IJAREEIE

e-ISSN: 2278-8875
p-ISSN: 2320-3765

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

Volume 12, Issue 5, May 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.317

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|

||Volume 12, Issue 5, May 2023||

|DOI:10.15662/IJAREEIE.2022.1205018 |

A LPC Based Voice Activity Detection Process

Nikhil Kumar, Sumit Dalal, Rohini Sharma

Student, Dept. of ECE, SKTIM, Bahadurgarh, India

Assistant Professor, Dept. of ECE, SKTIM, Bahadurgarh, India

Assistant Professor, Dept. of Computer Science, GPGCW, Rohtak, India

ABSTRACT: Voice activity detection (VAD) is used in several speech signal processing usages to separate a stream of audio into periods containing speech activities and periods containing no speech activities. A number of methods have been suggested for that reason. Few relies upon attributes obtained from the power spectral compactness, while others make use of the signal's frequency. We render an organised review of numerous recognised VAD elements that target various speech qualities. We also give a comprehensive examination of recent VAD processes based on energy thresholds, zero crossing, and other statistical variables. We have also proposed an enhanced LPC (linear predictive coding) approach for synthesized speech classification.

KEYWORDS*:* Detection of voiced and unvoiced samples, Linear predictive coding, signal energy, estimation of pitch

I. INTRODUCTION

With the latest developments in speech signal processing approaches, the necessity to reliably recognise the existence of speech in the incoming signal under diverse noise conditions has become a major industrial concern. Voice Activity Detectors (VAD) are employed to detached the speech segment from the non-speech part of an audio signal. VADs are a type of signal processing technology that identifies the existence or absenteeism of speech in brief audio signal fragments [1]. An incorporated VAD in a speech messaging system increases channel capacity, decreases co-channel disruption, and lowers power usage in handheld electronic devices in cellular radio networks, allowing for concurrent voice and data uses in multimedia telecommunications [2]. A VAD is utilised to acquire noise features and predict the noise spectrum in slowly fluctuating non-stationary settings where speech is damaged by noise. A simple VAD works by retrieving measurable features from an inbound audio signal that is separated into frames of short duration. These recovered audio signal components are then compared to a threshold cap, which is often derived from the noise only periods of the input signal, and a VAD conclusion is generated. If a component in the input frame exceeds the predicted threshold value, a VAD decision (VAD = 1) is generated, indicating the occurrence of speech. Alternatively, a VAD conclusion (VAD = 0) is generated, indicating that there is no speech in the input frame. Fig. 1 exhibits an instance of the basic VAD process. In general, A VAD approach is made up of two phases: the extraction of features and a discriminating prototype. Early research concentrated on energy-enabled features that may be integrated with the zero-crossing rate (ZCR) [3]. These characteristics, though, are greatly influenced by additive noise. As a result, additional characteristics such as autocorrelation-based characteristics [4], MFCCs [5] and many others [6,7,8,9]. Other approaches [10] relies upon a geometric framework of the Discrete Fourier Transform (DFT) factors. Additional methods take advantage of the idea that voice and noisy signals have distinct variance characteristics [11]. subsequently, a few investigations have looked into the use of various characteristics in parallel.

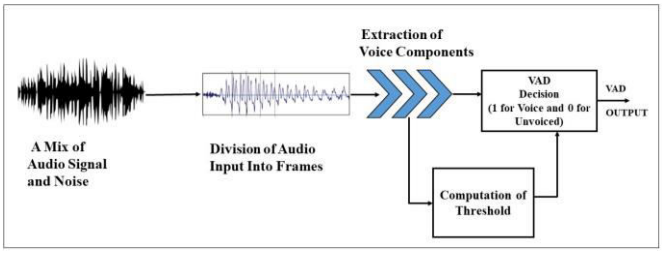


Fig. 1. An Instance of the VAD Process

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|

||Volume 12, Issue 5, May 2023||

|DOI:10.15662/IJAREEIE.2022.1205018 |

II. ELEMENTS OF VAD

A. *Voice Activity Detection Process*

The VAD process is dissented into two phases. To begin, attributes are taken from the distorted speech signal in order to create an illustration that differentiates between speech and noise. In the subsequent phase, the attributes are subjected to a recognition scheme, which results in the ultimate decision. Speech recognition has a restricted temporal resolution that is significantly less than the rate of sampling of the sound signal. Thus, normally, the selection is not made for every sample n of the signal s (n). Rather, the signal is fragmented into small frames that store N noisy signal patterns.as shown in (1).

