ISTANBUL UNIVERSITY CERRAHPASA

RESEARCH ARTICLE

Electrica 2019; 19(2): 91-100

Noise Robust Voice Activity Detection Based on Multi-Layer Feed-Forward Neural Network

Özkan Arslan[1], Erkan Zeki Engin[2]

[1]Department of Electronics and Communication Engineering, Tekirdağ Namık Kemal University, Tekirdağ, Turkey
[2]Department of Electrical and Electronics Engineering, Ege University, İzmir, Turkey

Cite this article as: Arslan Ö, Engin EZ. Noise Robust Voice Activity Detection Based on Multi-Layer Feed-Forward Neural Network. *Electrica*, 2019; 19(2): 91-100.

ABSTRACT

This paper proposes a voice activity detection (VAD) method based on time and spectral domain features using multi-layer feed-forward neural network (MLF-NN) for various noisy conditions. In the proposed method, time features that were short-time energy and zero-crossing rate and spectral features that were entropy, centroid, roll-off, and flux of speech signals were extracted. Clean speech signals were used in training MLF-NN and the network was tested for noisy speech at various noisy conditions. The proposed VAD method was evaluated for six kinds of noises which are white, car, babble, airport, street, and train at four different signal-to-noise ratio (SNR) levels. The proposed method was tested on core TIMIT database and its performance was compared with SOHN, G.729B and Long-Term Spectral Flatness (LSFM) VAD methods in point of correct speech rate, false alarm rate, and overall accuracy rate. Extensive simulation results show that the proposed method gives the most successful average correct speech rate, false alarm rate, and overall accuracy rate in most low and high SNR level conditions for different noise environments.

Keywords: Voice activity detection, time and spectral features, multi-layer feed-forward neural network

Corresponding Author:
Özkan Arslan

E-mail:
oarslan@nku.edu.tr

Received: 02.11.2018

Accepted: 12.03.2019

DOI: 10.26650/electrica.2019.18042

Introduction

Voice activity detection (VAD) is a signal processing technique that allows a speech signal to be seperated into voice and other parts. Non-speech regions include silence and noise signal parameters and VAD is used to effectively distinguish between speech and non-speech regions. The principal VAD applications are speech coding [1, 2], speaker and speech identification [3], noise reduction in digital hearing aid devices, mobile communications [4], and noisy speech enhancement [5, 6]. VAD methods used in many applications are based on thresholding energy, pitch detection, zero-crossing rate, linear predictive coefficients, and cepstral coefficients because of their simple structure [7-9]. However, in VAD applications using these criteria, it is difficult to detect speech regions in low signal-to-noise ratio (SNR) or non-stationary background noises. Up to recently, various features have suggested and this features including energy-entropy feature [10], cepstral feature [11, 12], and Teager energy feature [13]. In most of these methods, the noise signal is considered to be stationary over certain periods of time. For this reason, these methods are highly sensitive to SNR variations in the processed speech signal.

Voice activity detection methods are divided into three basic category based on time, frequency, and time-frequency domain. Time domain methods include energy, zero-crossing rate, and low frequency band power. The frequency domain methods include various spectral energy features which are energy-entropy, cepstral, and Teager energy features. Also, time-frequency methods include Wavelet transform and Empirical Mode Decomposition [14]. For noisy conditions, time domain, frequency domain, and time-frequency domain methods such as discrete wavelet transform, wavelet packet transform, and perceptual wavelet packet transform are insufficient in detecting speech/voice regions. In most of these methods, threshold values levels cannot be calculated accurately and precisely at all noise types and all SNR levels. Therefore, in recent years, empirical mode decomposition method, which separates speech signals

from high frequency to low frequencies into intrinsic modes, has been used in the determination of speech/voice regions [15, 16]. These traditional methods are insufficient in detecting speech/voice regions in noisy conditions.

Related Works

In recent years, various methods have been developed to detect voice or speech activity regions under different noisy conditions [17, 18]. The most known VAD approach in these methods, a statistical model for determining the voice and unvoiced regions has recommended by Sohn et. al [19]. The robust VAD algorithm, which use the likelihood ratio test based on statistical models, shows a better performance in the determination of the speech regions than the traditional methods. In this method, a speech decision rule was obtained from the generalized likelihood ratio test, assuming that noise statistics were a priority. In order to estimate time-varying noise statistics, a new noise-pattern adaptation algorithm was developed using soft decision information in the areas where the speech signal was present. This algorithm, which was designed to determine the voice regions of clean speech signals, cannot perform the same performance in speech and non-speech regions in noisy speech signals. The G729B, a modern standard VAD method, was developed for short-time frames and this method used line spectral frequencies, low band energy, zero-crossing rate, and full band energy parameters [20]. In this algorithm, a structure which was used together with the intuitively defined regions and borders was applied and decided for every 10 ms.

Most standard VAD algorithms are derived from short-term analysis frameworks and the VAD decision is made over these short-term frames. In contrast to the short-term framework for VAD, long-term spectral divergence (LTSD) was proposed as a long-term VAD analysis and a long-term spectral envelope was compared with the average noise pattern to establish a speech and non-speech decision rule [21]. In another long-term approach, a new and robust VAD algorithm was recommended using the long-term spectral flatness (LSFM) measure [22]. This new LSFM-based VAD increased the robustness of speech detection in various noisy environments using low noise spectrum estimation and adaptive threshold values. The proposed algorithm was evaluated under 12 types of noises and five types of SNR in core TIMIT test corpus. Comparisons based on the standard three VAD algorithms such as AMR1, AMR2 and G729B show that the LSFM-based VAD approach achieves the best average accuracy rate (88.95%).

