

Received May 11, 2021, accepted May 30, 2021, date of publication June 4, 2021, date of current version July 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3086364

# A Novel Voice Activity Detection for Multi-Channel Noise Reduction

RAMAZAN ÇOLAK<sup>1</sup> AND RAFET AKDENİZ<sup>2</sup>

<sup>1</sup>TRT İstanbul Television, 34347 İstanbul, Turkey

<sup>2</sup>Department of Electronics and Telecommunication Engineering, Çorlu Engineering Faculty, Tekirdağ Namık Kemal University, 59860 Tekirdağ, Turkey

Corresponding author: Rafet Akdeniz (rakdeniz@nku.edu.tr)


This work was supported by the Tekirdağ Namık Kemal University Scientific Research Project Commission under Grant NKÜBAP.06.YL.18.156.

**ABSTRACT** In this study, a voice activity detection technique is designed using features such as short-term energy, periodicity and spectral flatness. The desired results are obtained by using these three features, even at low signal to noise ratio values. In addition, performance of multi-channel noise reduction algorithms such as Wiener speech distortion weighted, spatial prediction, minimum variance distortion-less response are compared using the proposed voice activity detection. Two different audio signals and three different noise types are used in the experiment. Noisy speech and only detection of noisy areas have been performed by proposed voice activity detection algorithm. The filter coefficients have been calculated for each filter algorithm used after detection of noisy speech and only noisy areas. The calculated filter coefficients have been multiplied by the frequency components of the signal received from the reference microphone to obtain an enhanced signal. Segmental signal to noise ratio, an objective method, and mean opinion score as a subjective method have been used to evaluate the performance of the filters. Speech distortion weighted Wiener filter has been found to be the best filter for noise reduction performance.

**INDEX TERMS** Adaptive filter, noise reduction, speech enhancement, voice activity detection.

## I. INTRODUCTION

One of the most fundamental problems affecting the quality of speech in the communication industry is noise. There are many studies in the literature to deal with this noise problem. The VAD algorithm has a very important place in speech enhancement algorithms. In speech signals; The correct detection of loud speech and noise is very important in order to calculate the correct filter coefficients. Voice activity detection (VAD) is used to distinguish between the parts with the voice and the parts with noise in speech processing. This application is used in the first stage of speech processing processes such as speech enhancement, speech recognition and speech coding and it directly affects the performance of the process. In VAD algorithms, the features such as efficiency, robustness, and the simplicity of the algorithm are directly related to the availability of the algorithm. Features such as short-time energy (STE), periodicity, zero crossing rate (ZCR), spectral flatness (SF), most dominant frequency component, high-low frequency rate are some of the features used for the detection of noisy areas and noisy speech areas.

The associate editor coordinating the review of this manuscript and approving it for publication was Fang Yang .

In the early days, single-microphone studies were conducted in noise reduction methods [1], [2] and [3]. Better results were obtained from noise reduction methods performed with multiple microphones by benefiting from the correlation between signals received in the future.

A study was conducted by Meyer and Simmer [4] to reduce the noise inside the vehicle. First of all, the signal in the time plane was converted to the frequency plane and the new algorithm was produced by calculating and combining the spectral subtraction with the low pass filter and the Wiener filter coefficients with the high pass filter. With this new method, better results are obtained than traditional spectral subtraction and Wiener algorithms.

Rao *et al.* [5], a two-stage hybrid system was developed in a study. In spectral gain calculations, frequency indices are distributed non-homogeneously for ease of calculation. In the first stage of this system, the softened decision gain mechanism created to the Ephraim-Malah gain function was applied. In the second stage, psychoacoustic masking threshold was used for noise reduction. This proposed method has been compared with Spectral Subtraction and Spectral Weighting algorithms and it has been evaluated to be suitable for use in unstable noisy environments.

In a study by Chen *et al.* [6], A proposal is presented to parse the filtered audio signal in the time domain. The filtered audio signal is split into two unrelated components. This new parsing method has been tested with maximum signal to noise ratio (SNR), Wiener, minimum variance distortion-less response (MVDR) and Trade off filters. Experimental results and theoretical analysis It has been determined that the maximum SNR, Wiener and Trade off filter are identified with the MVDR filter using a scaling factor, but since this scaling parameter will cause distortion in the speech signal, the MVDR filter is recommended in speech enhancement applications.

The STE feature alone does not give the desired results, especially at low SNR values, therefore, an algorithm has been developed that yields better results at low signal to noise ratio (SNR) values by using SF and most dominant frequency components in the frequency domain, in addition to STE [7]. An efficient VAD algorithm has been proposed by K. Sakhnov *et al.* In this study has been used short-term features like periodicity, and high-low frequency rate of the voice [8].

The STE feature, as well as the periodicity, and the high-low frequency rate, have been used in another algorithm, which is new and easy to implement, developed by K. Sakhnov and E. Verteletskaya [9]. Using an algorithm similar to [9], a new VAD algorithm has been presented by using STE, most dominant frequency component, SF feature, and a different feature called peak-valley difference [10].

According to [11] an energy based VAD has been developed for in-ear listening devices such as earphones or headphones. This system allows consumers to hear external speech signals such as public announcements while listening to music without their listening devices.

In [12], a new voice activity detection method has been proposed based on the total spectrum energy. Because of the speech frequencies are in the lower frequency in the frequency spectrum, the noise energy from the higher frequency band is subtracted from the noisy speech spectrum in the lower frequency band and speech regions in the frequency spectrum is detected. In another study, using VAD concepts in the time domain and the frequency domain, the linear energy-based detector and the fuzzy logic and artificial neural network-based VAD performances have been compared [13].

In a study by Zaw and War [14], a combined parameter of  $D$  has been calculated for each frame of the audio signal by using STE, ZCR, spectral entropy, and the linear prediction error and the presence of speech has been determined by whether or not each audio frame is above the threshold level specified as  $D/D_{max}$ .

In another study, a new multi-channel speech enhancement algorithm has been proposed for use in hearing aids. In this algorithm, all noise components are considered as a matrix structure. In multiple speech signals, each speech signal is distorted by a noise signal. With the Wiener matrix structure, a filter coefficient is derived for each speech signal by evaluating the noise components entered into the matrix. With this

system, better efficiency is obtained from both single-channel and multi-channel Wiener filter applications [15] and [16].

In the study by Itzhak *et al.* [17] present a modified optimization criterion according to which the proposed filters may be derived, and compare their performances to conventional multichannel noise reduction filters. They show that the new approach is preferable, in particular when the input signal-to-noise ratio (SNR) is low or the number of sensors is small. In the study by Li *et al.* [18] propose a deep neural network - based generalized sidelobe canceller structure, which utilizes the traditional super-directive beamforming knowledge and the blocking technique to simultaneously perform localization and denoising. In the study by Wu *et al.* [19] introduce an end-to-end modeling version of unmixing, fixed-beamformer and extraction (UFE). To enable gradient propagation all the way, an attentional selection module is proposed, where an attentional weight is learnt for each beamformer and spatial feature sampled over space. In the study by Yang *et al.* [20], they propose a new framework for dereverberation by expressing the multichannel linear prediction filter as a Kronecker product of a temporal (interframe) filter and a spatial filter, and the lengths of the two filters correspond, respectively, to the order of the prediction filter and the number of microphones. In another study proposes a technique for improving statistical-model-based VAD in noisy environments to be applied in an auditory hearing aid. The proposed method is implemented for a uniform polyphase discrete Fourier transform filter bank satisfying an auditory device time latency of 8 ms. The proposed VAD technique provides an online unified framework to overcome the frequent false rejection of the statistical-model-based likelihood-ratio test (LRT) in noisy environments. This method is based on the observation that the sparseness of speech and background noise cause high false-rejection error rates in statistical LRT-based VAD—the false rejection rate increases as the sparseness increases [21].