$$s(f) = [s(fK - N + 1), \dots\dots., s(fK - 1), x(fK)\ ]^T \quad (1)$$

Additionally, the rate of frame is condensed by an integer component *K* related to the sampling rate. The purpose of VAD is to regulate if frame s (f) has voice or not. Let us assume that there are two types of states: S1 which is a combination of voice *v(f)* and noise samples *n(f)* and S2 which is a pure noise sample *n(f),* as shown in Eq. 2.

$$S1: s(f) = v(f) + n(f)$$
$$S2: s(f) = n(f) \quad (2)$$

Using Equation (3), the VAD is decided as follows:

$$VAD_{atr}(\acute{\eta}, f) = \begin{cases} 1, When\ S1\ is\ accepted\ state \\ 0, When\ S2\ is\ accepted\ state \end{cases} \quad (3)$$

Where ή is a threshold value. The S1 denotes detection of voice and S2 denotes absence of voice.

B. *Voiced and Unvoiced Patterns*

A two-dimensional (time and intensity) visualisation of a sound is a waveform. Waveforms are referred to as time domain sound descriptions since they depict variations in intensity over time. Speech sound sources are classified into two categories: (a) Voiced sample is generated by periodic vibrating of the vocal folds and (b) Aperiodic sound is created by disturbance in the vocal tract, leading to silent speech. In order to generate the diverse sounds of speech, these two sound sources have been altered by the frequency-selective (filtering) impacts of distinct vocal tract geometries. To generate all of the vowels as well as the numerous voiced consonants, the voiced source can be filtered ("modulated") by the location of the lips, tongue, and velum. In a comparable vein, aperiodic sources can be filtered to produce diverse unvoiced speech sounds, but the location of the constriction that causes turbulence has the most influence on the sound of such speech tokens. The voiced signal exhibits substantial periodicity in the time spectrum and high energy concentration in lower frequency spectrum features. In contrast, the unvoiced signal differs from the voiced sound in that it lacks distinct time-domain and frequency-domain features, reminiscent of Gaussian white noise [12].

III. PERFORMANCE MEASUREMENTS ATTRIBUTES

Because of the variety of speech qualities, it is preferable in practice to combine complementing aspects. As a result, we describe various speech qualities and examine how well they are expressed by the attributes.

A. *Short Period Power*

The power of the signal can be used as a preliminary indication of the existence of voice. Considering that the voice elements have larger power values than the noise in the environment, a threshold can be used to recognise speech. The Lombard reflex [13], which allows speakers to raise their voices in noisy situations, makes the assumption of increasing power plausible in many circumstances. A defined threshold, on the other hand, needs the level of noise and voice to be identified ahead of time. The normalization of the power improves the separation of voice and noise elements.

The short-time power of the n segments is represented by P_n, which is equivalent to the square sum of the small segment sample values. Regulate the primary sampling sequence for voice signals. The n frames of voiced signal achieved by combining windows and sub-frame pre-processing is denoted by Sn(m). This power of the n frame voice signal Sn(m) is specified with Pn demarcated as follows:

$$P_n = \sum_{m=0}^{N-1} S_n^2(m) \quad (4)$$

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|

||Volume 12, Issue 5, May 2023||

|DOI:10.15662/IJAREEIE.2022.1205018 |

The Pn function detects the modification in amplitude of a voiced signal, but it contains a fault that makes it particularly profound to high levels. The motive for this is that the measurement of short-term power is dependent on the square values of all samples, and square operations lead to an expansion of the amplitude variance between the amplitude values of each neighbouring samples, that adds needless challenges to a broad window selection. Since a wide window is required to find a virtuous levelling influence of the square variation between the sample values, a window that is too wide will end up in Pn being problematic to represent time-changing attributes of the sound signal power. With the advancement of speech analysis process, some enhanced energy-based VAD techniques have emerged. Authors in [14] proposed a spectrum energy based VAD algorithm. It is obtained using the overall power in the overlying sound window frames. The upper frequency band noise energy is eliminated from the lower frequency band raucous speech domain. A progressing average filter is also employed to uniform the energy waveform of the speech spectrum. The suggested method for detecting vocal activity is stable and functions well for a wide range of signal-to-noise ratio (SNR) values.

