International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 8, Issue.9, September 2020 E-ISSN: 2347-2693

Classification of Audio Segments using Voice Activity Detection

S. Kaur[1*], P. Mittal[2]

[1,2]Dept. of Computer Science and Engineering, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, Punjab, India

[*]*Corresponding Author: samanjeetkaur@gmail.com, Tel.: +91-96463-07945*

DOI: https://doi.org/10.26438/ijcse/v8i9.101105 | Available online at: www.ijcseonline.org

Received: 19/Sept/2020, Accepted: 23/Sep/2020, Published: 30/Sept/2020

***Abstract*—** Voice activity detection is classifying speech and non-speech frames. Effectively working and noise tolerant voice activity detection technique is responsible for better performance of many new speech technologies in the area of speech processing. In this paper, an unsupervised method for VAD is proposed to identify the segments of speech-presence and speech-absence in an audio. To make the presented algorithm effective and computationally fast, it is implemented by using long-term parameters that are extracted by using Petrosian algorithm used for fractal dimensions. This system plays a significant role in terms of achieving improved speech quality. Two types of datasets recorded in English and Arabic languages are used to analyses the output of the proposed algorithm. An Array of 85 audio signals of TIMIT Database, of different Signal to noise ratios is tested using the algorithm at once. The evaluated performance suggested that the proposed algorithm identifies segments in the audios with different SNR's

***Keywords*—**Fractal Dimensions

I. INTRODUCTION

In today's dynamic world, voice-over - Internet Protocol (VoIP) speech communication is attracting many scientists. Voice Activity Detector (VAD) is a method to separate out the voiced speech part and silence part in original speech signal. This method plays an important role in terms of obtain better quality of speech and to reduce bandwidth complexity, in many application such as speech recognition, speech compression systems, mobile communication etc. To understand the concept we can explore the application of mobile communication system. In number of mobile communication systems, when the dialogues are exchanged there are some silence periods in the interval. This silence period can be derelict or rejected in the network to save data or bandwidth by using VAD method. If the frame in speech is detected earlier to transmission then there is no need to broadcast the unwanted silence part [1] and only compressed speech can be transmitted which will also reduce the network traffic.

VAD used as a first step is an essential part of various speech processing applications, such as voice conferencing, echo deletion, voice recognition, voice editing, speaker recognition and hands-free telecommunications. The VAD facilitates quick data and voice applications in the areas of digital systems. Voice activity detection is a key in speech processing applications, since non-speaking segments are often removed. The VAD upsides involve relatively lower cellular phone energy usage. The improvements in Speech processing systems depend mainly on the percentage of pauses during speech and the reliability of the VAD used to detect these intervals.

The typical design of a VAD algorithm is as follows:

• Firstly there may be noise reduction stage, for example via spectral subtraction. This can be done by preprocessing of input audio signal.

• Then some features or quantities of the preprocessed audio frames are extracted to identify the required parameters for the algorithm.

• A classification rule is applied to classify the section as speech or non-speech. This classification rule is based on the threshold value calculated using the extracted features.

A. Voice Activity Detector Model

Voice Activity Detector model is the blueprint of common VAD techniques. A model of VAD shows the flow of any method with which a method works for Voice detection. It gives basic idea about the steps used in the technique of classification. This model is an ideology for binary decision making VAD. It gives output in the form of 0 or 1. Output 0 means non-speech frame and output 1 means speech frame. A VAD model shows many steps like input signal, preprocessing of the signal, feature extraction, decision making rule etc. these are basic steps in a VAD model that are common for all VAD techniques. These steps have multiple sub-steps. Various VAD techniques differ in the form of these sub-steps. For example one technique may use maximum margin clustering or other may use different probability density function for decision making.

In general, the basic rule of the VAD algorithm is to extract characteristics from the input signal and to compare the value with the threshold. If the threshold value is exceeded (VAD=1) indicates otherwise the presence of

speech (VAD=0) indicates the absence of speech. The VAD algorithm takes a binary choice of 20-40ms in output frame by frame basis.