In the study by Mahmmud *et al.* [22], Speech Enhancement Algorithms (SEAs) have been developed to deal with noisy signals, restore clean speech signals, improve speech quality and intelligibility, solve the noise pollution problem, and reduce listener fatigue.

In the study by Liang [23], A deep-learning method combining the attention mechanism for single-channel speech enhancement (SE) is proposed. Based on the traditional LSTM model, an attention-gate-based LSTM (Atten-LSTM) is proposed. In addition, the algorithm divides the speech band according to the bark scale; thus, the algorithm can simulate human auditory characteristics and improve speech quality. These bands gains, rather than the entire band gain, are used as training targets, thereby reducing the computational complexity. The performance comparison of different algorithms shows that the robustness and SE performance of the proposed algorithm are improved, thereby proving that the proposed attention model shows superior performance in SE. While maintaining low complexity, this method effectively

suppresses noise, obviously improves speech quality and is well generalizable to nonmatching samples.

In this study, an easy-to-implement VAD algorithm has been developed by using the features of STE, periodicity and SF to distinguish noisy speech and only areas with noise in a noisy environment. With proposed algorithm, noisy speech signals have been enhanced by using noise reduction methods. Section II provides basic information about microphone signals and the short-time features. Noise reduction algorithms used have been explained in section III. Section IV provides information about proposed method. In section V, the performances of the noise reduction filters and proposed algorithm has been evaluated.

## II. SIGNAL MODEL AND SHORT-TIME FEATURES

This section provides information about microphone signals and short-time features used in the proposed algorithm.

### A. SIGNAL MODEL

We consider a microphone array consisting of 3 microphones. The  $i$ th microphone signal  $Y_i(k, l)$  can be specified in the frequency domain as;

$$Y_i(k, l) = X_i(k, l) + V_i(k, l) \quad (1)$$

where  $X_i(k, l)$  is the speech component of the signal,  $V_i(k, l)$  is the noise component of the signal,  $k$  is frequency index,  $l$  is frame index,  $i$  is the number of microphones. Audio signal vectors used in the same number of microphones have been collected in the same matrix.

$$\mathbf{Y}(k, l) = [Y_1(k, l) \quad Y_2(k, l) \quad Y_3(k, l)]^T \quad (2)$$

where superscript  $T$  denotes transposition of the matrix. Autocorrelation matrices are as follows;

$$\mathbf{R}_y(k, l) = \varepsilon\{\mathbf{Y}(k, l)\mathbf{Y}^H(k, l)\} \quad (3)$$

$$\mathbf{R}_x(k, l) = \varepsilon\{\mathbf{X}(k, l)\mathbf{X}^H(k, l)\} \quad (4)$$

$$\mathbf{R}_v(k, l) = \varepsilon\{\mathbf{V}(k, l)\mathbf{V}^H(k, l)\} \quad (5)$$

where  $H$  denotes Hermitian transpose,  $\mathbf{R}_y(k, l)$  denotes autocorrelation matrix of noisy speech,  $\mathbf{R}_x(k, l)$  denotes autocorrelation matrix of clean speech signal,  $\mathbf{R}_v(k, l)$  denotes autocorrelation matrix of noise and  $\varepsilon\{\cdot\}$  represents the expectation operator. Assuming that speech and noise autocorrelation matrices are unrelated,  $\mathbf{R}_x(k, l)$  can be calculated from  $\mathbf{R}_y(k, l) = \mathbf{R}_x(k, l) + \mathbf{R}_v(k, l)$ . After calculating the autocorrelation matrices, the filter coefficients can be calculated. The signal received from the reference microphone is multiplied by the filter coefficients and the output signal  $\mathbf{Z}(k, l)$  is obtained.

$$\mathbf{Z}(k, l) = \mathbf{W}^H(k, l)\mathbf{Y}(k, l) \quad (6)$$

where  $Z(k, l)$  is the complete output from the multi-channel Wiener filter based noise reduction which may contain some residual noise depending on the estimated filter  $\mathbf{W}^H(k, l)$ .

### B. SHORT-TIME FEATURES

Operations on speech signals are performed at very small intervals. In this study has been used three different short-term features for each frame of voice in the proposed algorithm.

#### 1) SHORT-TIME ENERGY

The amplitude of speaking varies in speech signals depending on time. The STE value is higher in the parts including speech than those without a speech. Thus, it gives us important clues about the parts including speech. The equation of STE is as follows;

$$E(n) = \sum_{i=1}^N X_n^2(i) \quad (7)$$

where  $N$  denotes the length of the audio frame,  $X(i)$  denotes the original speech signal, and  $E(n)$  denotes the energy of the audio frame. As can be seen, the energy of an audio frame can be calculated by summing the squares of each audio sample. Fig. 1 shows a clean speech signal and the STE of this speech signal.

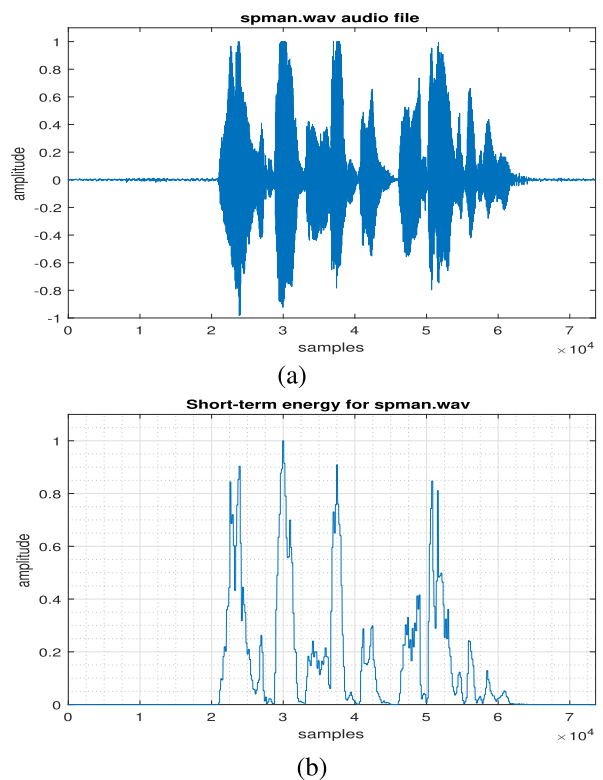


FIGURE 1. (a) spman.wav speech signal (b) its STE.

The linear energy based detector (LED), a traditional method, is a VAD algorithm that determines whether a frame of the audio signal is noisy speech or only noise. It is based on a comparing each frame to a certain STE threshold level.

$$\begin{aligned} \text{If } (E_j > K \times E_r), & \quad \text{Frame is voice} \\ \text{Else,} & \quad \text{Frame is noise} \end{aligned} \quad (8)$$

where  $K$  denotes scale factor ( $K > 1$ ),  $E_j$  denotes the energy level of the frame and  $E_r$  denotes the energy level of the noisy frame,  $K \times E_r$  is the threshold value used to determine whether the frame is noisy speech or only noise. The noise threshold value is updated differently in various LED algorithms [24]. The updating equation for the noisy frame is shown in (9).

$$E_{rnew} = (1 - p) \times E_{rold} + p \times E_{noise} \quad (9)$$

where  $E_{rnew}$  denotes the updated energy threshold level, and  $E_{rold}$  denotes the old energy threshold level,  $E_{noise}$  denotes the energy of the current frame with noise. The  $P$  parameter is a constant selected between 0 and 1.

