International Journal On Engineering Technology and Sciences – IJETS
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume VI, Issue V, May - 2019

Voice Activity Detection in the Speech Signal

Arun K.B
Dept. of Electronics and communication Engineering
Vimal Jyothi Engineering College
Kannur, Kerala, India
arunabc1994@gmail.com

Manoj K C
Dept. of Electronics and communication Engineering
Vimal Jyothi Engineering College
Kannur, Kerala, India
kcmanojkc@vjec.ac.in

Abstract

Voice activity detection (VAD) is used as a pre-processing step for speech enhancement, speech recognition, and speech transmission. VAD is described as the issue of differentiating speech from noise or silence. The variety and the varying nature of background noise and speech make it exclusively a challenging task. In the previous few years, many features indicating the differences between speech and noise have been proposed for their robustness. In this paper, a comparative study has been implemented to classify various VAD methods to get a clear idea of the better VAD method.

Index Terms- Voice activity detection (VAD), Mel-frequency cepstral coefficients (MFCC), SNR, Neurograms, Teager-Kaiser Energy (TKE) operator

1. Introduction

Voice is considered as the best common messenger of human and is expected to turn into the major form of upcoming computer-human communication [1]. Speech signal is a sequence of successive portions of speech and silence [2]. VAD is used for detecting the speech or non-speech events in speech signals. In a speech signal corrupted by noise, silence regions are considered as noisy part. In speech processing applications, such as speaker recognition, speech enhancement, and speech recognition VAD is used as a pre-processing step for speaker recognition, speech enhancement and speech recognition. When the speech signal is distorted by noise, VAD accuracy is extremely degraded. Therefore, in order to improve the performance of VAD and ensuing speech processing applications, a reliable VAD algorithm with robustness against noise is required [3].

In the speech recognition process first, we have to deal with the audio data. VAD is a crucial step in speech recognition. The goal of VAD is to obtain the appropriate audio segment, which means the section between the start-endpoints, from the impending audio signal. If we can detect the start points and end points precisely from the speech signal, the extent of data and computation from the ensuing model building and feature extraction steps can be reduced [4].

Hearing aids or hands-free telephony make use of audio enhancement algorithms. These applications depend on features of noise, predicted in time intervals when only noise is present (speech is absent). Robust voice activity detection is essential for removing speech segments from the noise estimates and to shorten the artifacts generated because of intrusive noise contraction in the audio signal. In order to ensure simultaneousness between input and output signal, latencies have to be kept small. Additionally, the capabilities of the hardware are limited, so the usage of memory and CPU have to be scaled correspondingly., In mobile networks, Speech transmission is mainly concentrated on speech sections. Less data is transferred, during pauses and instead, comfort noise is included [5].

Voice activity detection system consists of a 'feature extraction' section and a 'speech/ non-speech judgment' section. The extraction part represents the discriminative attributes of speech comparing to noise from the speech signal by extracting features. The decision part uses these features, with a set of decision criteria to make a decision on whether the speech segment is speech/non-speech.

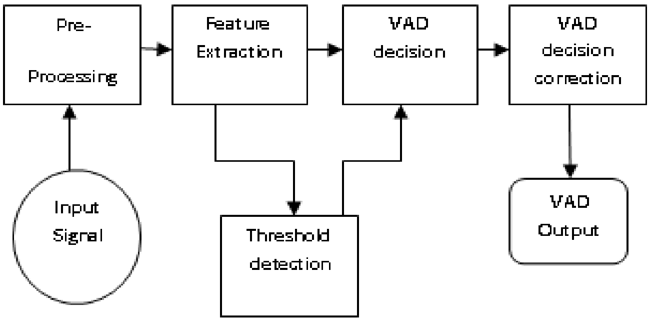

Figure1: Block diagram of VAD

Energy, spectral flatness, periodicity, spectral distance, Linear Predictive Coding (LPC), cepstral coefficients and zero crossing rate are some of the features which are generally used in VAD techniques. However, these features are sensitive to the Signal to Noise Ratio (SNR) value; and the detection accuracy degrades in accordance with the decrease in SNR value. So, different algorithms have been developed for VAD that uses various features to compromises between accuracy, computational cost, latency, and sensitivity. Various voice activity detection methods usually make use of a combination of different VAD features in order to retain the VAD accuracy [6].