Recently, a novel voice activity detection (VAD) scheme employing differential entropy at each frequency bin of power spectral estimates of past and present overlapping speech frames were proposed [23]. In this method, frequency domain long-term differential entropy (FLDE) was used to determine the speech and non-speech regions. They evaluated the performance of the proposed FLDE scheme, considering 12 types of noises and 5 different SNRs which were artificially added to speech samples from the SWITCHBOARD corpus. Graf et al. [24] mentioned about harmonicity, power, SNR, formant structure, modulation, and stationarity features of VAD algorithms. Pasad et al. [25] showed that speech and non-speech regions were difficult to detect when background sounds such as breath, mic pops, etc. were present in the signal.

In most of these methods, threshold values cannot be calculated accurately and precisely at all noise types and all SNR levels. Therefore, in recent years, the classification of speech and non-speech regions is performed through artificial neural networks. Farsinejad et al. [26]. proposed a VAD algorithm using short-time power, zero-order maximum likelihood parameter, and pitch-period difference and Probability Neural Network (PNN). In another VAD algorithm based on Support Vector Machine (SVM) [13], the determination of voice and unvoiced regions, the noisy speech signal was decomposed by Wavelet Packet Transform (WPT) and Teager energy of each wavelet packet coefficients was calculated. The energy values of noisy speech signal used as input for SVM.

In [27], fuzzy entropy and a SVM-based VAD approach have been proposed to detect voice regions in noisy conditions. In this method, the Fuzzy entropy (FuzzyEn) features obtained from noisy speech signals were used to train and test the SVM model in determining speech and non-speech regions. In various noisy conditions, speech and non-speech regions were determined with a accuracy of 93.29%. However, the performance of this method is insufficient for low SNRs. In another VAD study, Deep Neural Network (DNN) was used to determine the speech regions [18]. In this VAD approach, six different features extracted from speech signals were used to train and test the network. In VAD performance evaluation, the area under the curve (AUC) and F-measure were used and absolute increments were calculated as 10.41% and 8.56%, respectively. Similarly, a VAD approach was proposed for the speech recognition system, which allows the classification of speech and non-speech regions and DNN model was used for classification [28]. In this DNN-based VAD approach, a 22% improvement in equal error rate (EER) was achieved compared to traditional VAD approaches. Wang et al. [29] suggested a DNN-based VAD approach on the CENSREC-1-C database. They used mel-frequency cepstral coefficients, instantaneous frequency derivative, power normalised cepstral coefficient, and magnitude information and found an equal error rate of 19.44%.

A VAD method that uses line spectral frequency based statistical features with extreme learning based classification has been proposed [30]. In this study, data having a duration of more than 350 hours were used and an overall accuracy rate of 99.43% was obtained. Bouguelia et al. [31] proposed a new active learning method (ALM) that questions the label of samples based on how they affect the outcome of the classification mode. An unsupervised classification method was proposed in order to determine the presence and absence of speech regions in speech signals and this method was successful at the boundary points where the detection was quite important [32]. In another VAD approach, a VAD system has been

proposed in a multi-speaker environment. In this approach, Mahalanobis classifier and attributes of speech energy signals were taken into consideration [33]. Shi et al. [34] used a neural network-based approach with maximum short-term automatic correlation and spectrum variance for end-point detection of speech signals. In this VAD approach, 200 clean speech signals were corrupted with three different noise types such as white, babble, and car for SNR values ranging from -5 to 10 dB. The average correct rates of 93.59, 85.11 and 90.29% were obtained for the white, babble, and car noise types, respectively .

Any VAD algorithm basically consists of two parts which are 'feature extraction' and 'speech / non-speech decision mechanism'. The parameters are extracted in the first section from the speech signals that can represent the distinctive features of the speech relative to the noise. In the second part, these voice parameters are used to decide whether the region is to be speech / non-speech based on a set of decision rules.

In this study, voice activity detection algorithm was improved by using time and spectral features based on MLP-NN. The time and frequency domain features of segmented speech signal distorted by different noise type at 0, 5, 10 and 15 dB SNR levels used as input of MLP-NN. The proposed method was tested on core TIMIT database and its performance was compared with SOHN, G.729B, and LSFM VAD methods in point of correct speech rate, false alarm rate, and overall accuracy rate.

Methodology

Basic Structure of Proposed VAD

The steps of the proposed method is given in Figure 1. Firstly, noise was added to the clean speech signals and the signals were separated into frames. Then, time and spectral features were extracted for each frame. These features of noisy speech signals were applied to MLF-NN and the network was tested with speech data. Finally, the performance of the system was evaluated.

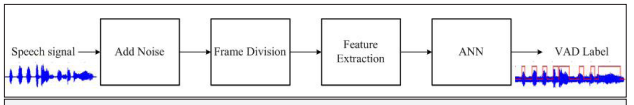

Figure 1. A schematic for the proposed method
30x3 mm (600x600 DPI)

Features for VAD

Speech signals are inherently non-stationary signals that change over time. Therefore, speech signals were processed by dividing into frames. In the proposed method, six different features per frame was used. These features were the short-time energy, zero-crossing rate, spectral entropy, roll off, flux, and centroid. These features were applied to the input of MLF-NN for training and the network was determined if the speech was either speech or non-speech.