B. SNR Based Estimation

The SNR is a popular way to normalize power. Measurements of the (SNR) of speech are significant in noise suppression and comprehension forecasts based on the speech transmission index. To identify speech and non-speech portions, VAD algorithms must be utilised either directly or implicitly during SNR estimations. During SNR estimates, most studies have fixed the threshold decision for VAD for non-speech and peech categorizations. According to the authors of [15], fixing the threshold choice for all testing scenarios is not optimum for regulating the false reception and missed-recognition rates of sound. In this research, they present SNR estimations based on an optimisation on a receiver operating characteristic (ROC) curve, the compromise between false voice acknowledgment and miss detection rates. Moderately establishing the evaluation level in VAD for all situations, they ideally predict the conclusion threshold for each SNR condition using a ROC curve. Thresholds in subband signals are optimised using a huge training data set comprised of varied circumstances and noise kinds. After detecting sound patterns, SNR is calculated by adding the sub-band powers of voice and noise from all spectrums. A speech signal's global SNR assesses the comparative level of energy between needed voice and contextual noisy signals. It is expressed as a signal power ratio in terms of the power of speech P_S and power of noise P_N as follows:

$$GSNR = 10log_{10}\left(P_S/P_N\right) \quad (5)$$

Given that noise and speech signals can occur at various times, exact SNR computations should account for speech and noise detectors through guesses of noise and speech powers. The powers of speech and noise in time t are PS (t) and PN (t) respectively. The global SNR is redefined using Eq. 6:

$$\overline{GSNR} = 10log_{10}\left(\frac{\sum_0^T P_S(t)\ S1(t)}{\sum_0^T P_N(t)\ S2(t)}\right) \quad (6)$$

Where T is the span of the speech signal. *S*1 (*t*) and *S2* (*t*) are the hypotheses for deciding the present time signal as voice or noise.

C. The ROC Curve

In recognition theory, performance is frequently finalized using a ROC curve. Here, there is a tradeoff among incorrect reception and miss detection of the wanted signal. Setting a threshold choice for VAD entails fixing the functioning criterion as a point on the ROC curve. Under distinctive noisy circumstances, we must adjust distinct performance circumstances on the ROC curve. Authors in [15] developed a voice and non-voice recognition technique based on optimising the relationship between speech incorrect acceptance and miss perception rates. Furthermore, because noise considerations have distinct impacts within various sub-bands, the threshold decision was improved in sub-band signals on a training data sample. After detecting voiced and unvoiced signals in all sub-bands, the standard SNR was calculated by adding the powers of signals from all sub-bands.

D. Zero Crossing Rate (ZCR)

At ZCR the signal goes from +ive to 0 to -ive or from -ive to 0 to +ive. Its utility in identifying percussive sounds has been widely applied in voice recognition and music information extraction [16]. For speech signals the ZCR is defined as follows [17]:

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn\,[x(m)] - sgn[x(m-1)]w(n-m)| \quad (7)$$

Where,

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|

||Volume 12, Issue 5, May 2023||

|DOI:10.15662/IJAREEIE.2022.1205018 |

$$sgn[x(n)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases} \quad (8)$$

And the window function w(n) is defined as:

$$w(n) = \begin{cases} \frac{1}{2N}, 0 \leq n \leq N-1 \\ 0, otherwise \end{cases} \quad (9)$$

Where the window sample size is *N*.

According to the speech creation prototype, the power of voice sound is intense under roughly 3 kHz due to the domain fall induced by the glottal wave, while the majority of the power of unvoiced speech is discovered at upper frequencies. Because high frequencies indicate high ZCRs and low frequencies indicate low ZCRs, there is a sturdy relationship between the ZCR and the power distribution with frequency. An acceptable generalisation is that when the ZCR is large, the sound signal is unvoiced, and when it is low, the sound signal is voiced.

E. Pitch and harmonicity

In the speech, pitch refers to the relative elevation or lowering of a tone as experienced by the ear, which is determined by the number of oscillations/ second generated by the vocal chords[18]. The two related issues of voice source analysis are determining the pitch and uttering of sound signals. The vocal cords tremble in a quasiperiodic pattern during spoken speech. Unvoiced innervation in speech is caused by airflow that is turbulent at confinement or by the exhalation of a closing in the vocal territory. The factors that we must decide are the mode of excitement, i.e., the existence of a voiced excitement and the existence of a voiceless excitement, An issue we'll call voicing decision, and the rate of vocal cord trembling, that is commonly described as pitch identification or basic frequency decision in the research field. Pitch, which is defined as the basic frequency F_0 and the basic period T_0 of signal, can be determined using a variety of methods. If a signal is perfectly stable and periodic, all procedures - if used appropriately - produce similar results. However, because both speech and harmonious signals are inconstant and time variation, elements of every approach like the initial location of the measuring, the span of the evaluating period, the method of averaging, or the functioning domain start affecting the outcomes and may result in anticipates that vary from procedure to procedure, even if all of these results are right and precise.