2. Literature review

Hongzhi Wang et al. proposed a method for VAD, based on the similarity of Mel-frequency cepstral coefficients (MFCC) [1]. In this method first, the voice signal is divided into a number of frames. Then, MFCC is extracted for every frame. Later, MFCC Euclidean distance and MFCC correlation coefficient of the background noise and the test frame is found to compare the similarity between them. Then MFCC similarity curve is obtained, which is used to compute start- end boundaries of the noisy speech.

The MFCC extraction can be achieved by either using LPC (Linear Predictive Coding) or FFT (Fast Fourier Transform) [7]. In this paper, FFT is used for MFCC extraction. MFCC extraction steps consist of Pre-emphasis, Frame blocking, Windowing, Fast Fourier Transform (FFT), Triangular window band-pass filter, and Discrete Cosine Transform (DCT). Pre-emphasis is used to eliminate the effect of lips and vocal cords in the sound system and remunerate for the suppressed high-frequency section. The purpose of the hamming window is to reduce the discontinuity in both sides of the frame. The Output energy of each filter is computed using Triangular window band-pass filter. MFCC is obtained by taking DCT on the logarithmic energy of the triangular window band-pass filter outputs. After calculating MFCC vector for all frames first, ten frames of the audio data are assumed as background noise for finding the initial estimate of MFCC vector of background noise by calculating mean of the MFCC vector for the first ten frames. Then the MFCC similarity curve is obtained by calculating the similarity between the MFCC vector of the noise frame and test frame. This MFCC similarity curve is used to find the boundaries of the speech sections.

Mel Frequency Cepstral Coefficients called MFCC is one of the most widely used spectral based parameter in the speech recognition process [8]. MFCCs are obtained from the cepstral depiction of the audio signal and these coefficients collectively make up a Mel frequency cepstrum. When the frequency is less than 1 kHz, perception ability is linear and is logarithmic, otherwise. Specifically, the human ear's perception is rather ambiguous to high-frequency components and sensitive to low-frequency components. Mel frequency reflects hearing attributes of ear by converting spectrum into non-linear spectrum and then convert to the spectrum domain depending on the Mel frequency coordinates, [1].

Actual and Mel frequency are related by:

$$\mathrm{Mel}(f) = 2595\lg(1 + f / 700)$$

The results show that in noisy environments, MFCC similarity based method is superior to traditional methods (zero crossing rate and short-term energy method). But at low SNR circumstances, the detection accuracy reduces. The suggested method has enough area for advancement since an improved similarity measure method can substantially enhance the algorithm.

Soudeh A. Khoubrouy et al. introduced a VAD method using Teager-Kaiser Energy Measure [9]. This method makes use of a modified Teager-Kaiser Energy (TKE) operator, which is selected as a feature for VAD. The modified Teager-Kaiser Energy (TKE) operator is a combination of TKE operators of various orders.
Generalized TKE operator is defined as:

$$TKE_k(n) = x^2(n) - x(n+k)x(n-k)$$

In order to conserve the easiness of the suggested TKE method, two different TKE operators with distinct resolution parameters are designed, i.e. one TKE operator is with the resolution parameter, k=1 and the other one is with k=10. k=1 is selected to emphasize the higher frequencies and to facilitate the detection of high- frequency part of the audio data. k=10 is chosen to emphasize on the lower frequencies and to enhance the detection of low-frequency parts of the audio data.