Short-Time Energy

Short-time energy (STE) is commonly used to determine the voice and unvoiced regions. Voice regions have high energy, whereas non-speech regions have low energy values. However, this feature loses its efficiency especially in lower SNRs. In general, the STE is defined as follows [35]:

$$E_m = \sum_{k=-\infty}^{\infty} [x(k)w(m-k)]^2 \quad (1)$$

where $x(k)$ and $w(m–k)$ denote a segment of the sequence and time-shifted window sequence, respectively.

Zero-Crossing Rate

Zero-crossing rate (ZCR) is a measure of how many times the speech signal has passed through zero. In speech and non-speech regions, zero-crossing rate has lower and higher values, respectively. The ZCR is defined as [35]:

$$Z_m = \frac{1}{2}\sum_{n=0}^{N-1} \left|\mathrm{sgn}[x(n)] - \mathrm{sgn}[x(n-1)]\right| w(n-m)$$
$$\mathrm{sgn}[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases} \quad (2)$$

where $x(n)$ is the time domain signal for frame m.

Spectral Entropy

The spectral entropy is a measure of uncertainty for intrinsic characteristics of speech spectrums. Many existed experiments have proved that spectral entropy is superior to the time domain features for speech activity detection [36]. Spectral entropy has low and high values for speech and non-speech regions, respectively. Spectral entropy is calculated for each frame and each frame separated into four sub-frames. Fourier Transform of speech frame for each sub-band is defined as [37]:

$$X(k) = \sum_{n=0}^{N-1} x(n)\exp(-j\frac{2\pi nk}{N}), \quad k = 0,1,2,...,N-1 \quad (3)$$

where $X(k)$ represents the spectral magnitude of the k th frequency bin and N is the total number of frequency bins. The spectral energy of each frame $E(k)$ is expressed as:

$$E(k) = \left|X(k)\right|^2, \quad k = 0,1,2,...,N/2 \quad (4)$$

and the probability measure in the spectral domain can be written as:

$$p(i) = \frac{E(i)}{\sum_{k=1}^{N/2} E(k)}, \quad i = 0,1,2,...,N/2 \tag{5}$$

Then, the spectral entropy for each frame is calculated as:

$$H = -\sum_{i=1}^{N/2} p(i)\log[p(i)] \tag{6}$$

Spectral Centroid

Centroid is a measure of the spectral shape. The spectral centroid value in the high frequency range is greater than in the low frequency range. Therefore, the spectral centroid measure are used to identify voice and unvoiced regions. While the spectral centriod values are low in the speech regions, the spectral centroid values of the non-speech regions are high. The spectral centroid value of a signal spectrum can be expressed as follows [38]:

$$C_r = \frac{\sum_{k=1}^{N/2} f[k]|X_r[k]|}{\sum_{k=1}^{N/2}|X_r[k]|}, \quad k = 0,1,...,N-1 \tag{7}$$

where $X_r(k)$ is the magnitude of the Fourier transform at frame r and $f(k)$ is the frequency at bin k.

Spectral Roll-off

The spectral roll-off could be defined as the frequency at which 85% of the spectrum magnitude is concentrated and this value is also a measure of the spectral shape. It has higher values in the higher frequency ranges than in the low frequency ranges. Because of this property, this criterion can be used in voice activity detection. The spectral roll-off measure can be formulated as follows [38]:

$$\arg\min_{f_c \in \{0,...,N-1\}} \sum_{k=0}^{f_c} |X_r[k]| \geq 0.85 \sum_{k=0}^{N-1} |X_r[k]| \tag{8}$$

where f_c is the roll-off frequency and $X_r(k)$ is the magnitude of the *k-th* frequency component at frame *r*.

Spectral Flux

The flux could be defined as the average variation of spectrum between the two adjacent frames. The spectral flux of the noise signal has higher values than the periodic signals. Therefore, spectral flux can be used to identify the voice and unvoiced regions. Spectral flux can be expressed as follows [38]:

$$F_r = \sum_{k=0}^{N/2} (|X_r[k]| - |X_{r-1}[k]|)^2 \tag{9}$$

where $X_r(k)$ and $X_{r-1}(k)$ are the normalized magnitude of the Fourier transform at frames r and $r-1$.

Artificial Neural Network for VAD

Neural Networks have also been shown to be applicable in the area of speech detection or voice activity detection. The most commonly used neural network is the multi-layer feed-forward (MLF) neural network and it is very useful in practice since it could represent a set of very non-linear functions. A MLF neural network is a few layered structures consisting of neurons as seen in Figure 2. A MLF neural network has an input layer, one or more hidden layers, and an output layer. The most widely used network structure has one hidden layer [39]. The MLF networks could be trained with a back-propagation learning algorithm [40].