### 2) PERIODICITY

Periodicity is an important feature to detect parts with voice in the audio signals such as speech and music. The periodicity of the signal can be determined by pitch estimation (the shortest repeatable interval). The normalized auto-correlation function  $R(\tau)$  can be calculated by using (10).

$$R(\tau) = \frac{\sum_{n=1}^{N-m-1} X(n)X(n + \tau)}{\sqrt{\sum_{n=1}^{N-m-1} X^2(n + \tau)}} \quad (10)$$

$$C = \max(R(\tau)) \quad (11)$$

where  $\tau$  denotes the lag value. In (10),  $X(n)$ ; ( $n = 0, 1, \dots, N$ ) denotes the length of the input signal's frame. The autocorrelation function is calculated by using the values of  $\tau$  between  $T_{min}$  and  $T_{max}$ . In the proposed algorithm, we select  $T_{min} = 16$  samples and  $T_{max} = 64$  samples, respectively. The maximum  $R(\tau)$  value,  $C$ , for each audio frame gives the periodicity of that audio frame. When  $C = 1$  the signal can be said to be completely periodic, and when  $C = 0$  the signal can be said to be a random signal. Fig. 2 shows a clean speech signal and the periodicity of this speech signal.

### 3) SPECTRAL FLATNESS

SF is a feature used in the frequency domain. SF values tend to approach zero in the parts which include just noise. It tends to go to  $-\infty$  (minus infinity) in the parts with the speech. SF can be calculated in  $dB$  by using (12).

$$SF_{dB} = 10 \log_{10}(G_m/A_m) \quad (12)$$

where  $G_m$  denotes geometric mean and  $A_m$  denotes arithmetic mean. In order to calculate SF, the audio signal is separated into the frames and each frame is undergone Fast Fourier Transform (FFT) separately, then the arithmetic mean and geometric mean values for each frame are calculated. Fig. 3 shows a clean speech signal and the SF of this speech signal.

## III. NOISE REDUCTION METHODS

In this section, basic information about four different adaptive filters which were used in the study, has been given.

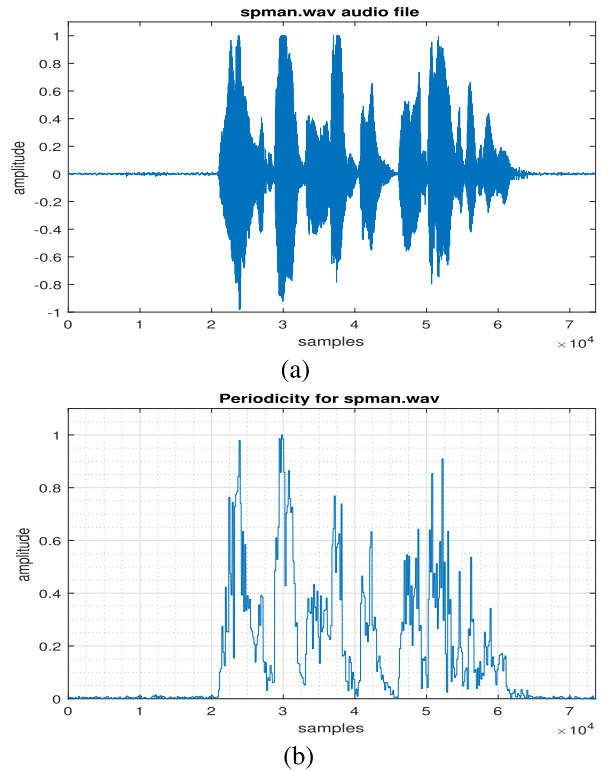


FIGURE 2. (a) spman.wav speech signal (b) its periodicity.

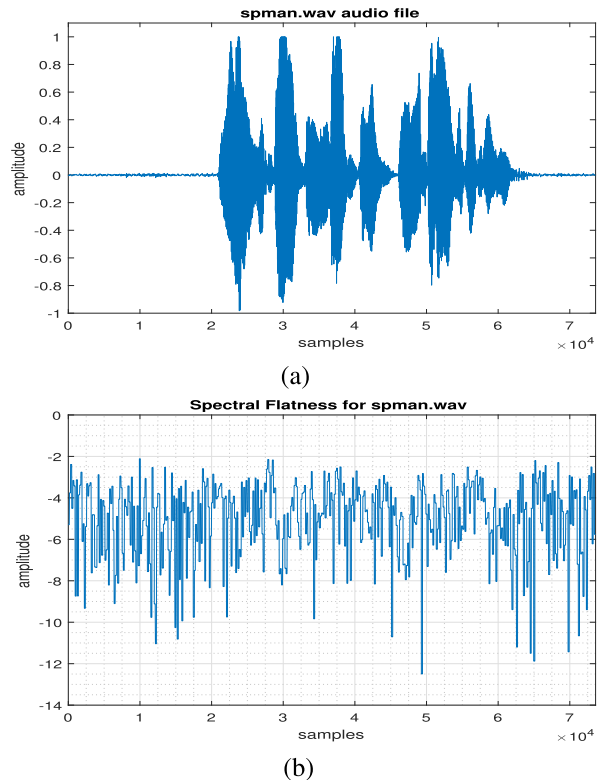


FIGURE 3. (a) spman.wav speech signal (b) its SF.

### A. MULTI-CHANNEL WIENER FILTER

By minimizing the minimum mean square error (MMSE) in the Wiener filter, the error in the filter output has been kept

as minimum [25].

$$\mathbf{W}_{MMSE}(k, l) = \arg \min_{\mathbf{W}(k, l)} \varepsilon\{|X_1(k, l) - \mathbf{W}^H(k, l)\mathbf{Y}(k, l)|^2\} \quad (13)$$

In (13), the basic formula for the calculation of Wiener filter coefficients have been given.  $X_1(k, l)$  represents the desired clean speech, and one of the microphones (usually the first microphone) is considered as a reference and this clean speech is received from that microphone. In (13), desired filter coefficients can be obtained to minimize noise by taking the derivative according to  $\mathbf{W}(k, l)$ .

$$\frac{\partial J_{MMSE}(\mathbf{W}(k, l))}{\partial(\mathbf{W}(k, l))} = -2\varepsilon\{\mathbf{Y}(k, l)X_1^H(k, l)\} + 2\varepsilon\{\mathbf{Y}(k, l)\mathbf{Y}^H(k, l)\mathbf{W}(k, l)\} \quad (14)$$

Assuming that speech and noise signals are uncorrelated;

$$\varepsilon\{\mathbf{V}(k, l)X_1(k, l)\} = 0 \quad (15)$$

By dissolving (14), the formula Multi-channel Wiener filter (MWF) has been obtained as in (16).

$$\mathbf{W}_{MMSE}(k, l) = [\mathbf{R}_x(k, l) + \mathbf{R}_v(k, l)]^{-1}\mathbf{R}_x(k, l)e_1 \quad (16)$$

where  $e_1$  represent the first column of the unit matrix as long as the number of microphones ( $e_1 = [1 \ 0 \dots 0]^T$ ).