The basic harmonic waveform can be applied to identify F0 in the sound. If present, this harmonic is cut off from the signal in the preprocessor by substantial low-pass filtering. There are three key separators: The most basic is zero-crossings analysis, followed by nonzero threshold study, and ultimately threshold study with hysteresis. When the zero axis is crossed with a specified polarity, the zero-crossings evaluation basic extractor places a marking. This necessitates that the input waveform has just two zero crossings for each period. When a nonzero threshold exceeds the value, the threshold analysis basic extractor places a marker.

Percussion of the vocal cords produces a harmonically rich resonance with a unique pitch from 50 to 250 Hz for voiced phonemes [19]. This harmonic form, which is prominent in speech, is shared by all vowels and consonants. Characteristics that represent the harmonic pattern are trustworthy speech indicators. Unvoiced speech segments, including certain fricatives, cannot be recognised using harmonicity or pitch-enabled attributes alone. Furthermore, music and other harmonic noise elements may be interpreted incorrectly as speech. The harmonic pattern of speech is captured by the (standardized) auto-correlation function (ACF) [20]. The ACF is a core technique for various pitch-enabled speech identification features. This characteristic is used by characteristics that depict the highest ACF peak, the ACF cyclicity, or the disparity between maximum and minimum values. The ACF reflects the voice cords' harmonic excitation. Yet it additionally demonstrates vocal tract properties. The cepstrum can be used instead to isolate the two reactions [21].

IV. ANALYSIS OF ENHANCED LINEAR PREDICTIVE CODING FOR VOICE DETECTION AND PITCH ESTIMATION

The signal of speech should be treated to eliminate noise before obtaining the relevant features in speech recognition. The aim of attribute extraction is to depict a sound signal through a fixed quantity of signal components. This is due to the fact that tackling all of the data contained in the sound wave takes too much time, while some of the details are irrelevant to the determination task. Attribute mining is achieved by altering the sound wave to a parametric illustration at a lower data rate for later handling and study. It converts the intermediate resonance signal into a short yet analytical depiction that is more discriminatory and accurate than the original signal. Because the front end is the first component in the series, the overall performance of the following characteristics (pattern harmonizing and amplifier modelling) is heavily influenced by it.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|

||Volume 12, Issue 5, May 2023||

|DOI:10.15662/IJAREEIE.2022.1205018 |

Attribute mining methods naturally provide a multiplanar feature vector for each spoken signal. To parametrically characterize the sound signal for perception purposes, a variety of techniques are available, including linear prediction coding (LPC) [22], and MFCC [23]. In this work, we have explored the LPC method for speech detection and pitch estimation of sound. Feature abstraction minimises the size of the spoken signal without affecting its power. Several preprocessing processes must be completed before the features may be retrieved. The preprocessing phase is highlighted. It is accomplished by running the signal over a FIR filter [24]. This is followed by frame blocking, which divides the speech signal into frames. It eliminates the sound interface that exists at the beginning and conclusion of the spoken signal.

After that, the framed voice signal is windowed. A bandpass filter is an appropriate window that is used to minimise disjunction at the beginning and end of every frame. Hamming and Rectangular windows are the two most well-known types of windows [25]. The Hamming window is a window function that is often used in speech analysis to minimise abrupt changes and unwanted frequencies in the framed speech. It sharpens harmonics and removes signal discontinuities by tapering the commencement and termination of all frames by zero. It also mitigates the spectral alteration caused by overlap. In this work, we have used Hamming Window for the generation of frames.

Linear prediction, also known as linear predictive coding (LPC), is driven by the notion that the speech signal may be estimated at any time instant by a linear amalgamation of its previous values [26]. The linear prediction theory enables us to discover precisely what is expectable in the signal and eliminate those details from the voice signal before it is transmitted over the digital channel.

