine Learning*, Springer Science Business Media.

Schubert, E. and Wolfe, J. (2006) 'Does timbral brightness scale with frequency and spectral centroid?', *Acta Acustica United with Acustica*, Vol. 92, No. 5, pp.820–825.

Sehgal, A. and Kehtarnavaz, N. (2018) 'A convolutional neural network smartphone app for real-time voice activity detection', *IEEE Access*, Vol. 6, pp.9017–9026.

Shin, J.W., Chang, J.H. and Kim, N.S. (2010) 'Voice activity detection based on statistical models and machine learning approaches', *Computer Speech Language*, Vol. 24, No. 3, pp.515–530.

Silva, D.A., Stuchi, J.A., Violato, R.P.V. and Cuozzo, L.G.D. (2017) 'Exploring convolutional neural networks for voice activity detection', *Cognitive Technologies*, Springer, Cham, pp.37–47.

Yoo, I.C., Lim, H. and Yook, D. (2015) 'Formant-based robust voice activity detection', *IEEE/ACM Transactions on Audio, Speech, and language Processing*, Vol. 23, No. 12, pp.2238–2245.

Zaw, T.H. and War, N. (2017) 'The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection', *Proceedings of the 20th International Conference of Computer and Information Technology (ICCIT)*, IEEE, pp.1–5.

Zhang, X.L. and Wu, J. (2012) 'Deep belief networks based voice activity detection', *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, No. 4, pp.697–710.

Zhang, X.L. and Xu, M. (2022) 'AUC optimization for deep learning-based voice activity detection', *EURASIP Journal on Audio, Speech, and Music Processing*, No. 1, pp.1–12.