In this method, first Teager-Kaiser Energy (TKE) is computed for each frame using the generalized TKE operator. Then the first few frames are assumed as background noise, in order to compute the Teager-Kaiser Energy (TKE) of noise. For frequencies in between 2 kHz and 6 kHz, k=1 is generally chosen and for the other frequency values, k=10 is selected. Then the mean value of TKE of the noise frame is calculated and is chosen as a threshold. Then the energy of each test frame is compared with this threshold. If the energy of the test frame is greater than the threshold value, it is considered as a speech frame, other ways it is considered as the noise frame.

Under various Signal to Noise Ratios (SNRs), modified TKE operator shows good robustness and performance in voice activity detection from noisy speech signal. Simulation results indicate that the modified TKE feature outperforms the energy feature (usually applied in VAD methods) in terms of accuracy. Also, the TKE operator has less dependency on the background noise. Although it provides it shows better performance than MFCC based VAD method when SNR is too low, its detection accuracy degrades.

Sun Yiming et al. proposed a method, known as the improved dual-threshold method for VAD. This method uses zero crossing rate and short time average energy for processing the speech signal [4]. These two methods are time domain parameters used to process the speech signal. A method called dual threshold method, which depend on these methods, is further commonly used. In the dual-threshold method, the start points and end points of the audio data are determined by combining the zero-crossing rate and short-term energy methods to form appropriate spatial or temporal features [11].

The traditional dual-threshold method for VAD consists of partition of the speech signal into a number of frames and computation of the short-time zero-crossing rate and the short time average energy for each frame. From the analysis of these parameters, a value as T1 is determined as threshold, for the detection of the starting point of the voice segments. Then, T2 is established as a threshold for the detection of the ending point of the VAD by analyzing the noise using short time average energy method. Then the short time average zero rate of noise is calculated and set up a third threshold value T3 to find out the exact start point and end point position of the audio data. The simulation results show that the traditional dual-threshold method is detected with a number of serious errors. To overcome the limitations of the traditional method, improved dual-threshold method is introduced. In this method, the threshold value T3 is reduced and then the zero-crossing rate part is found which values lower than T3. Also, particularly, noise in the audio data is suppressed by spectral subtraction method. Then, voice of short-time average energy and zero crossing rate is calculated. Finally, these parameters are smoothened using a median filter.

In the simulation, vehicular speech overlapped with the noisex92 library is used with different signal-to-noise ratios. Low error rate is detected in improved dual-threshold method, as compared to traditional dual-threshold method of VAD. The main disadvantage of this method is, when the Signal to Noise Ratio (SNR) is quite low, performance with the white noise is poor.

Jing Pang proposed a unique VAD method, named Spectrum Energy Based VAD [6]. This method makes use of the total spectrum energy in the overlapping speech window frames. Generally, noise energy is mainly concentrated in higher frequency band. In this method, noise energy in the higher frequency band is deducted from the lower frequency band. Lower frequency band is composed of noisy speech signal. It uses the short-time time-frequency attributes of the speech signal for VAD.

In this method, speech signals which are sampled at a rate of 8 kHz is tested. The spectrum band is of 4 kHz, which is divided into two frequency bands: LFB and HFB (LFB-lower frequency band ranges between 0 kHz and 1.5 kHz and HFB- higher frequency band ranges between 2.5 kHz and 4 kHz). First, the audio data is divided into window frames of duration 20 ms with an overlap of 10 ms, after the removal of the mean value. Then, in order to remove discontinuities between the frames hamming window is applied. Then the spectrum of each frame is obtained using FFT. Afterward, absolute LOG magnitude of the FFT values is taken. Then the resulting negative values, are considered as zeros. Afterward, the total values of the LOG spectrum of each speech window frame are calculated and energy in the LFB is obtained as EnLFB and that in HFB as EnHFB. Also, a moving average filter is applied for filtering these energy values. The energy in higher frequency band EnHFB, indicate the noise energy in the audio data. Generally, Gaussian speech noise is has a bandwidth of 4 kHz. Then M_EnHFB, which is the mean of EnHFB obtained in all speech frames, is subtracted from EnLFB for every audio frame. This result can be used for the detection of start-end points of VAD.