### B. SPEECH DISTORTION WEIGHTED WIENER FILTER

Derived from the classic multi-channel Wiener filter, in this filter  $\mu$  parameter provides a relation between speech distortion and noise reduction [26]. Speech distortion weighted (SDW) multi-channel Wiener filter coefficients are calculated as follows:

$$\mathbf{W}_{MWF\mu}(k, l) = [\mathbf{R}_x(k, l) + \mu\mathbf{R}_v(k, l)]^{-1}\mathbf{R}_x(k, l)e_1 \quad (17)$$

As it is clear in (17) when it is  $\mu = 1$  the formula of the classical Wiener filter is obtained. When it is  $\mu > 1$ , noise component in the filtered signal further reduced, but causes more speech distortion.

### C. SPATIAL PREDICTION FILTER

In this approach, the noisy speech signal received from the first microphone is considered as a reference signal and a vector is created which shows its relationship with the other microphone signals [27].

$$\mathbf{X} = \begin{bmatrix} H_{1,ref} \\ H_{2,ref} \\ \vdots \\ H_{N,ref} \end{bmatrix} X_{ref} = \mathbf{H}X_{ref} \quad (18)$$

In this way, the SP filter is designed by creating a spatial prediction vector of speech components. In (18), the spatial prediction vector  $\mathbf{H}$  can be found in the Wiener sense by minimizing as shown in (19).

$$\min_H \varepsilon\{(\mathbf{X} - \mathbf{H}X_{ref})^H(\mathbf{X} - \mathbf{H}X_{ref})\} \quad (19)$$

The spatial vector  $\mathbf{H}$  is found as shown in (20).

$$\mathbf{H}(k, l) = \frac{1}{u^H\mathbf{R}_x(k, l)u}\mathbf{R}_x(k, l)u \quad (20)$$

Which means one column of the speech correlation matrix is selected and divided by the first element of the speech correlation matrix.  $u$  is a vector of length  $N$  with the first element equal to one the other elements are zero ( $u = [1 \ 0 \dots 0]^T$ ). By using (18), speech error  $E_x$  can be written as follows;

$$E_x = (\mathbf{W} - u)\mathbf{H}\mathbf{X} = (\mathbf{W}^H\mathbf{H} - 1)X_{ref} \quad (21)$$

So that  $E_x = 0$  if  $\mathbf{W}^H\mathbf{H} = 1$ , According to this optimization problem; filter coefficients are calculated as follows;

$$\mathbf{W}_{SP}(k, l) = \frac{1}{\mathbf{H}^H\mathbf{R}_v^{-1}(k, l)\mathbf{H}(k, l)}\mathbf{R}_v^{-1}(k, l)\mathbf{H}(k, l) \quad (22)$$

### D. MINIMUM VARIANCE DISTORTION-LESS RESPONSE FILTER

In this approach based on Beam-forming technique proposed by Capon [28], to calculate the filter coefficients, a steering vector of the target speech signal has been generated and the filter coefficients have been calculated with the steering vector changing at each time interval. While trying to minimize the error in the output of the minimum variance distortion-less response (MVDR) filter, it is requested not to disturb the target speech signal. The optimization formula for this filter is as follows;

$$\min_{\mathbf{W}(k, l)} \mathbf{W}^H(k, l)\mathbf{R}_x(k, l)\mathbf{W}(k, l), \quad e^H(k, l)\mathbf{W}(k, l) = 1 \quad (23)$$

where  $e^H(k, l)$  is steering vector. Based on the principle that the product of the filter coefficient vector and the steering vector elements are equal to one, it is tried to calculate the filter coefficient vector from the known steering vector. The closest solution to (23) is given by Capon [26].

$$\mathbf{W}_{MVDR}(k, l) = \frac{1}{e^H\mathbf{R}_x^{-1}(k, l)e(k, l)}\mathbf{R}_x^{-1}(k, l)e(k, l) \quad (24)$$

### IV. PROPOSED ALGORITHM

In the proposed algorithm, three properties such as STE, Periodicity and SF are used to detect areas with noisy speech and noisy. The threshold value of the Periodicity has been set to 0.5. STE and SF threshold values have been calculated using noisy areas in noisy speech. Firstly, the audio signal has been divided into frames with 128 samples. The average of the STE values of the first 20 only noisy frames has been considered as the initial threshold level of the energy of the noise ( $STE_{threshold}$ ). The three properties described above have been calculated for each frame. If one or more of these three features is above the specified threshold level, the presence of the speech has been assumed, whereas it has been assumed that there is no speech if none of these features is above the threshold level. In this study, three different types of noise have been added on different speech samples, and the proposed algorithm was examined and similar



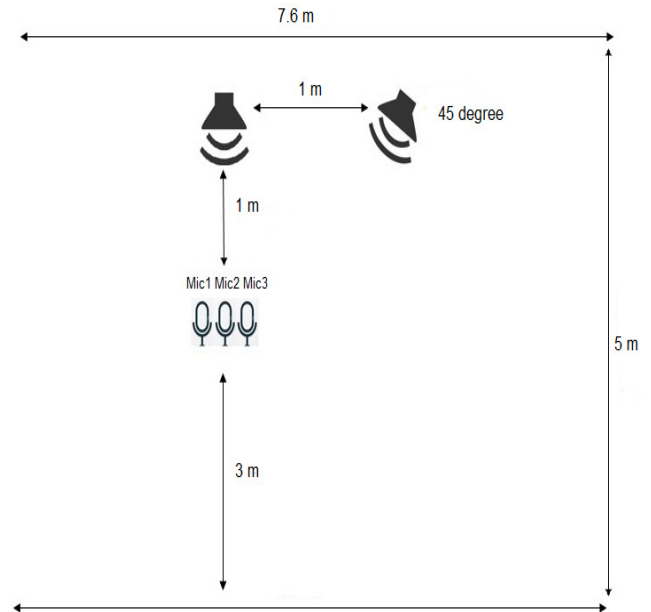
**Algorithm 1** The Script Cycle of the Proposed Algorithm

1. The size of the audio frame has been considered to be 128 sample.
2. The threshold values for the used features have been determined externally.
  - $K$  constant for STE ( $K > 1$ )
  - Periodicity threshold value
  - $SF_{th}$  value
3. for  $i=1$  to the number of frames
  - 3.1 Compute STE( $i$ )
  - 3.2 Compute Periodicity( $i$ )
  - 3.3 Apply FFT for each frame.
    - 3.3.1 Compute SF value for each frame.
  - 3.4 After determining threshold values for STE, Periodicity and SF, the average of the energy of the first 20 frames has been considered as the threshold value of the noise, and the average of the SF values has been considered as the mean SF value of the noise, and then  $STE_{noise}$  and  $SF_{noise}$  values are computed.
    - 3.4.1  $SF_{noise} = SF(1, 1 : 20)/20$
    - 3.4.2  $STE_{noise} = STE(1, 1 : 20)/20$
  - 3.5 Counter=0
    - If  $STE(i) \geq K \times STE_{noise}$  ;  
counter=counter+1
    - If  $SF(i) - SF_{noise} \leq SF_{th}$  ;  
counter=counter+1
    - If  $Periodicity(i) \geq 0.5$  Then ;  
counter=counter+1
  - 3.6 If counter  $\geq 1$  ; The audio frame is a speech with noise;  
If counter=0 ; Audio frame is noise;
  - 3.7 If the audio frame is noise, the threshold value of STE is updated using the following equation.  
 $STE_{th-new} = (1 - p) \times STE_{th-old} + p \times STE_{noise}$
4. If there are less than 10 consecutive frames of silence, it is updated with 1.
5. 5. If there are less than 5 consecutive frames of speech, it is updated as 0.

results were obtained. Pre-defined noises are added to a few speech examples and the results are shown in this article. The script cycle of the proposed algorithm has been given in Algorithm 1.