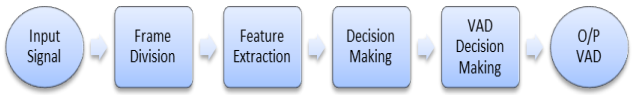


Figure 1. Voice Activity Detection Model

The basic model of VAD is shown in Figure 1. As shown in the figure, the first input speech signal which is digitized and the processing of the input speech signal through which certain parameters are extracted in second step. The extracted parameters are helpful to distinguish the voiced part and silence part of the speech. This Based on extracted information, decision making circuit play its role to decide the threshold value. Based on decision making part of VAD voiced frame and silence frames are segregated. VAD is widely used in speech processing applications such as speech enhancement, speech coding and speech recognition etc. Voice differentiates into voice or silenced depending on the features of voice. The VAD algorithm needs a few good features, such as: good decision law, background noise resilience.

The purpose of the study is that an efficient VAD tool with the ability to separate voice samples into voice-presence and voice-absence has many possible applications in various science disciplines, such as speech recognition systems (SRSs), medical diagnostics systems, and audio forensic analysis. In a Speech Processing Systems, detecting segments of speech-presence is important to generating accurate acoustic examples of different words or phonemes depending on the nature of the Speech Processing system being created. Inaccurate detection by the VAD method of segments of speech-presence and speech-absence will reduce the recognition accuracy and, consequently, the system performance.

The function of the VAD approach in the program for diagnosing vocal folds disorder (VFD) is critical. Due to the non-existence of segments of speech absence, most published studies have used continuous vowel for detection of VFD. Sustained vowel contains only segments of the voice presence which facilitate speech analysis. Long term features such as simple pitch, glow become unstable in the presence of fragments of speech-absence.

The VAD approach is thus necessary for computation of long-term features in order to prevent segments of the speech absence. Detection of voice or unvoiced frames also plays a key role in controlling the estimation techniques used in echo canceling and noise removal algorithms. In audio forensics, i.e., VAD methods are equally relevant in Audio authentication, fraud identifier and audio translation. It is therefore suggested that the effectiveness of the various speech recognition applications, speaker recognition systems or other Digital signal processing systems relies on the effectiveness of the voice activity detection algorithm used. Therefore the effectiveness of the VAD algorithm affects the efficiency of the applications for signal processing.

Rest of the paper is organized in the following manner as Section I contains the introduction of voice activity detection and purpose of the voice activity detection, Section II contain the related work of existing studies on VAD, Section III contain the some design parameters and objectives of study. Section IV contains methodology that is to be followed for the proposed method of VAD, Section V contains the results of the proposed method and performance evaluation steps, Section VI explains the final conclusion and discussion about the research study.

II. RELATED WORK

Some recently released VAD algorithms use parameters to formulate the decision rule frame by frame to define the difference between speech and non-speech frames. The different measures which are used in VAD methods include correlation coefficients, Cepstral coefficient, spectral slope and short-term energy, likelihood ratio-test [8] etc. Various VAD methods are different in terms of feature extraction step.

For the application to variable rate speech coding, a voice activity detection technique is developed. The developed VAD uses the method of decision-directed parameter estimation for the probability ratio test. In addition, the authors have suggested an efficient scheme that takes into account previous observations by modeling speech occurrences through a first-order Markov method. This method was tested efficient for low signal to noise ratio [1].

Another research introduced a new voice activity detector (VAD) to correctness of voice detection in noisy conditions and the efficiency of voice recognition systems. The method specified an optimum test of probability ratio (LRT) involving several, independent observations. The decision rule used in this algorithm improved the accuracy of speech systems over the other methods [2].

This algorithm detecting voice activity is based on several predictive methods. It uses various probability density functions such as complex Laplacian and Gamma functions. These are used to analyse the predictive properties. In addition to the Gaussian model, authors also incorporated the complex Laplacian and Gamma probability density functions to the analysis of statistical properties. They evaluated the statistical properties of the DFT spectra of the noisy speech under different noise conditions using goodness-of - fit tests. The probability ratio test under the specified predictive methods is based on the objective of VAD, based on the statistical analysis. The statistical features are differently affected by different type of noise and the levels of noise. This method was efficient for stationary and non-stationary noise [3].

Since the statistical characteristics of the speech signal are differently affected by the noise types and levels, to cope with the time-varying environments, this approach is aimed at finding adaptively an appropriate statistical model in an online fashion. The performance of the proposed VAD approaches in both the stationary and non-stationary noise environments is evaluated with the aid of an objective measures [4].