The frequency-spectrum technique (FT) and the time- spectrum technique (LP) can be used to analyse speech signals. Though the frequency/time-domain methods seem to be distinct, there is a strong relationship amid the 2 for constant signals. This relationship is merely tenuously extended to non-constant signals, like voice which properties change gradually over time. The LPC cofficients mimic the humanoid vocal band and deliver sturdy voice attributes. It assesses the audio signal by evaluating the formants, subtracting their impact from the signal, and evaluating the quantity and periodicity of the remnant that remains. The formant frequencies are the periodicity at which the resonant crests happen. Thus, the places of the formants in a sound signal can be predicted using this approach by determining the LPC above a slithering window and positioning the peaks in the band of the ensuing LP filter. By lowering the error between the input voice and the anticipated voice, the LP approach is used to produce the filter coefficients equal to the vocal band. The LPC evaluation of speech signals estimates any speech sample at a given time as a combination of previous samples. The speech generation LP prototype is presented as [27]:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{10}$$

Where $\hat{s}(n)$ denotes the forecast sample, s denotes the sound sample, and p denotes the predictor measurements. As a result, every frame of the windowed signal is related, and the order of the LP analysis is determined by the highest autocorrelation value. The LPC analysis follows, in which every frame is turned into an LPC varibles group consisting of the LPC coefficients. Fig. 2 depicts an overview of the LPC application process.

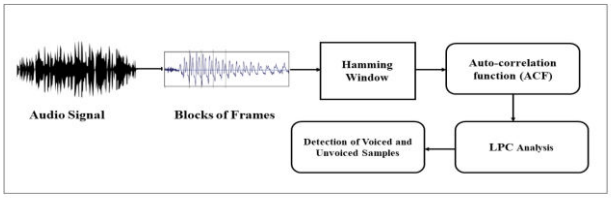


Fig. 2: A depiction of Hamming Window Based LPC Process

V. SIMULATION RESULTS

The enhanced method has been implemented in MATLAB 2014 and results are analyzed. Here we have analyzed VAD attributes like ZCR, and Short term power. Also, we have successfully classified an audio signal into voiced and

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|

||Volume 12, Issue 5, May 2023||

|DOI:10.15662/IJAREEIE.2022.1205018 |

unvoiced signals. Fig. 3 shows a speech sample in terms of the Time and Amplitude domain. Fig. 4 shows an instance of the Spectrogram of a Hamming Window of Speech Sample. Fig.5 shows the division of the input signal into various frames. Fog. 6 shows the short term power and ZCR concerning frames of input signal. Fig. 7 illustrates a comparative analysis of the detection of voice in speech signal by LPC and DFT techniques. The DFT is a complex-valued function of frequency that translates a fixed number of evenly spaced samples of a function into a equal length series of evenly spaced samples of the DTFT.

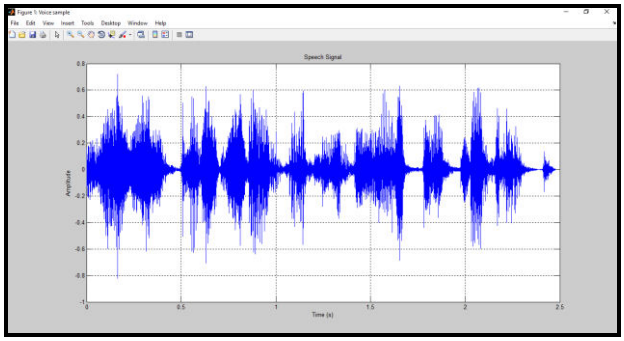


Fig. 3: Speech Sample

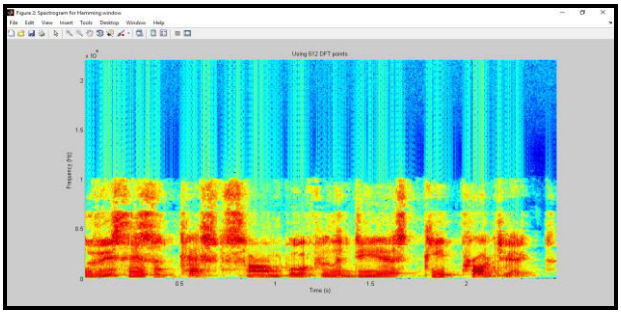


Fig. 4: Spectrogram of a Hamming Window of Speech Sample

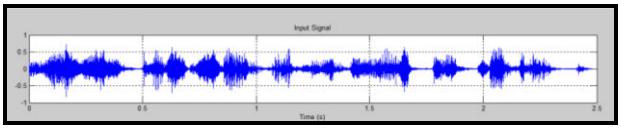


Fig. 5: Division of Input Signal into Frames.