## V. EXPERIMENTAL RESULTS

In this study; Computer with Intel Core i5-5200U CPU 2.2 GHz processor, 8 GB RAM memory and 64 bit Windows 10 operating system, MATLAB R2016a program, audio cable, speaker cable, three microphones, three microphone stands, speaker stand, speaker, audio card and Adobe Audition audio editing program were used. The audio files used in the study are the sounds of a female and a male speaker recorded in the studio. The noise files used have been taken from the AURORA database [29]. Speech and noise audio



**FIGURE 4.** Recording environment.

files have been sampled with a sampling frequency of 16 kHz. The speech signals used has been recorded at  $0^\circ$  angle to the three microphones and at a distance of one meter. Noise signals has been recorded at  $45^\circ$ . The recording environment created using these equipment is shown in the Fig. 4.

In this section, experimental results of proposed algorithm have been obtained. Using the proposed algorithm, the performances of four different noise reduction algorithms have been evaluated. Noise and clean speech signals have been mixed according to segmental SNR principle. Fig. 5 and Fig. 6 show the results of the VAD algorithm for noisy man speech and noisy woman speech.

Speech examples used are as follows:

sp09.wav: “Hurdle the pit with the aid of a long pole.”

sp11.wav: “He wrote down a long list of items.”

spwoman.wav: “For things to do, I have to go home early.”  
(in Turkish)

spman.wav: “I will go to Ankara at 14:00 tomorrow.”  
(in Turkish)

According to Fig. 5 and Fig. 6, it has been observed that the detection of speech areas in 10 dB noisy speech is better than 5 dB noisy speech. As can be understood from this, only when the energy level of noisy areas is high, it negatively affects the results. Table 1, Table 2, Table 3 and Table 4 show the results of the assessments for spwoman.wav, spman.wav, sp09.wav and sp11.wav signals by using  $HRI$  (Speech Hit Rate),  $HRO$  (Silence Hit Rate) and the average of these two which are widely used to measures the efficiency of the VAD algorithms.

According to all tables, the detection of speech and noisy speech areas in white noisy speech is better than airport speech and car noisy speech. The variability of the noise level in airport and car noisy speeches have been affected the results.

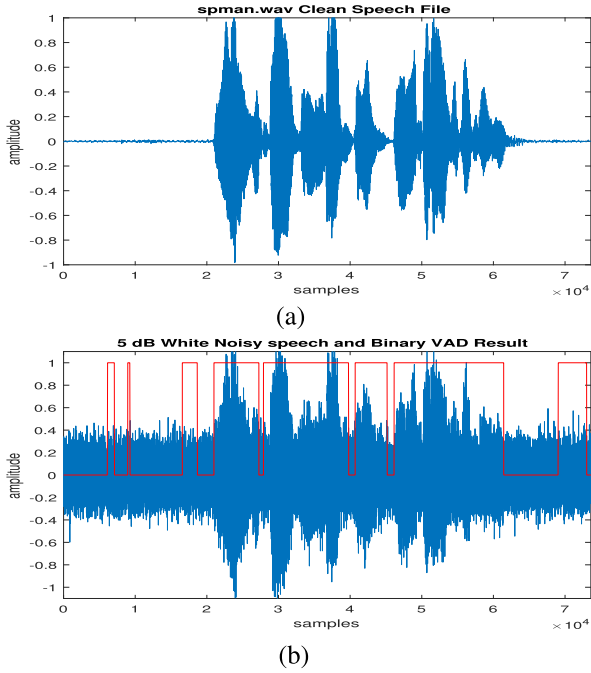


FIGURE 5. (a) Waveform results of 5 dB White noisy man speech (b) VAD result of proposed algorithm.

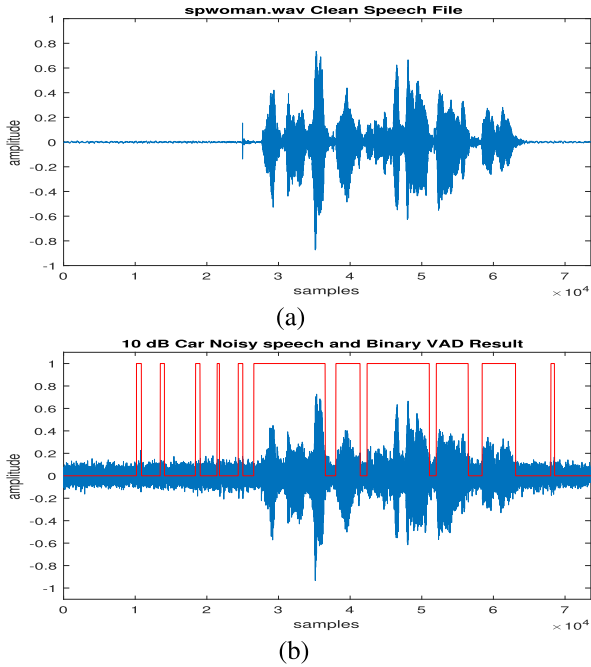


FIGURE 6. (a) Waveform results of 10 dB car noisy woman speech (b) VAD result of proposed algorithm.

In practice, the length of the frames has been taken as 128 samples (8 ms). 50% Overlapping method has been applied. The Hanning windowing method has been used. Noisy speech and only noisy areas have been calculated by proposed VAD algorithm. Correlation matrices have been calculated after the VAD algorithm determined whether the frames are noise or noisy speech. Elements of the first frame of the  $\mathbf{R}_y$  autocorrelation matrix have been calculated using 100 frames with noise and elements of the first frame of

TABLE 1. Experimental results of noisy woman speech.

Type of noise	Input SNR	HR1(%)	HR0(%)	A(%)
Airport	0 dB	89.72	42.73	66.22
	5 dB	89.11	62.02	75.56
	10 dB	96.77	75.34	86.05
Car	0 dB	83.87	52.68	68.27
	5 dB	90.12	66	78.06
	10 dB	95.16	86.98	91.07
White Noise	0 dB	92.14	69.37	80.75
	5 dB	96.17	72.74	84.45
	10 dB	97.98	86.52	92.25

TABLE 2. Experimental results of noisy man speech.

Type of noise	Input SNR	HR1(%)	HR0(%)	A(%)
Airport	0 dB	90.98	54.73	72.85
	5 dB	92.3	63.08	77.68
	10 dB	94.92	81.45	88.18
Car	0 dB	83.61	46.75	65.17
	5 dB	91.48	47.12	69.29
	10 dB	93.44	56.96	75.2
White Noise	0 dB	88.85	64.94	76.89
	5 dB	95.9	77.18	86.54
	10 dB	96.07	90.72	93.39

TABLE 3. Experimental results of sp09.wav speech.