Another technique based on Maximum margin clustering that is unsupervised technique developed for VAD. In this technique they used multiple observation compounds to enhance the accuracy of the method. This comprises of two characteristics which are multiple signal-to-noise ratios of observation and the maximum probability of multiple observations [5].

III. DESIGN PARAMETERS

A. Frame Duration

The choice of frame duration is one of the important considerations for VAD. The transmitted active frames are stored in a form of packet buffer in the receiver which allows the speech to continue playing even after any delay in network. Suppose, a buffer 4-5 packets in a VoIP system having frame duration of 10ms will allow the system to play after 30-40ms from the time the queue starts building up in the receiver‘s end. There will be delay of 150-200ms for frame Length of 50ms that is not sufficient or intolerable. Therefore it is necessary to set the frame length properly to prevent these conditions. The specifications for toll quality encoding of speech for all VAD algorithms are [1]: 8 kHz sampling frequency, 256 levels of linear quantization (8 Bit PCM) [2], Single channel (mono) recording.

B. Initial threshold value

The presented VAD algorithm trained for small period that contains only pre-recorded sample initial stage of threshold will track the background noise. Threshold values are considered from computing the mean energy of the samples of speech. For example if we consider initially 200ms of the samples does not contain any speech i.e. inactive frames. Thus care must be taken while deciding threshold value otherwise may lead to poor performance. The qualified VAD algorithm for a small duration which contains only the initial threshold stage of the pre-recorded sample can monitor background noise.

Following are the main research objectives: -

1. To study various Voice Activity Detection techniques for speech segmentation
2. To propose algorithm for Voice Activity Detection technique using Petrosian fractal algorithm.
3. To compare various Voice Activity Detection techniques with proposed technique using parameters like accuracy, Initial threshold value, Frame Duration.

IV. METHODOLOGY

A. Pre-processing of Audio signal

Pre-processing of speech signal is the first stage in the proposed procedure. This includes the split of audio signals into frames [F_1, F_2, F_3 ..... F_n]. To detect speech presence and speech absence, audios are split into frames of defined length. Frame intervals are kept shorter in order to accurately distinguish all segment types. The downside of long frame length is that it may contain speech in certain parts, and may contain silence or brief delays in the remaining parts. Therefore the fixed frame length is taken 0.01ms in the proposed algorithms. In this step the sampling frequency (Fs) is also set for the audio input. The sampling frequency has been down to 22.5 KHz.

The parameters are measured in the next step such that parts of speech-presence and speech-absence are correctly identified. The measured properties are the dimensions of the audio fractals. The steps mentioned for calculating the fractal dimension of each frame in the paragraph below.

B. Estimation of Fractal Dimensions

The fractal dimensions for every segmented frame are calculated. Using Petrosian algorithm the fractal dimensions are calculated in the proposed technique. Petrosian makes use of a fast estimate of the fractal dimension. But this estimate is actually the FD of a binary sequence, as it was originally defined by Katz”Ref. [5]”. Since waveforms are Analog signals, a binary signal is extracted from four separate methods, denoted with letters a, b, c, and d respectively. Often used in a fourth form but it is the same as for an adjustable parameter. Method a produces the reference signal by assigning those when the waveform value is greater than the data window average is considered, and zero if it is lower than the average. In method b, the binary sequence is generated by specifying one each time the value of the signal is outside the mean plus and minus the standard deviation unit, and assigning zero otherwise. Method c constructs the reference signal by subtracting successive samples from the sound wave record. The binary sequence of subtractions is generated by assigning, or based on if the subtraction result is positive or negative, simultaneously [6].

The proposed algorithm is based on the method d .In method d, the differences between consecutive waveform values are given the value of one or zero depending on whether their difference exceeds or not a standard deviation magnitude. A variation of this method consists of utilizing an a priori chosen threshold magnitude different from the standard deviation, is denoted by Petrosian as method c. The FD of any of the previous binary sequences is then computed as:

$$D = \frac{\log_{10} N}{[\log_{10} N + [\log_{10} (\frac{N}{(N + 0.4M)})]} \quad (1)$$

Where N is the length of the sequence (number of points), and M is the number of sign changes (number of dissimilar pairs) in the binary sequence generated. M is multiplied with the a-priory chosen threshold value.