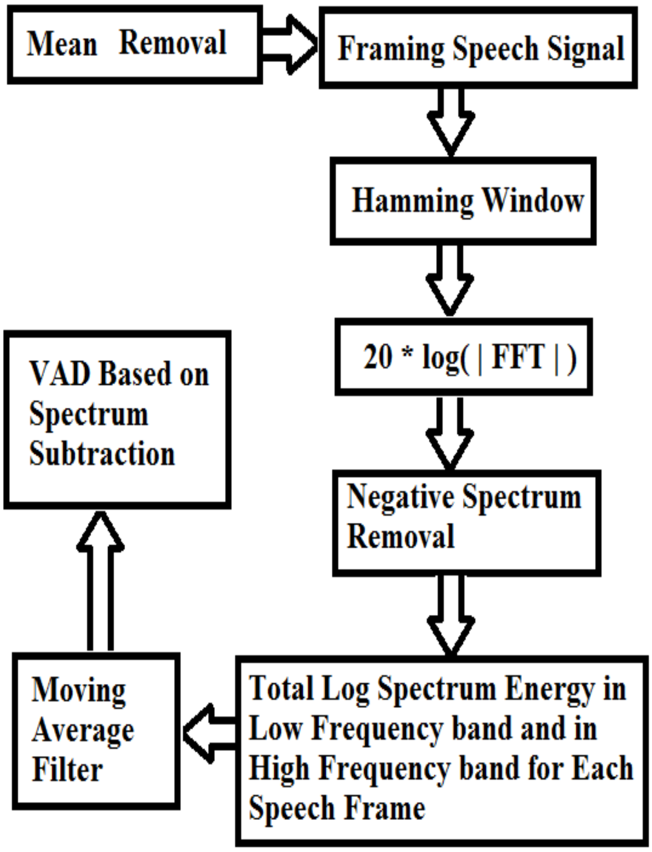
Mean Removal
Framing Speech Signal
Hamming Window
20 * log(| FFT |)
Negative Spectrum Removal
Total Log Spectrum Energy in Low Frequency band and in High Frequency band for Each Speech Frame
Moving Average Filter
VAD Based on Spectrum Subtraction

Figure 2: Block diagram of VAD.

Jing Pang’s method is easy to implement and has low complexity. In particular, for a positive signal to noise ratio (SNR) value, no particular selection of threshold is needed, and this promotes the automation of the voice activity detection. For signal to noise ratio (SNR) value greater than or equal to -5 dB, the suggested method shows good performance. This method is also feasible to attain the approximate locations of voiced segments in the noisy speech signal, for SNR of 10dB. The drawback of this method is that when the SNR is -10 dB or less, in order to obtain better VAD results, processing Gaussian noise variance using different advanced strategies is required.

Wissam A. Jassim et al. proposed a VAD method based on neurograms [3]. Since, neurograms represent most of the non-linear behaviors in the auditory periphery system, as compared to other 2D representations such as spectrogram, they are more effective. Applications like speech quality, emotion classification in speech and estimate of speech intelligibility makes use of the features extracted from neurograms. Neurograms are a form of signal depiction which is derived from the response of the human Auditory-Nerve (AN) system. This method improves the VAD accuracy by simulating neurograms using a computational model of the auditory nerve system for a range of Characteristic Frequencies (CFs). Discrete Cosine Transform (DCT) is used to extract features from neurograms. Then VAD prediction is made using a Multilayer Perceptron (MLP) classifier which is trained with these features.

Neurogram based voice activity detection method composed of stages, such as a training stage and testing stage. During the training stage, neurograms are simulated using auditory nerve model for the input speech signals and then feature extraction is applied. In the testing stage, features from true voice activity detection events is used to train the MLP classifier, which predicts the VAD events for an input feature set.