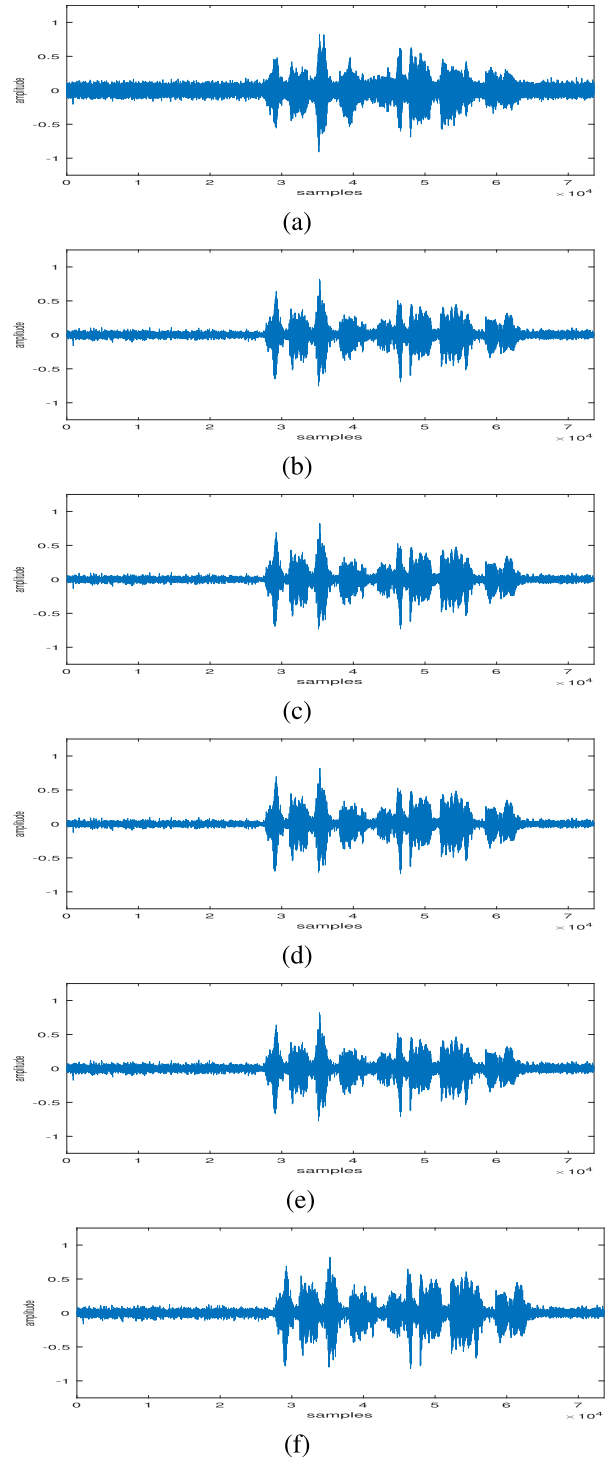
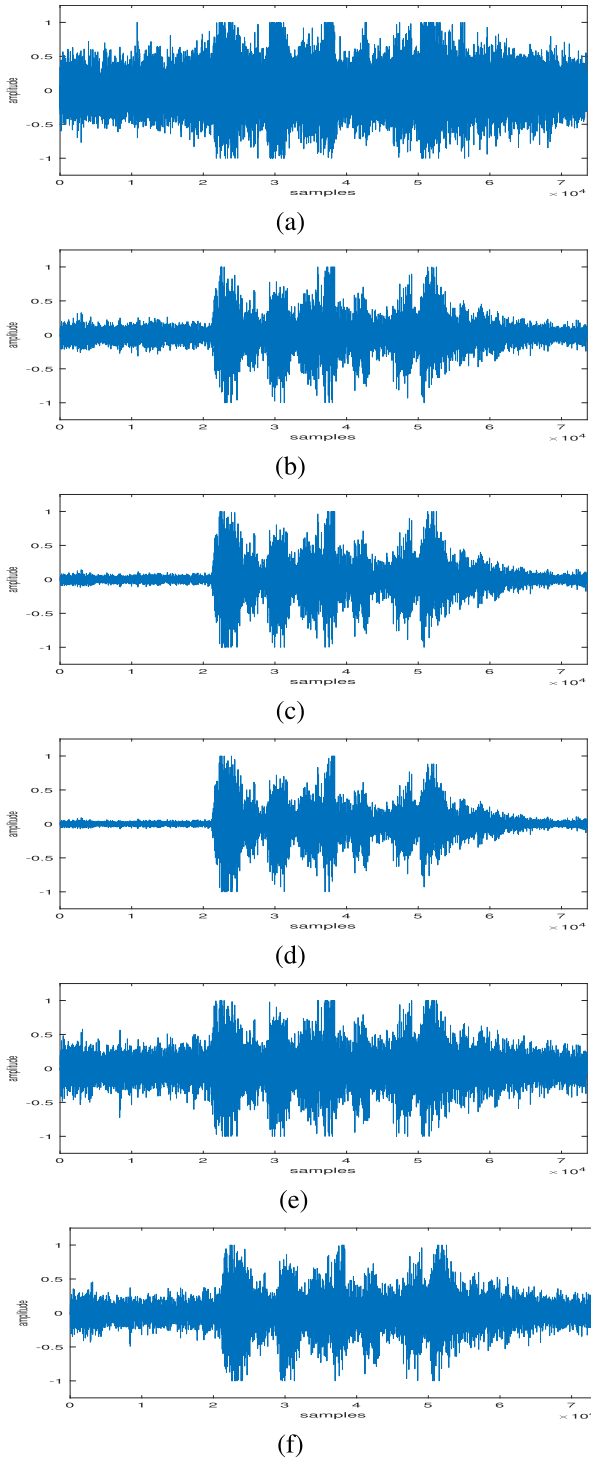
Type of noise	Input SNR	HR1(%)	HR0(%)	A(%)
Airport	0 dB	93.04	65.17	79.10
	5 dB	92.65	80.34	86.49
	10 dB	96.52	80	88.26
Car	0 dB	76.21	74.83	75.51
	5 dB	90.33	70.69	80.50
	10 dB	81.04	83.79	82.41
White Noise	0 dB	90.14	66.55	78.34
	5 dB	95.55	73.45	84.50
	10 dB	95.55	94.14	94.84

TABLE 4. Experimental results of sp11.wav speech.

Type of noise	Input SNR	HR1(%)	HR0(%)	A(%)
Airport	0 dB	87.28	59.81	73.54
	5 dB	89.02	67.76	78.38
	10 dB	98.46	78.50	88.48
Car	0 dB	83.24	65.42	74.32
	5 dB	86.32	67.29	76.80
	10 dB	91.91	82.24	87.07
White Noise	0 dB	91.33	50.47	70.89
	5 dB	99.23	64.49	81.85
	10 dB	98.27	84.11	91.18

the  $\mathbf{R}_y$  autocorrelation matrix have been calculated using the 70 frames with noisy speech.

$$\mathbf{H}_0(k, l) : \begin{cases} \mathbf{R}_y(k, l+1) = \alpha_n \mathbf{R}_y(k, l) + (1 - \alpha_n) \mathbf{Y}(k, l) \mathbf{Y}^H(k, l) \\ \mathbf{R}_y(k, l+1) = \mathbf{R}_y(k, l) = \mathbf{R}_y(k, l) \end{cases} \quad (25)$$



**FIGURE 7.** (a) Waveform of 0 dB car noisy man speech. (b) enhanced speech by wiener filter algorithm. (c) enhanced speech by SDW ( $\mu = 3$ ) filter algorithm. (d) enhanced speech by SDW ( $\mu = 5$ ) algorithm. (e) enhanced speech by MVDR filter algorithm. (f) enhanced speech by SP filter algorithm.