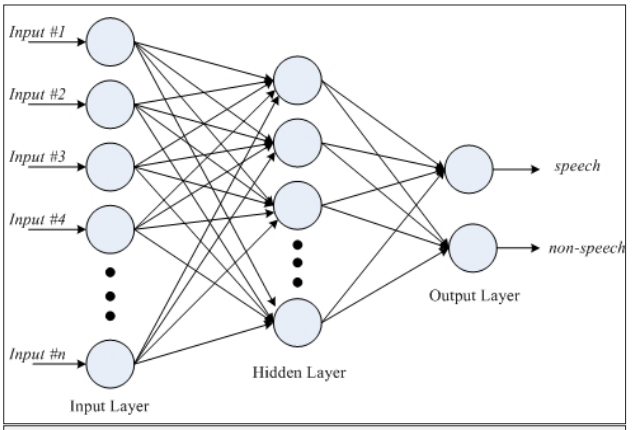

Figure 2. A structure of multi-layer feed-forward neural network
128x83 mm (96x96 DPI)

Each neuron sums its inputs from the previous layer of neurons and passes it through a transfer function. The first layer is the input layer and is completely connected to the hidden layer through a set of weights. Similarly, the hidden layer is completely connected to the output layer through another set of weights. A clean speech sequence can be applied at the input layer and the decision of speech or non-speech sequence will be obtained at the output layer for an appropriately trained neural network.

Transfer Function

Within any single processor, the summation of the incoming connection activity is performed by calculating the sum of the products of the respective weights and input activity:

$$x_j = \sum_i w_{ji} * y_i \tag{10}$$

where *j* is the index number for the processor in the layer under consideration, y is the output level of the th processor in the previous layer, and w_{ji} *is* the weight for the connection between the *i* th and *j* th processors. The result of this weighted summed process is the activation level, X_j, which is then operated on by the transfer function, *f(x)*, to produce the new output for that processor, $Z_j = f(x_j)$. The transfer function is usually chosen to be non-linear to assist in the formation of a complex representative space. The most commonly used transfer function types are sigmoid or hyperbolic functions. The following log-sigmoid transfer function was chosen and is defined as:

$$f(x) = \frac{1}{1+\exp(-x)} \tag{11}$$

Training and Generalization

The MLF neural network is based on a dual working principle of training and prediction. There is a need of two data sets which are train and test and optimal weights are determined in during the training phase. In prediction (test) mode, network generates an output based on input values with its trained network [40].

Experimental Results

Dataset

The proposed algorithm was tested on English sentences which were taken from TIMIT database [41]. TIMIT corpus consists of 24 individual speakers (16 males and 8 females) of eight different dialects, each with 10 phonetically balanced English sentences which are approximately 3-seconds long. The speech records in the TIMIT database are sampled with 16 kHz. In many studies, the TIMIT sentences were downsampled to 8 kHz [42-45]. In addition, as the sampling rate of the G729B and Sohn VAD was 8 kHz, all data was resampled to 8 kHz for a fair comparison [46]. The TIMIT database consists of 720 different speech samples and 432 and 288 utterances were used for training and testing, respectively. The clean speech signals were corrupted by white, car, babble, airport, street, and train noises from the NOISEX-92 noise database [47].

Performance Evaluation

The proposed VAD algorithm was evaluated according to correct detection P_d, false alarm P_f, and the overall accuracy rate (Acc). To obtain P_d and P_f rates, the speech (voice) and non-speech (silence or noise) regions of clean speech signals were primarily manually labeled. The speech correct hit rate P_d is calculated as the rate of correctly detected labeled speech regions by the VAD algorithm, while the false alarm P_f is calculated as the rate of non-speech is identified as speech. The correct speech hit rate (P_d) and the false alarm (P_f) can be expressed as:

$$P_d = \frac{N_{1,1}}{N_{1,1}^{ref}} \qquad P_f = \frac{N_{0,1}}{N_{1,1}^{ref}} \tag{12}$$

where $N_{1,1}$ represents correctly classified number of speech samples by VAD methods, $N_{0,1}$ represents non-speech samples which are erroneously identified as speech samples and $N_{0,0}$ represents correctly classified number of non-speech samples by VAD methods. N_1^{ref} and N_0^{ref} represent correct speech and non-speech samples that manually labeled samples taken from the database, respectively. The overall accuracy rate (Acc) is calculated as:

$$Accuracy = \frac{N_{1,1} + N_{0,0}}{N_1^{ref} + N_0^{ref}} \tag{13}$$

Since there are always trade-off relationships among these two metrics, we use the mean of overall accuracy, P_d and P_f as the final metric for better performance comparison. For an ideal VAD algorithm, the speech or non-speech correct detection rate and overall accuracy rate should be maximized, while the false alarm rate should be minimized.

Implementation

The time resolution is limited in determining the speech regions and is much lower than the sampling rate of the speech signal. Therefore, the decision rule for determining the regions of speech was not calculated for each sample of the signal. Instead, the speech signals were processed by dividing them into short frames [24]. In this study, the speech signals were divided into overlapping frames by using Hamming window. As with most VAD applications, the duration of overlapping frames were 20 ms, which implies that each frame consists of 160 samples, and shifted by 10 ms between frames [45, 48, 49]. In TIMIT database, speech frames (segments) were manually marked for identifying speech or non-speech regions. In this study, Gradient Descent (GD), Gradient Descent with Adaptive (GDA), the Broyden–Fletcher–Goldfarb–Shanno (BFGS) Quasi-Newton and Levenberg-Marquardt (LM) learning algorithms were used to train the MLF neural network. The learning algorithm was used in the training of the network, since there was a very small difference between the minimum error values reached by these learning algorithms (approximate error 10^{-6}) and the same error value was reached with less iteration by the LM algorithm. Log-sigmoid transfer function was chosen as the activation function in MLF neural network structure, which was trained with Levenberg-Marquardt learning algorithm.