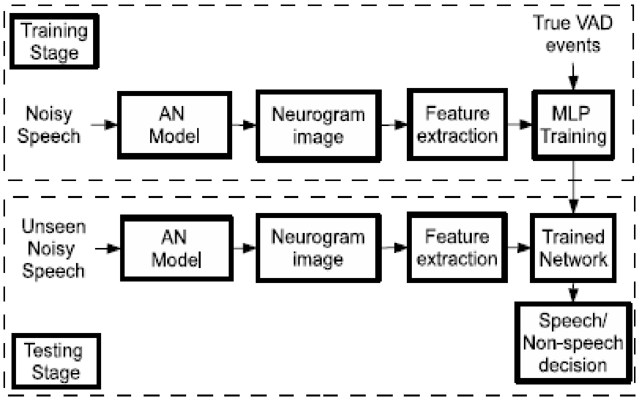

Fig. 3.VAD block diagram of the proposed method.

The proposed VAD method is evaluated and the using NIST scoring and QUT-NOISE-TIMIT corpus algorithm. The neurons in the peripheral auditory system exhibit phase-locking property, which attributes robustness to this method. The neurogram based method attained overall better results as compared to most of the existing VAD methods. The simulation results also indicated that the overall performance of VAD can be improved by mixing the neurogram features with other features.

3. Conclusion

The approach outlined by Wissam A. Jassim et al., which is based on neurograms achieved the best and most consistent performance at all SNRs and different noise types. The phase locking property of neurons in the peripheral auditory system leads to greater accuracy. The overhang scheme implemented was also very robust as there were very few errors associated with this method. There are however some concerns about the Wissam A. Jassim algorithm. The main concern is the computational complexity of the algorithm. Despite this shortcoming, this VAD method was shown to perform exceptionally well against other VADs in a standard testing framework.

References

[1] H. Wang, Y. Xu, and M. Li, “Study on the MFC similarity-based voice activity detection algorithm,” 2nd Int. Conf. on Artificial Intelligence, Management Science and Electronic Commerce, pp. 4391 – 4394, 2011.

[2] K. R. Borisagar, D. G. Kamdar, B. S. Sedan, and G. R.Kulkarni, “Speech enhancement in noisy environment using voice activity detection and wavelet thresholding,” IEEE Int. Conf. on Computational Intelligence and Computing Research, pp. 1 – 5, 2010.

[3] Jassim, W. A., & Harte, N. “Voice Activity Detection Using Neurograms”. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[4] S. Yiming, and W. Rui, “Voice activity detection based on the improved dualthreshold method,” Int. Con. on Intelligent Transportation, Big Data and Smart City, pp. 996 – 999,2015.

[5] Graf, S., Herbig, T., Buck, M., & Schmidt, G. “Features for voice activity detection: a comparative analysis”. EURASIP Journal on Advances in Signal Processing, 2015.

[6] Jing Pang. “Spectrum energy based voice activity detection”. IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), 2017.

[7] C. K., Pandiyan, P. M., Yaacob, S., & Saudi, A. “Mel-frequency cepstral coefficient analysis in speech recognition”. International Conference on Computing & Informatics, 2006.

[8] Godino-Llorente J.I., Gomez-Vilda P., and Blanco Velasco M., “Dimensionality Reduction of a pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short Term Cepstral Parameters”, IEEE Transaction on Biomedical Engineering, 53(10):1943-1953, 2006. 14

[9] S. A. Khoubrouy, and I. M. S. Panahi, “Voice activation detection using Teager Kaiser energy measure,” 2013 8th Int. Symposium on Image and Signal Processing and Analysis, pp. 388 – 392, 2013.

[10] Nitin N Lokhande, Navnath S Nehe and Pratap S Vikhe, k “Voice Activity Detection Algorithm for Speech Recognition Applications,” International Conference in Computational Intelligence (ICCIA) 2011.