**FIGURE 8.** (a) Waveform of 10 dB white noisy woman speech. (b) enhanced speech by wiener filter algorithm. (c) enhanced speech by SDW ( $\mu = 3$ ) filter algorithm. (d) enhanced speech by SDW ( $\mu = 5$ ) algorithm. (e) enhanced speech by MVDR filter algorithm. (f) enhanced speech by SP filter algorithm.

$$\mathbf{H}_1(k, l) : \begin{cases} \mathbf{R}_y(k, l + 1) = \alpha_y \mathbf{R}_y(k, l) + (1 - \alpha_y) \mathbf{Y}(k, l) \mathbf{Y}^H(k, l) \\ \mathbf{R}_v(k, l + 1) = \mathbf{R}_v(k, l) = \mathbf{R}_v(k, l) \end{cases} \quad (26)$$

where  $\mathbf{H}_1(k, l)$  indicates the VAD result is 1;  $\mathbf{H}_0(k, l)$  indicates that it is 0.  $\alpha_n$  and  $\alpha_n$  are forgetting factors of noise

and noisy speech matrices and in this application they have been taken as  $\alpha_n = \alpha_y = 0,99$ . Using  $\mathbf{H}_0(k, l)$  and  $\mathbf{H}_1(k, l)$ ,  $\mathbf{R}_y$  and  $\mathbf{R}_v$  autocorrelation matrices have been calculated for each frame along the noisy speech signal according to (25) and (26). Finally, the estimated autocorrelation matrix of the



**TABLE 5. Segmental SNR results of noisy woman speech (dB).**

Type of noise	Input SNR	Wiener	SDW ( $\mu = 3$ )	SDW ( $\mu = 5$ )	MVDR	SP
Airport	0 dB	5.56	6.86	7.77	3.05	6.88
	5 dB	8.91	10.71	11.50	6.50	10.39
	10 dB	12.25	13.81	14.68	11.49	13.94
Car	0 dB	7.20	10.08	10.70	2.95	3.45
	5 dB	11.89	13.59	13.99	9.81	9.13
	10 dB	17.43	17.95	18.16	15.51	15.06
White Noise	0 dB	10.42	13.13	13.90	7.70	8.42
	5 dB	14.50	16.80	17.37	12.68	12.90
	10 dB	15.03	16.08	16.46	14.35	15.53

**TABLE 6. Segmental SNR results of noisy man speech (dB).**

Type of noise	Input SNR	Wiener	SDW ( $\mu = 3$ )	SDW ( $\mu = 5$ )	MVDR	SP
Airport	0 dB	9.38	12.87	14.12	5.12	10.04
	5 dB	13.00	15.73	16.59	10.71	13.02
	10 dB	16.55	18.34	19.09	15.22	17.25
Car	0 dB	10.63	15.90	17.78	5.51	7.60
	5 dB	15.58	17.73	18.64	9.75	10.23
	10 dB	19.47	21.20	21.90	14.93	15.81
White Noise	0 dB	10.66	12.83	13.47	8.43	8.21
	5 dB	12.20	13.89	14.45	11.19	11.25
	10 dB	15.96	16.63	16.77	15.43	16.09

clean speech signal has been calculated as shown in (27).

$$\mathbf{R}_x(k, l) = \mathbf{R}_y(k, l) - \mathbf{R}_v(k, l) \quad (27)$$

In order to evaluate the results obtained, Mean Opinion Score (MOS) and Segmental SNR have been used [30]. Segmental SNR is calculated by using (28).

$$SNR_{seg} = \frac{1}{L} \sum_{l=0}^{L-1} 10 \log_{10} \left( \frac{\sum_{k=1}^M X(lM+k)^2}{\sum_{k=1}^M [Y(lM+k) - X(lM+k)]^2} \right) \quad (28)$$

where  $M$  is the number of samples in a frame,  $L$  is number of frames,  $Y$  is noisy speech signal,  $X$  is the clean speech signal. Noisy speech signals as 0 dB, 5 dB and 10 dB have been applied to all algorithms and the results have been shown in Table 3 and Table 4. In SDW algorithm  $\mu$  constant,  $\mu = 3$  and  $\mu = 5$  has been used in two different values. Thus, five different results have been obtained for each noisy speech signal.

According to Table 5 and Table 6, the best results have been obtained SDW ( $\mu = 5$ ) algorithm. For MVDR filter, noise reduction performance is less than other algorithms. In Fig. 7 and Fig. 8, noisy speech signals and enhanced speech signals for each algorithm have been shown. In order to MOS, the noisy speech files applied to the algorithms and the enhanced speech files have been played to nine people working professionally in the field of sound at Turkish Radio Television (TRT). In the MOS evaluation, the listeners have been rated the spwoman.wav higher in all algorithms, and they have been given the highest score to the SDW algorithm

**TABLE 7. MOS results.**

Audio file	Type of noise	Wiener	SDW ( $\mu = 3$ )	MVDR	SP
spman.wav	Airport	2.90	3.47	2.97	2.77
	Car	2.71	2.88	2.74	2.59
	White Noise	2.81	2.77	2.68	2.70
spwoman.wav	Airport	3.25	3.56	3.20	2.78
	Car	3.10	3.29	3.10	2.95
	White Noise	2.81	2.79	2.73	2.61

compared to other algorithms. The averages of MOS results obtained from these nine people have been shown in Table 7.

## VI. CONCLUSION

In this study, an easy-to-implement and efficient VAD algorithm has been developed. Three basic short-time features (STE, Periodicity, and SF) have been used in the algorithm. The files generated by adding noises of various type and SNR values to the male and female voices have been tested using the proposed algorithm with MATLAB software. According to the experimental results, the best results have been obtained with white noise and quite good results have been provided in detection of the frames with speech but some parts of the algorithm need to be improved for detecting the noisy frames. This flaw can be solved by using some other features which are more robust to this condition. Using the proposed VAD algorithm, performance of some noise reduction algorithms in frequency domain is compared. Segmental SNR and MOS have been used as performance evaluation criteria. According to the results of the noise reduction algorithms, the best results have been obtained with the SDW Wiener filter algorithm.

In speech enhancement algorithms; the type of noise added to the sound signals and the gender of the people to whom the sound signals belong are factors that affect the results. The closer the frequency characteristic of noise and speech, the lower the performance of the algorithm, however; if the frequency characteristics of the noise and speech signals is different, the algorithm results give better results. For this reason, the best results were obtained in speech samples with added white noise in all algorithms. As the upper frequencies of airport noise and female voices are higher, there are less improvements in female speech samples in all algorithms. However, when the same noise was added to male speech samples, better results were obtained in speech enhancement algorithms.

As a result, it was shown that the VAD algorithm proposed in this study can be used in adaptive filter applications. According to the results; although Wiener and SDW algorithms give better results than SP and MVDR algorithms for noise reduction; SP and MVDR algorithms provide better results in terms of preserving speech quality. In this context, SP and MVDR algorithms can be given priority in

applications where speech quality is more important, and SDW algorithm can be used in applications where noise reduction is prioritized.

The future research idea is to study further improvement methods from two aspects of features and models by analyzing the differences between speech and noise.