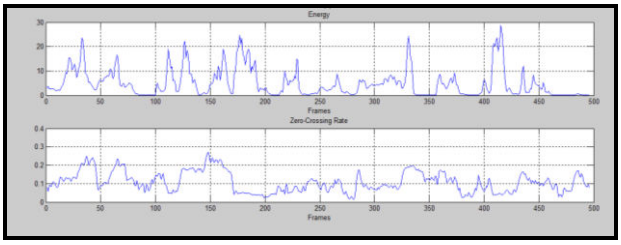


Fig. 6: Short Term Power and ZCR of Input Signal

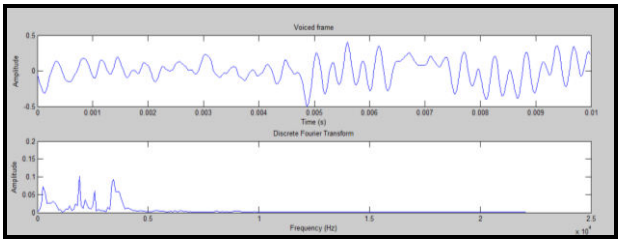


Fig.7: Detection of Voiced Frames by LPC and DFT

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|

||Volume 12, Issue 5, May 2023||

|DOI:10.15662/IJAREEIE.2022.1205018 |

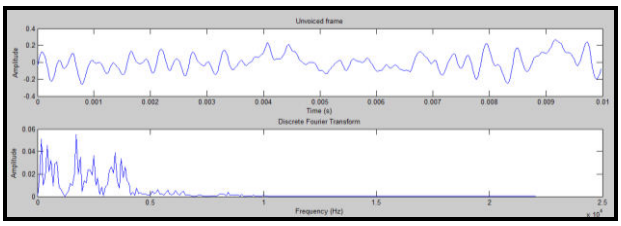


Fig.8: Detection of Unvoiced Frames by LPC and DFT

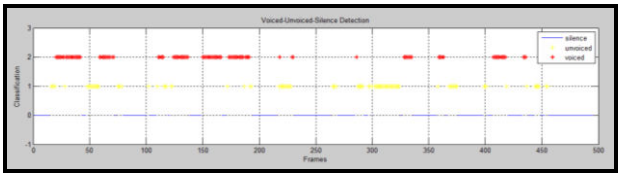


Fig. 9: Detection of voiced, unvoiced and silence from signal

The outcomes illustrate that the LPC technique is more effectual than the DFT method in regards of detection of silence-voiced-unvoiced patterns of input audio signal.

VI. CONCLUSION

Choosing a set of attributes that reflect discriminatory aspects of speech/noise is a crucial part of constructing a VAD algorithm. If the sound qualities portrayed by the attributes are not hidden by contextual noise or other intervening signals, appropriate VAD findings can be predicted. In order to be useful in real-world situations, the features must meet extra standards imposed by constrained hardware and latency restrictions. LPC is useful for encoding high eminence speech at low bit rates. Noise is a significant difficulty in the procedure of attribute extraction and speaker perception in general. As a result, scholars have made many adjustments to the aforementioned procedures in order to make them less vulnerable to noise, more strong, and time-consuming. These techniques have also been employed in sound identification. The mined data will be used as input to the classifier for recognition. The approaches of attributes extraction outlined above can be applied in MATLAB. Through enhanced LPC, we have successfully separated the voiced and unvoiced patterns. We have compared the enhanced LPC method with the DFT method, and concluded that LPC method is more efficient than DFT method for the detection of voice in an audio input signal.

REFERENCES

[1] D. S. Jat, A. S. Limbo, and C. Singh, Voice Activity Detection-Based Home Automation System for People With Special Needs, Intelligent Speech Signal Processing, Academic Press, 2019, pp. 101-111.

[2] Ji .Wu and X. Zhang, “An efficient voice activity detection algorithm by combining statistical model and energy detection,” EURASIP Journal on Advances in Signal Processing, Article No. 18, 2011.

[3] B. Kotnik, Z. Kacic, B. Horvat, “A multiconditional robust front-end feature extraction with a noise reduction procedure based on improved spectral subtraction algorithm,” in Proc. 7th Europseech, pp. 197-200, 2001.

[4] S.O. Sadjadi, J. Hansen, “Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux,” IEEE Signal Processing Letters, vol. 20, pp. 197-200, 2013.

[5] T. Kristjansson, S. Deligne, P. Olsen, “Voicing features for robust speech detection,” in Proc. Interspeech, pp. 369-372, 2005.