In this study, the input layer consists of 6 neurons, the single hidden layer consists of 15 neurons and the output layer consists of 2 output neurons. The most appropiate number of neurons in the hidden layer were chosen by testing different

combinations. In the MLF-NN network, the first neuron of the output layer yields the speech regions (voice) and the other output neuron yields the non-speech regions. The weights were initially adjusted to some small random values close to zero and these weights were updated during the training of the network and the output produced was matched with the correct output. Only clean speech signals were used in training and the signals used in training were not used in the test. The estimated VAD results of the network output were compared with manually labeled VAD results.

Figure 3 shows the short-time energy, zero-crossing rate (ZCR), spectral entropy (SE), spectral flux (SF), spectral centroid (SC), and spectral roll-off (SR) acoustic parameters of segmented noisy speech in white noise condition. In the determination of speech or voice segments, high energy, low ZCR, SE, SF, SC, and SR values were used.

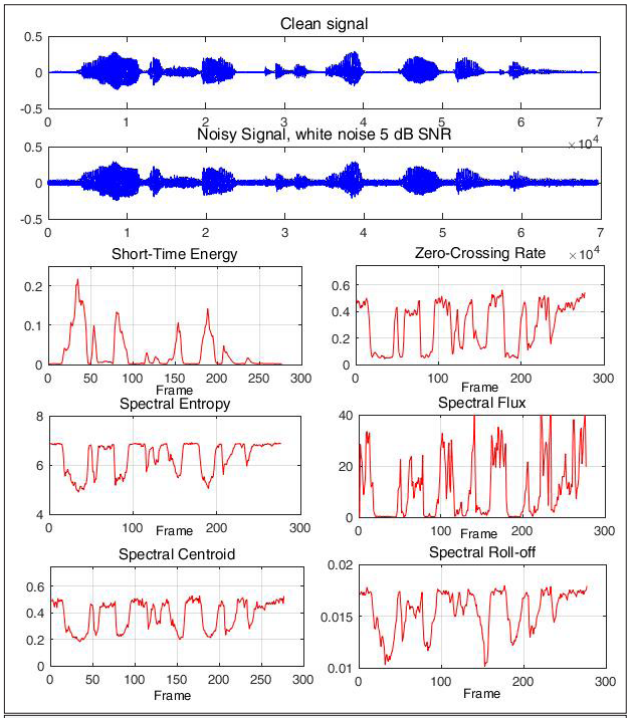

Figure 3. Results of short-time energy, ZCR, SE, SF, SC and SR for speech signal corrupted by white noise at 5 dB SNR level
240x245 mm (72x72 DPI)

The proposed, LSFM, SOHN and G729B VAD algorithms with a clean speech ("He picked up the dice for a second roll") and its noisy version (SNR = 0 dB) is shown in Figure 4. First, the voice regions of noisy speech signal was manually labeled. Then, the neural network structure was trained with clean speech signals and the network were tested for noisy signals. In Figure 4, the voice and unvoiced segments were determined with 92.14% accuracy rate for the proposed VAD algorithm.

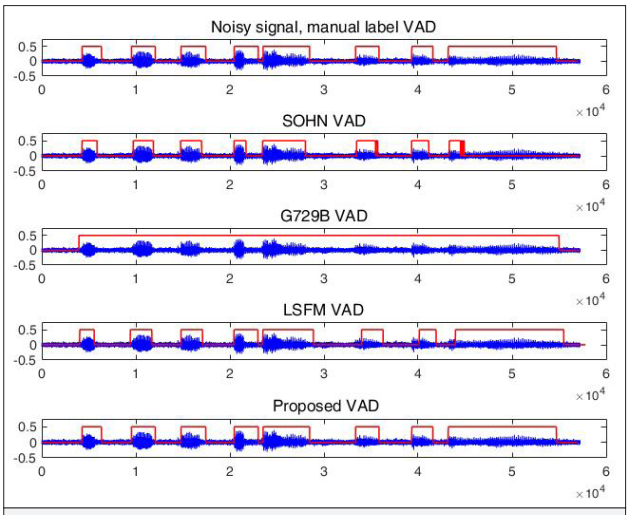

Figure 4. From top to bottom: Clean signal, Noisy signal, Manually labeled VAD, SOHN VAD, G729B VAD, and proposed VAD method results for 0 dB SNR white noise conditions
270x203 mm (72x72 DPI)

The accuracy performances of SOHN, G729B, LSFM and the proposed VAD algorithms in six different noise types and four different SNR level conditions are given in Table 1. As given in the table, although the correct detection rate of the speech regions using SOHN and G729B VAD algorithms is high, the false alarm rate and overall accuracy rates are not reasonable. This means that in traditional VAD methods, the speech regions are determined as non-speech regions. The recommened VAD algorithm plays an important role to overcome this problem. Also, LSFM VAD is able to determine speech regions with high accuracy, especially at low SNR levels. The proposed method has effective and successful results in terms of high correct speech rate (P_d), low false alarm (P_d), and high overall accuracy rate compared with other methods for various noisy environments.