## REFERENCES

- [1] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 273–276.
- [2] J. Benesty, M. Souden, and J. Chen, "A study of multichannel noise reduction linear filters in the time domain," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Sep. 2011, pp. 1–6.
- [3] X. Yousheng and H. Jianwen, "Speech enhancement based on combination of Wiener filter and subspace filter," in *Proc. Int. Conf. Audio, Lang. Image Process.*, Jul. 2014, pp. 459–463.
- [4] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Apr. 1997, pp. 1167–1170.
- [5] C. V. R. Rao, M. B. R. Murthy, and K. A. Sheela, "A new technique for street noise reduction in signal processing applications," in *Proc. IEEE Region 10 Conf. (TENCON)*, Nov. 2008, pp. 1–5.
- [6] J. Chen, J. Benesty, Y. Huang, and T. Gaensler, "On single-channel noise reduction in the time domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 277–280.
- [7] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time voice activity detection algorithm," in *Proc. 17th EUSIPCO*, Aug. 2009, pp. 2549–2553.
- [8] K. Sakhnov, E. Verteletskaya, and B. Simak, "Low-complexity voice activity detector using periodicity and energy ratio," in *Proc. 16th Int. Conf. Syst., Signals Image Process.*, Jun. 2009, pp. 1–5.
- [9] E. Verteletskaya and K. Sakhnov, "Voice activity detection for speech enhancement applications," *Acta Polytechnica*, vol. 50, no. 4, pp. 1–8, 2010.
- [10] M. H. Moattar, M. M. Homayounpour, and N. K. Kalantari, "A new approach for robust realtime voice activity detection using spectral pattern," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2010, pp. 4478–4481.
- [11] N. Lezzoum, G. Gagnon, and J. Voix, "Voice activity detection system for smart earphones," *IEEE Trans. Consum. Electron.*, vol. 60, no. 4, pp. 737–744, Nov. 2014.
- [12] J. Pang, "Spectrum energy based voice activity detection," in *Proc. IEEE 7th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2017, pp. 1–5, doi: 10.1109/CCWC.2017.7868454.
- [13] Y. K. Bharath, S. Veena, K. V. Nagalakshmi, M. Darshan, and R. Nagapadma, "Development of robust VAD schemes for voice operated switch application in aircrafts: Comparison of real-time VAD schemes which are based on linear energy-based detector, fuzzy logic and artificial neural networks," in *Proc. 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT)*, pp. 191–195, Jul. 2016.
- [14] T. H. Zaw and N. War, "The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection," in *Proc. 20th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2017, pp. 1–5.
- [15] N. Modhave, Y. Karuna, and S. Tonde, "Design of matrix Wiener filter for noise reduction and speech enhancement in hearing aids," in *Proc. IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol. (RTEICT)*, May 2016, pp. 1–5.
- [16] N. Modhave, Y. Karuna, and S. Tonde, "Design of multichannel Wiener filter for speech enhancement in hearing aids and noise reduction technique," in *Proc. Online Int. Conf. Green Eng. Technol. (IC-GET)*, Nov. 2016, pp. 1–4.
- [17] G. Itzhak, J. Benesty, and I. Cohen, "Nonlinear kronecker product filtering for multichannel noise reduction," *Speech Commun.*, vol. 114, pp. 49–59, Nov. 2019.
- [18] G. Li, S. Liang, S. Nie, W. Liu, Z. Yang, and L. Xiao, "Deep neural network-based generalized sidelobe canceller for robust multi-channel speech recognition," in *Proc. Interspeech*, , Shanghai, China, Oct. 2020, pp. 51–55.
- [19] J. Wu, Z. Chen, J. Li, T. Yoshioka, Z. Tan, E. Lin, Y. Luo, and L. Xie, "An end-to-end architecture of online multi-channel speech separation," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 81–85.
- [20] W. Yang, G. Huang, J. Chen, J. Benesty, I. Cohen, and W. Kellermann, "Robust dereverberation with kronecker product based multichannel linear prediction," *IEEE Signal Process. Lett.*, vol. 28, pp. 101–106, 2021.
- [21] S. M. Kim, "Auditory device voice activity detection based on statistical likelihood-ratio order statistics," *Appl. Sci.*, vol. 10, no. 15, p. 5026, Jul. 2020, doi: 10.3390/app10155026.
- [22] B. M. Mahmmod, A. R. Ramli, T. Baker, F. Al-Obeidat, S. H. Abdullhussain, and W. A. Jassim, "Speech enhancement algorithm based on super-Gaussian modeling and orthogonal polynomials," *IEEE Access*, vol. 7, pp. 103485–103504, 2019.
- [23] R. Liang, F. Kong, Y. Xie, G. Tang, and J. Cheng, "Real-time speech enhancement algorithm based on attention LSTM," *IEEE Access*, vol. 8, pp. 48464–48476, 2020.
- [24] P. Pollak, P. Sovka, and J. Uhler, "Noise suppression system for a car," in *Proc. 3rd Eur. Conf. Speech Commun. Technol., 3rd Eur. Conf. Speech Commun. Technol. (EUROSPEECH)*, Jan. 1993, pp. 1073–1076.
- [25] K. Ngo, "Digital signal processing algorithms for noise reduction dynamic range compression and feedback cancellation in hearing aids," Ph.D. dissertation, Dept. Elect. Eng., Katholieke Univ., Leuven, Belgium, 2011.
- [26] A. Spriet, M. Moonen, and J. Wouters, "Stochastic gradient based implementation of spatially pre-processed speech distortion weighted multichannel Wiener filtering for noise reduction in hearing aids," *IEEE Trans. Signal Process.*, vol. 53, no. 3, pp. 911–925, Mar. 2005.
- [27] B. Cornelis, M. Moonen, and J. Wouters, "Comparison of frequency domain noise reduction strategies based on multichannel Wiener filtering and spatial prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 129–132.
- [28] J. Capon, "High resolution frequency wave number spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
- [29] *A Noisy Speech Corpus for Evaluation of Speech Enhancement Algorithms*. Accessed: Feb. 13, 2018. [Online]. Available: <http://ecs.utdallas.edu/loizou/speech/noizeus>
- [30] R. Zelinski and P. Noll, "Adaptive transform coding of speech signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, no. 4, pp. 299–309, Aug. 1977.



**RAMAZAN ÇOLAK** received the M.Sc. degree in electronics and telecommunication engineering from Tekirdağ Namık Kemal University, Tekirdağ, Turkey, in 2019. From 2008 to 2009, he was a Lecturer with the Çorlu Vocatioanal School, Tekirdağ Namık Kemal University. He is currently working with Turkish Radio and Television Corporation, İstanbul, Turkey. His research interests include the speech enhancement, speech recognition, and speech analysis.



**RAFET AKDENİZ** received the B.S. degree in electrical-electronics from Gazi University, Ankara, Turkey, in 1985, the M.S. degree in electronics from Marmara University, in 1989, and the Ph.D. degree in electronics engineering from İstanbul University, İstanbul, Turkey, in 1998.

From 1985 to 1989, he was a Lecturer with Trakya University, Edirne, Turkey. From 1990 to 1991, he was a Visiting Instructor with Ferris State University, Big Rapids, MI, USA. From 1992 to 1998, he was a Lecturer with Trakya University, where he was also an Assistant Professor, from 1999 to 2006. From 2006 to 2015, he was an Assistant Professor with Department of Electronics and Telecommunication Engineering, Tekirdağ Namık Kemal University, Tekirdağ, Turkey, where he has been an Associate Professor, Since 2015. His research interests include signal processing, image processing, and speech coding. He received the Best Paper awards, such as BILISIM'2000, 17th National Informatics Congress, İstanbul, in September 2000, and the IEEE IC-BNMT'2010, 3rd International Conference on Broadband Network and Multimedia Technology, Beijing, China, in October 2010.

• • •