[6] M. Marzinzik, B. Kollmeier, “Speech pause detection for noise spectrum estimation by tracking power envelope dynamics,” IEEE Transactions on Speech and Audio Processing, vol. 10, pp. 109-118, 2002.

[7] J. Haigh, J. Mason, “A voice activity detector based on cepstral analysis,” in Proc. Eurospeech, pp. 1103-1106, 2003.

[8] E. Nemer, R. Goubran, S. Mahmoud: Robust voice activity detection using higher-order statistics in the LPC residual domain, IEEE Transactions on Speech and Audio Processing, vol. 9, pp. 217-231, 2001.

[9] K. Sakhnov, E. Verteletskaya and B. Simak, "Low-Complexity Voice Activity Detector Using Periodicity and Energy Ratio," 2009 16th International Conference on Systems, Signals and Image Processing, Chalkida, Greece, 2009, pp. 1-5.

[10] J. Ramirez, J. Segura, M. Benitez, L. Garcia, “A. Rubio: Statistical voice activity detection using a multiple observation likelihood ratio test,” IEEE Signal Processing Letters, vol. 12, pp. 689-692, 2005.

[11] P. Ghosh, A. Tsiartas, S. Narayanan, “Robust voice activity detection using long-term signal variability,” IEEE Transaction Audio Speech Language Processing, vol. 19, pp. 600-613, 2011.

[12] S. L. Miller and D. Childers, CHAPTER 12- Simulation Techniques, Probability and Random Processes (Second Edition), Academic Press, 2012, pp. 517-546.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 8.317|

||Volume 12, Issue 5, May 2023||

|DOI:10.15662/IJAREEIE.2022.1205018 |

[13] J-C Junqua, "The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex," Speech Communication, vol. 20, pp.13–22, 1996.

[14] J. Pang, "Spectrum energy based voice activity detection," *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, 2017, pp. 1-5.

[15] S. Morita, X. Lu and M. Unoki, "Signal to noise ratio estimation based on an optimal design of subband voice activity detection," The 9th International Symposium on Chinese Spoken Language Processing, Singapore, 2014, pp. 560-564.

[16] https://www.analyticsvidhya.com/blog/2022/01/analysis-of-zero-crossing-rates-of-different-music-genre-tracks/.

[17] R.G. Bachu, S. Kopparthi, B. Adapa and B.D. Barkana, "Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy ," In: Elleithy, K. (eds) Advanced Techniques in Computing Sciences and Software Engineering. Springer, pp. 1-7.

[18] W.J. Hess, "Pitch Determination of Speech Signals — A Survey," n: Simon, J.C. (eds) Spoken Language Generation and Understanding. NATO Advanced Study Institutes Series, vol 59. Springer, pp.263-278.

[19] DJ Nelson, J Pencak, "Pitch-based methods for speech detection and automatic frequency recovery," in *Proc. of SPIE's 1995 International Symposium on Optical Science, Engineering, and Instrumentation*. (InternationalSociety for Optics and Photonics, San-Diego, California, USA, 1995).

[20] T Kristjansson, S Deligne and P Olsen, "Voicing features for robust speech detection," in Proc. of INTERSPEECH. (ISCA, Lisbon, Portugal, 2005).

[21] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," in IEEE Transactions on Speech and Audio Processing, vol. 7, no. 3, pp. 333-338, May 1999.

[22] S. A. Alim and N. K. Rashid, "Some Commonly Used Speech Feature Extraction Algorithms," From Natural to Artificial Intelligence, 2018.

[23] P. P. Kumar, K.S.N Vardhan and K.S.R. Krishna, "Performance evaluation of MLP for speech recognition in noisy environments using MFCC & wavelets," International Journal of Computer Science & Communication (IJCSC), vol.1, 2010, pp.41-45.

[24] M.A. Al-Alaoui, L. Al-Kanj, J. Azar, E. Yaacoub, "Speech recognition using artificial neural networks and hidden Markov models," IEEE Multidisciplinary Engineering Education Magazine, vol. 3, 2008, pp.77-86.