C. Performance Evaluation

The evaluation of the proposed approach is measured using two speech databases reported in separate languages, English and Arabic. The method is analysed with clean audio signals to observe its reliability. To test the output of the proposed method, two databases in different languages are used. The first is the Massachusetts Institute of Technology (TIMIT) database of Texas Instruments, and the second is the Arabic language database for King Saud University (KSU) [9]. TIMIT's language is English, while the KSU speech database language is Arabic.

Sound recordings from both datasets will be segmented into frames, and the parameter (FD) for each frame will be determined. The results of the proposed technique will demonstrate a significant difference between the voice-presence and speech-absence segments of the FD's. The value computed of FD's of the parts of speech-presence will be roughly equal to 1 and equals 0 for the sections of silence. After this discrete conversion, a threshold selected a-priory value will be compared to label that the segment of particular frame is voiced or unvoiced (U) portion.

V. RESULTS

The effectiveness of the proposed method is evaluated by using speech database recorded in English language. The method is evaluated by using clean audio recordings.

Results of proposed algorithm by using TIMIT Database

The Texas Instruments Massachusetts Institute of Technology (TIMIT) database [6] is used to evaluate the proposed method. The database is developed to provide speech data for obtaining acoustic-phonetic knowledge. The TIMIT database is recorded at Texas Instruments and then recorded at the Massachusetts Institute of Technology. The database is distributed by the National Institute of Standards and Technology. The database has been widely used in VAD systems and various speech-related applications .The TIMIT database is recorded by 630 male and female speakers from eight dialect regions in the United States. The language of the TIMIT database is English, and each speaker recorded 10 sentences at 16 KHz sampling frequency and 16-bit rate. For each speaker, the rest two sentences are fixed, while the remaining sentences vary from one speaker to another. All sentences are read by the speakers, and they are recorded in one session in a soundproof room. Audio samples of the TIMIT database are partitioned into frames, and the fractal dimension is calculated for each frame. A clear difference between the fractal dimension of speech- presence and speech-absence segments can be observed in Figure 3. The fractal dimensions are approximately equal to 1 for the silence parts and greater than 1 for the speech-presence parts.

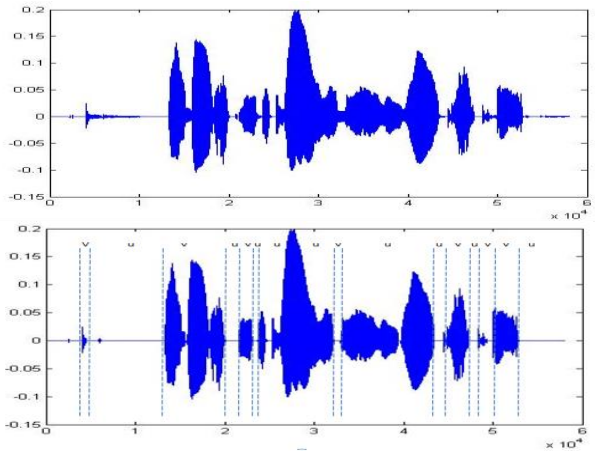
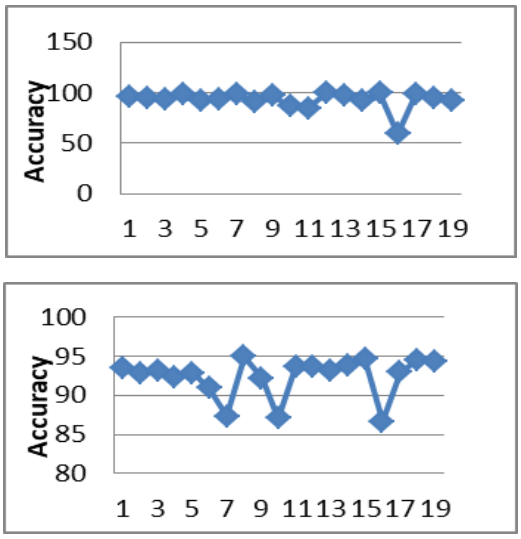


Figure 2. Automatic segmentation of audio from TIMIT Database recorded in noise free environment