Conclusion

This article has proposed a voice activity detection method based on time and spectral features using multi-layer feed-forward neural network classification. In traditional VAD methods, it is necessary to determine an appropriate threshold value to distinguish between voice and other regions. In this way, VAD methods based on the threshold value may be insufficient in noisy environments. In order to overcome this problem, VAD method based on a classification algorithm was used to distinguish between speech and non-speech regions in both clean and noisy environments. We suggested time and spectral based features for speech/non-speech decisions. The time domain features which are short-time energy and zero-crossing rate and spectral domain features which are entropy, centroid, flux, roll-off and a well-trained MLF-NN were utilized to identify and label parts that are speech and non-speech. Finally, the recommended method was objectively evaluated in six kinds of noises at different SNR levels. The suggested approach was also compared with LSFM, SOHN, and G729B VAD techniques

Table 1. The correct speech rate P_d, false alarm P_f and overall accuracy rate of the proposed, Long-Term Spectral Flatness (LSFM), SOHN and G729B VAD for different noisy environments

		Proposed VAD			LSFM VAD			SOHN VAD			G729B VAD		
Noise	**SNR**	**P_d**	**P_f**	**Acc**	**P_d**	**P_f**	**Acc**	**P_d**	**P_f**	**Acc**	**P_d**	**P_f**	**Acc**
	15	99.32	2.64	96.22	96.52	3.61	92.85	86.85	16.60	85.92	99.97	41.12	58.16
White	10	97.58	2.32	95.52	94.68	3.25	91.05	81.83	9.41	84.00	99.93	42.19	57.73
	5	94.00	1.97	93.70	92.72	2.56	89.32	71.17	3.16	81.43	97.90	42.38	57.75
	0	91.54	1.20	90.11	91.68	1.92	88.17	52.28	2.69	71.75	95.73	42.40	57.68
	15	98.38	2.62	95.43	96.39	3.48	92.30	85.89	27.60	72.61	99.92	42.36	57.79
Babble	10	97.02	2.98	93.78	94.81	4.27	91.53	79.64	27.30	70.63	98.90	42.35	57.77
	5	93.18	6.82	89.22	93.35	4.82	88.14	68.46	28.70	65.64	98.80	42.32	57.82
	0	90.15	8.85	83.71	91.05	6.16	84.72	51.97	32.60	57.75	96.60	42.34	57.76
	15	98.01	2.53	93.06	95.71	3.78	91.75	79.06	22.08	74.61	99.90	38.60	62.12
Car	10	97.95	2.55	91.77	93.60	4.16	88.66	63.60	19.90	69.32	99.72	40.23	61.68
	5	94.72	6.50	87.89	90.45	7.03	84.36	61.62	19.21	60.12	98.10	40.81	60.25
	0	90.80	9.14	80.73	87.12	11.08	75.26	58.99	20.20	50.29	95.21	40.68	60.12
	15	97.74	2.91	93.20	93.19	8.42	89.14	81.08	26.32	71.98	99.91	40.95	60.15
Train	10	96.06	4.96	90.41	91.95	9.61	86.90	68.35	25.20	67.89	99.70	40.30	61.15
	5	94.57	5.57	80.10	90.02	10.15	83.64	58.77	26.20	59.86	98.86	39.48	62.22
	0	91.44	6.54	78.32	88.83	11.78	80.36	55.56	31.82	49.60	96.00	38.78	62.61
	15	98.71	1.57	91.52	95.79	5.71	90.13	82.78	27.74	71.44	99.91	41.99	58.43
Airport	10	98.67	1.60	90.79	94.36	6.08	87.91	71.01	27.60	67.27	99.83	41.91	58.54
	5	97.52	2.78	86.41	91.97	8.13	83.27	62.76	29.93	64.45	98.60	41.78	58.72
	0	93.89	6.13	78.63	89.31	8.86	80.49	51.27	24.49	59.81	95.12	41.67	58.83
	15	98.43	1.99	93.21	96.34	2.76	91.53	75.54	35.35	72.43	96.97	41.94	60.65
Street	10	98.44	1.92	91.61	95.05	2.91	89.28	71.98	39.75	68.55	95.92	42.55	59.12
	5	95.11	5.85	87.03	93.72	4.03	85.75	70.31	33.46	61.54	94.74	47.72	58.94
	0	92.02	8.01	77.08	91.24	7.12	80.46	63.91	26.87	51.42	92.70	48.35	58.32

VAD: voice activity detection; LSFM: Long-Term Spectral Flatness; SNR: signal-to-noise ratio

for better ratification of its accomplishments and capabilities. Although only clean speech signals were used in training and the training duration of the MLF network is long, this VAD algorithm based on time and spectral speech parameters can be recommended for high probabilities of voice activity, high accuracy rate and low probabilities of false-alarm in determining the regions of speech or non-speech activity. Experimental results show that the proposed algorithm performs well in most low and high SNR level conditions for different noise environments.

Peer-review: Externally peer-reviewed.

Conflict of Interest: The authors have no conflicts of interest to declare.

Financial Disclosure: The authors declared that the study has received no financial support.

References

1. D. Freeman, G. Cosier, "The voice activity detector for the Pan-European digital cellular mobile telephone service", Acoustics, Speech, and Signal Processing, Glasgow, UK, 1989, pp. 369-372.