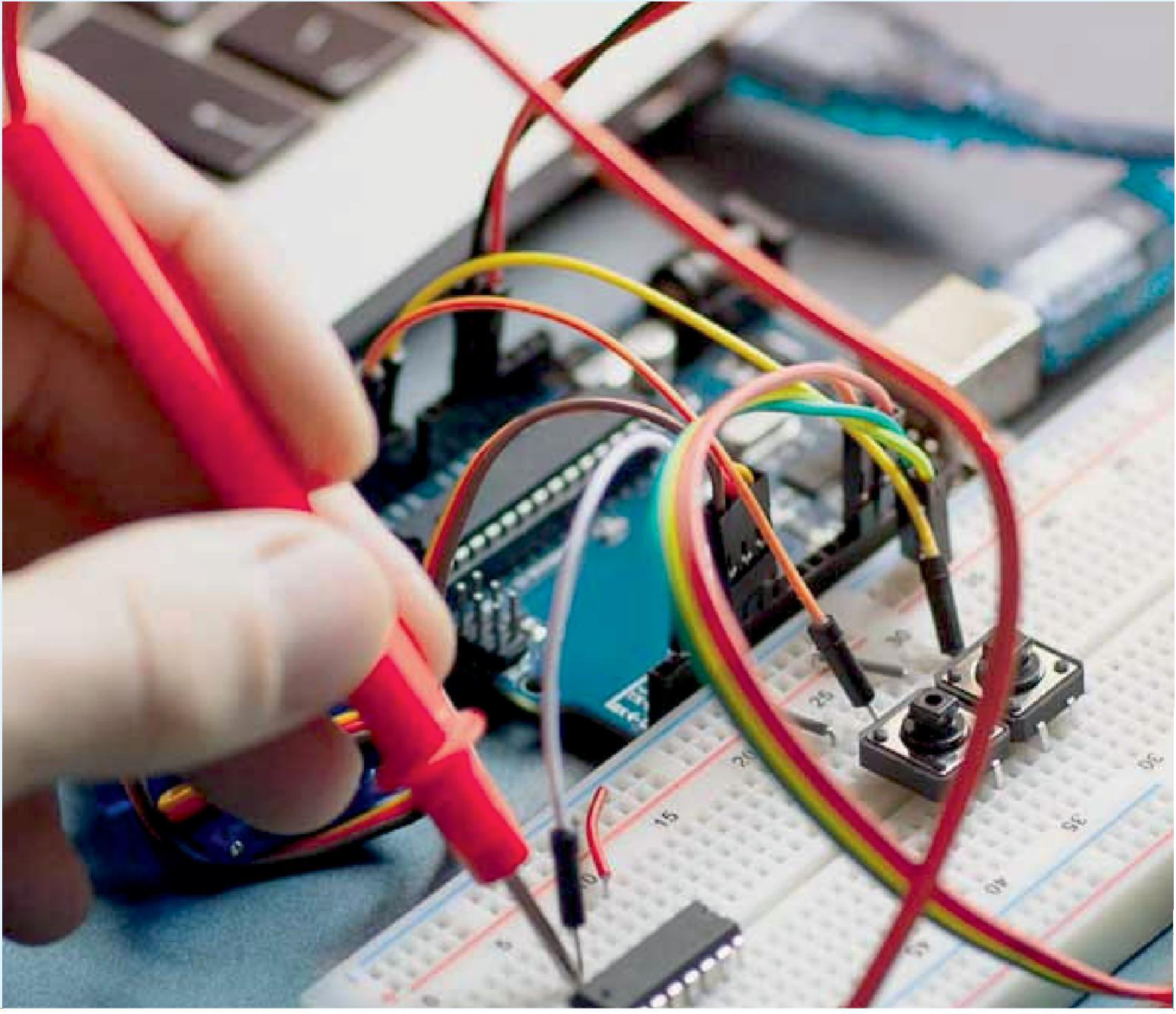
[25] A.M. Othman, M.H. Riadh, "Speech recognition using scaly neural networks," World academy of science. Engineering and Technology. vol. 38: pp.253-258, 2008.

[26] B. S. Atal, "Speech Synthesis Based on Linear Prediction," Encyclopedia of Physical Science and Technology (Third Edition), Academic Press, 2003, pp. 645-655.

[27] S. Alim, Alang and N.K. Rashid, "Some Commonly Used Speech Feature Extraction Algorithms," 2018, 10.5772/intechopen.80419.

[28] F. Ernawan, N. Abu and N. Suryana, "Spectrum analysis of speech recognition via discrete Tchebichef transform," in Proc. of SPIE - The International Society for Optical Engineering," 2011.

[29] https://www.kaggle.com/datasets/lazyrac00n/speech-activity-detection-datasets



INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.317

International Journal of Advanced Research

in Electrical, Electronics and Instrumentation Engineering

9940 572 462 6381 907 438 ijareeie@gmail.com

Scan to save the contact details