A fixed Threshold value is considered for decision making that a frame has voiced segments or unvoiced segments. The performance of the proposed method is assessed using the metric set out in Equation (2). This metric is previously used in [7]. TP stands for true positive, meaning a speech frame was identified as a voice frame by the system. FN reflects the false negative, indicating that the system identified the voice frame as a non-voiced context. TN stands for the true negative, meaning that a non-voiced frame is perceived as an unvoiced frame. Finally, FP reflects the false positive, indicating that the system identified a voiceless frame as a voiced frame. The precision of the method for some clean audio of the TIMIT database is shown in Fig. 4. The first sentence of the TIMIT database is used. The average accuracy of the randomly selected 85 audio is 92.45%. The approach is also tested by adding various forms of noise to audios. In the audios three types of noise are introduced with specific signal-to-noise ratios (SNRs) to determine the robustness of the proposed system. This lists the accuracy of the process with noisy audio in Table 1.

$$\text{Accuracy} = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} * 100 \qquad (2)$$



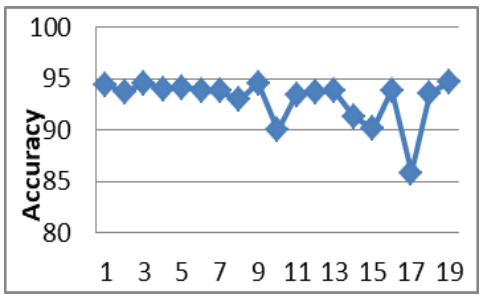


Figure 2. Accuracy of clean audio signals of TIMIT database with 5dB, 15 dB and 25 dB noise

The approach is evaluated by using SNR's of 5dB, 15dB, 25dB. Results of the proposed algorithm shows that the performance of this algorithm also good in noisy audio. The noisy audios are generated by adding white noise in the clean audios. The noise of different SNRs is added.

Table 1. Accuracy of audio signals of TIMIT database

SNR	5dB	15dB	25dB
Accuracy	93.36%	92.35%	93.14%

VI. DISCUSSION

The approach proposed classifies parts of speech-presence and speech-absence of an audio based on the computed features. The characteristics are segment fractal dimension, which defines segment type. Fractal dimension tests a waveform's complexity. The waveform with the higher amplitude has a greater fractal dimension compared to a waveform exhibiting the lower amplitude. There are many algorithms to measure a waveform's fractal dimension. An established method's success depends on the fractal estimation algorithm choosing the right path. The aim of the method developed is to identify segments of speech-presence and speech-absence in order to be used in different speech-related applications. The segments of speech-presence have higher amplitude relative to segments of speech-absence. There are various methods discovered in previous studies to calculate fractal dimensions of waveform. The proposed technique used Petrosian algorithm to calculate fractal dimensions and identify the speech-presence and speech-absence segments. This algorithm is based on difference between amplitude of continuous waveforms. This algorithm takes less computation time to calculate fractal dimensions as compared to other algorithms. Table1 takes the results of the proposed method. The accuracies in Fig. 4 they are for minimal noise (5dB, 15dB, 25dB). Comparing the accuracies with previous studies showed in ensures that the method proposed outperforms current VAD methods.

REFERENCES

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

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

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

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

[5] S. Mudaliar , T.Tahilramani, "*Techniques of voice activity detection: A review‖ in "*IJSRD - International Journal for Scientific Research & Development‖ Vol. **5**, Issue **02, 2017**

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

[7] Z.Ali, M.Talha, " *Innovative method for unsupervised voice activity detection and classification of audio segments, in IEEE Int. Conf.*, Special section on radio frequency identification and security technique , Vol no**.6** April **2018**.

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

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

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

[11] T. R. Senevirathne, E. L. J. Bohez, and J. A. Van Winden, "*Amplitude scale method: New and efficient approach to measure fractal dimension of speech waveforms*,'' Electron. Lett., vol**. 28**, no. **4**, pp. **420–422**, Feb. **1992**.

AUTHORS PROFILE

Samanjeet Kaur received Bachelors in Information Technology (2016) from Guru Nanak Dev Engineering College, Ludhiana in India. She is currently a Masters in Computer Science and Engineering candidate in the Department of Computer Science at Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, India. Her current research interests include Speech Processing, audio segmentation, signal segmentation into syllables.

Puneet Mittal received Masters in Computer Engineering (2012) from PTU Jalandhar and Bachelors in Computer Science Engineering (2005) from PTU Jalandhar. She is currently a PhD research scholar in Speech based command and system for mobile phone application in Punjabi language at Punjabi University. She is also an Assistant Professor at Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, and India. She is teaching UG and PG computer engineering students from the last over 12-years. She is available at puneet.mittal@bbsbec.ac.in