2. L. Zhang, Y. C. Gao, Z. Z. Bian, C. Lu, "Voice activity detection algorithm improvement in adaptive multi-rate speech coding of 3GPP", International Conference on Wireless Communications, Networking and Mobile Computing, Wuhan, China, 2005, pp. 1257-1260.
3. L. Karray, A. Martin, "Towards improving speech detection robustness for speech recognition in adverse conditions", Speech Communication, vol. 40, no. 3, pp. 261-276, May, 2003. [CrossRef]
4. A. Sangwan, R. Sah, R. V. Prasad, V. Gaurav, "VAD Techniques", Time, pp. 46-50, 2002.
5. K. Woo, T. Yang, K. Park, C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum", Electronics Letters, vol. 36, no. 2, pp. 180-181, Jan, 2000. [CrossRef]
6. Y. Zhang, Z. Tang, Y. Li, Y. Luo, "A Hierarchical Framework Approach for Voice Activity Detection and Speech Enhancement", Scientific World Journal, vol. 2014, no. 2014, May, 2014. [CrossRef]
7. B. V. Ilarsha, "A noise robust speech activity detection algorithm," International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, China, 2004, pp. 322-325. [CrossRef]
8. C. Babu, P. Vanathi, "Performance analysis of voice activity detection algorithms for robust speech recognition", International Journal of Computing Science and Communication Technologies, vol. 2, no. 1, pp. 288-293, 2009.
9. R. G. Bachu, S. Kopparthi, B. Adapa, B. D. Barkana, "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal", Advanced Techniques in Computing Sciences and Software Engineering, pp. 279-282, 2010. [CrossRef]
10. P. Khoa, "Noise robust voice activity detection", MSc Thesis, pp. 77, 2012.
11. J. A. Haigh, J. S. Mason, "Robust voice activity detection using cepstral features", Proceedings of TENCON '93. IEEE Region 10 International Conference on Computers, Communications and Automation, Beijing, China, 1993.
12. K. Chung, S. Y. Oh, "Voice Activity Detection Using an Improved Unvoiced Feature Normalization Process in Noisy Environments", Wireless Personal Communications, vol. 89, no. 3, pp. 1-13, 2015. [CrossRef]
13. S. H. Chen, H. Te Wu, Y. Chang, T. K. Truong, "Robust voice activity detection using perceptual wavelet-packet transform and Teager energy operator", Pattern Recognition Letters, vol. 28, no. 11, pp. 1327-1332, 2007. [CrossRef]
14. S. Chen, R. C. Guido, T. Truong, Y. Chang, "Improved voice activity detection algorithm using wavelet and support vector machine," Computer Speech & Language, vol. 24, no. 3, pp. 531-543, 2010. [CrossRef]
15. C. Z. Chong Feng, "Voice activity detection based on ensemble empirical mode decomposition and teager kurtosis," 12th International Conference on Signal Processing (ICSP), Hangzhou, China, 2014, pp. 455-460.
16. Y. Kanai, S. Morita, M. Unoki, "Concurrent processing of voice activity detection and noise reduction using empirical mode decomposition and modulation spectrum analysis", Proceedings of the Annual Conference of the International Speech Communication Association, Lyon, France, 2013, 742-746.
17. M. Sahidullah, G. Saha, "Comparison of Speech Activity Detection Techniques for Speaker Recognition", arXiv, no. arXiv:1210.0297, pp. 1-7, 2012.
18. G. Ferroni, R. Bonfigli, E. Principi, S. Squartini, F. Piazza, "A Deep Neural Network approach for Voice Activity Detection in multi-room domestic scenarios", International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 2015. [CrossRef]
19. J. Sohn, "A statistical model-based voice activity detection", IEEE Signal Processing Letters, vol. 6, no. 1, pp. 1-3, 1999. [CrossRef]
20. A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, J. P. Petit, "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, 1997. [CrossRef]
21. J. Ramírez, J. C. Segura, C. Benítez, Á. De la Torre, A. Rubio, "Efficient voice activity detection algorithms using long-term speech information", Speech Communication, vol. 42, no. 3-4, pp. 271-287, 2004. [CrossRef]
22. Y. Ma, A. Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure", EURASIP Journal on Audio, Speech, and Music Processing, vol. 21, pp. 1-18, 2013. [CrossRef]
23. D. Ghosh, R. Muralishankar, S. Gurugopinath, "Robust voice activity detection using frequency domain long-term differential entropy", Interspeech, pp, 1220-1224, Sept, 2018. [CrossRef]
24. S. Graf, T. Herbig, M. Buck, G. Schmidt, "Features for voice activity detection: a comparative analysis", EURASIP Journal on Advances in Signal Processing, vol. 2015, no. 1, pp. 91, 2015. [CrossRef]
25. A. Pasad, K. Sabu, P. Rao, "Voice activity detection for children's read speech recognition in noisy conditions", 2017 23rd National Conference on Communications (NCC), Chennai, Indi, 2017. [CrossRef]
26. M. Farsinejad, M. Mohammadi, B. Nasersharif, A. Akbari, A. Framework, "A Model-based Voice Activity Detection Algorithm using probabilistic neural networks", Computer Engineering, vol. 326, pp. 8-11, 2008.
27. R. Johny Elton, P. Vasuki, J. Mohanalin, "Voice Activity Detection Using Fuzzy Entropy and Support Vector Machine", Entropy, vol. 18, no. 8, pp. 298, 2016. [CrossRef]
28. F. Bie, Z. Zhang, D. Wang, T. F. Zheng, "DNN-based Voice Activity Detection for Speaker Recognition", CLST Technical Report, pp. 1-11, 2015.
29. L. Wang, K. Phapatanaburi, Z. Oo, S. Nakagawa, M. Iwahashi, J. Dang, "Phase Aware Deep Neural Network for Noise Robust Voice Activity Detection", Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10-14 July, 2017, pp. 1087-1092. [CrossRef]
30. H. Mukherjee, S. M. Obaidullah, K. C. Santosh, S. Phadikar, K. Roy, "Line spectral frequency-based features and extreme learning machine for voice activity detection from audio signal", International Journal of Speech Technology, vol. 21, no. 4, pp. 1-8, 2018. [CrossRef]
31. M. R. Bouguelia, S. Nowaczyk, K. C. Santosh, A. Verikas, "Agreeing to disagree: active learning with noisy labels without crowdsourcing", International Journal of Machine Learning and Cybernetics, vol. 9, no. 8, pp. 1307-1319, 2018. [CrossRef]
32. Z. Ali, M. Talha, "Innovative Method for Unsupervised Voice Activity Detection and Classification of Audio Segments", IEEE Access, vol. 6, pp. 15494-15504, 2018. [CrossRef]
33. L. K. Hamaidi, M. Muma, A. M. Zoubir, "Robust distributed multi-speaker voice activity detection using stability selection for sparse non-negative feature extraction", 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, 2017, pp. 161-165, 2017. [CrossRef]
34. Y.Q. Shi, R.W. Li, S. Zhang, S. Wang, X.Q. Yi, "A speech endpoint detection algorithm based on BP neural network and multiple features", Applied Mechanics, Mechatronics and Intelligent Systems, pp. 393-402, 2016.
35. M. Jalil, F. A. Butt, A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals", 2013 The

International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), Konya, Turkey, 2013, pp. 208-212. [CrossRef]

36. M. H. Moattar, M. M. Homayounpour, "A weighted feature voting approach for robust and real-time voice activity detection", ETRI Journal, vol. 33, no. 1, pp. 99-109, 2011. [CrossRef]
37. P. Renevey, A. Drygajlo, "Entropy based voice activity detection in very noisy conditions", Seventh European Conference on Speech Communication and Technology, 2001.
38. M. N. Stolar, M. Lech, S. J. Stolar, N. B. Allen, "Detection of Adolescent Depression from Speech Using Optimised Spectral Roll-Off Parameters", Biomed J Sci & Tech Res, vol. 5, no. 1, pp. 1-10, 2018. [CrossRef]
39. M. Hill, "Notes on Multilayer, Feedforward Neural Networks Fall 2007", pp. 1-7, 2007.
40. D. Svozil, V. Kvasnička, J. Pospíchal, "Introduction to multi-layer feed-forward neural networks", Chemometrics and Intelligent Laboratory Systems, vol. 39, no. 1, pp. 43-62, 1997. [CrossRef]
41. V. Zue, S. Seneff, J. Glass, "Speech database development at MIT: Timit and beyond," Speech Communication, vol. 9, no. 4, pp. 351-356, 1990. [CrossRef]
42. P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum", IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 857-869, 2005. [CrossRef]
43. T. Drugman, Y. Stylianou, Y. Kida, M. Akamine, "Voice Activity Detection: Merging Source and Filter-based Information", IEEE Signal Processing Letters, vol. 23, no. 2, pp. 252-256, 2016. [CrossRef]
44. N. Dhananjaya, B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs", IEEE Signal Processing Letters, vol. 17, no. 3, pp. 273-276, 2010. [CrossRef]
45. N. Lezzoum, G. Gagnon, J. Voix, "Voice activity detection system for smart earphones", IEEE Transactions on Consumer Electronics, vol. 60, no. 4, pp. 737-744, 2014. [CrossRef]
46. D. Ying, Y. Yan, J. Dang, F. K. Soong, "Voice activity detection based on an unsupervised learning framework", IEEE Transactions on Audio, Speech and Language Processing, vol. 19, no. 8, pp. 2624-2632, 2011. [CrossRef]
47. A. Varga, H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems", Speech Communication, vol. 12, no. 3, pp. 247-251, 1993. [CrossRef]
48. M. H. Moattar, M. M. Homayounpour, N. K. Kalantari, "A new approach for robust realtime Voice Activity Detection using spectral pattern", Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, pp. 4478-4481, 2010. [CrossRef]
49. S. Dwijayanti, K. Yamamori, M. Miyoshi, "Enhancement of speech dynamics for voice activity detection using DNN", Eurasip Journal on Audio, Speech, and Music Processing, vol. 2018, no. 1, 2018. [CrossRef]

Özkan Arslan received B.Sc. degree in Electronics and Communication Engineering from Süleyman Demirel University in 2010, M.Sc. and Ph.D degree in Electrical and Electronics Engineering from Ege University in 2014 and 2018, respectively. He is currently Research Assistant at Electronics and Communications Engineering Department of Tekirdağ Namık Kemal University. His research interests include biomedical signal processing, voice/speech processing/analysis and machine learning applications.

Erkan Zeki Engin received B.Sc., M.Sc. and Ph.D. degree in Electrical and Electronics Engineering from Ege University, 2000, 2003 and 2010, respectively. He is currently Assistant Professor at Electrical & Electronics Engineering Department of Ege University. His research interests include biomedical signal and image processing and speech processing/analysis.