Voice Activity Detection using windowing and updated K-Means Clustering Algorithm

Shilpa Sharma
Department of Computer Science & Engineering
CT University
Ludhiana, India
shilpa13891@gmail.com

Anurag Sharma
Department of Computer Science & Engineering
GNA University
Phagwara, India
er.anurags@gmail.com

Rahul Malhotra
Department of Electronics Communication Engineering
CT Group of Institutions
Jalandhar, India
blessurahul@gmail.com

Punam Rattan
Department of Computer Application
CT University
Ludhiana, India
Punamrattan@gmail.com

2021 2nd International Conference on Intelligent Engineering and Management (ICIEM) | 978-1-6654-1450-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICIEM51511.2021.9445371

***Abstract*—Voice Activity Detection (VAD) is a method of detecting speech and non-speech in noisy environments. Vaious methods for this purpose have also been proposed. In general, the research has been divided into supervised and unsupervised speech recognition and produced various algorithms to depict the occurring of speech signal. Research aims to examine window overlapping and detection of speech and non-speech segments. A speech signal seems to be a slowly non stationary signal, and its characteristics are short time constant when examined over a short span of time (between 10 and 30 ms). As a result, frames windowing is used to enable us to use a speech signal and interpret its characteristics. However, a widespread study is required in the selection of techniques from predefined VAD and problems and opportunities to increase research in the emerging region. The advantage of the new unsupervised K-means approach over the supervised method is that it will not have to pre-train classifiers and pre-know any previous knowledge about audio streams.**

Keywords—Speech Recognition, VAD, K-Means, Feature Extraction, Zero Crossing Rate, Mel-Frequency Cepstrum

I. INTRODUCTION

Speech recognition is a field that has a wide range of applications and uses in our everyday lives. In general, a speech recognizer is a device that can understand human speech and operate on it. It can be used in a car setting, for example, to voice control non-critical operations like dialing a phone number. Another scenario is on-board navigation, which presents the driving route to the driver and uses voice control to improve traffic safety. A different aspect of speech recognition is that it can be beneficial for people with functional disabilities or other types of handicaps to use voice modulation to make their everyday tasks easier. They could turn on/off the light switch, the coffee machine, and other household appliances with their voice. This leads to a debate regarding intelligent homes, in which these functions can be made accessible to both the general public and the handicapped [1]. The main goal of VAD is to separate noise and speech content from each other. This area has been used in many different fields including speech enhancement, speech surveillance, speech coding, and speech recognition. The paper focuses on the discrimination paradigm and various feature extraction techniques. At first, more energy-based features were integrated into zero crossing rate (ZCR) technology that was highly affected by additive noise. To combat the additive interference effects, several other techniques have been implemented by researchers such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Prediction Residual, Line Spectral Frequencies and Periodic Features. Besides these mathematical models, discrete Fourier transform, such as the one depicted, have also been suggested. However, only a few studies have examined the variability in speech and noise parts [15] while hybrid methods [17-18] have been cited. After that, artificial intelligence came into play [15-16] to overcome the shortcomings of traditional speech recognition systems with Gaussian Mixture Models based Hidden Markov Models to depict spoken word. The peaks found in speech signal is shown in figure 1.

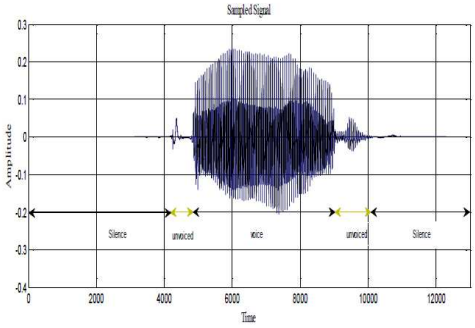

Fig. 1. Peaks found in speech signal [43]

This paper presents a comprehensive study on the various VAD techniques with overview of speech recognition and comprise of a suitable theoretical background to understand the topic. Authors believe that this paper will be help future researcher to identify new research directions to fill the gap in existing VAD approaches.

978-1-6654-1450-0/21/$31.00 ©2021 IEEE

II. BACKGROUND

The major field of VAD is speech recognition which includes various types of feature such as speaker recognition, emotional recognition, age identification and gender recognition. The design of a speech recognition device shows its general architecture in figure 2. An analysis of some studies to gain awareness of how speech recognition issues can be addressed today. The first attempts to design systems for automatic speech recognition by machine were made in the 1950s, when various researchers tried to exploit the fundamental idea of acoustic –phonetics. In 1952, at Bell laboratories, Davis Biddulph and Balashek built a system for isolated digit recognition for a single speaker. Another attempt was made by Forgie in 1959, when ten vowels embedded in the a-b-/vowel/-t format were recognized in a speaker independent manner at MIT Lincoln laboratories. The field of isolated word or discrete utterance recognition became a viable and functional technology in the 1970s, thanks to fundamental studies by Velichko and Zagoruyko in Russia, Sakoe and Chiba in Japan, and Itakura in the United States. Russian research aided in the advancement of pattern recognition ideas in speech recognition, Japanese research demonstrated how complex programming methods could be successfully implemented, and Itakura's research demonstrated the concept of linear predicting coding (LPC) [42].

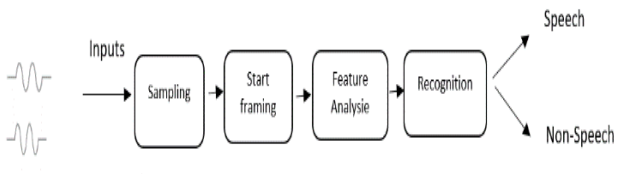

Fig. 2. Design of speech recognition system

Speech recognition has many different uses in the field of computers, instruction instead of texting, space stations, disabled persons, Smart house, and more. Generally, automated speaker identification is divided into two phases of speech recognition and speaker recognition. The field of emotion recognition are discussed in [22-23] and [24-25]. Speech recognition systems have additional applications in areas such as expression recognition where language is recognized while language is challenging. Another new field of VAD is age estimation and/or gender identity in recognition of voice. VAD algorithms fall into the following general categories: feature extraction, and classifier detection schemes.

III. PROPOSED TECHNIQUE

Speech is a non-stationary signal due to deviations in phoneme's spectral characteristics, changes in prosody, and spontaneous variations in the vocal tract. However, since the speech signal is thought to be stationary over a short time period (generally 10 to 20 ms), it is analyzed over these shot-time windows. As a result, the frame blocking technique entails splitting the speech signal into N-sample short frames that overlap by M samples with neighboring frames.

Every frame is multiplied with a window to reduce spectral distortions when blocking the speech signal. This signal processing function reduces the amplitude at the centre part of the signal and increases it at the edges.

- To choose the frame size (n), and N successive frames which are separated by m.
- e.g. An 12 KHz sampling, 8ms window has 322 samples, (neighboring shift) 132 samples

A. Frame and overlapping

Because our ears can't respond to very rapid changes in voice, we normalize the data by converting it into frames. Frames are .WAV files that can be overlapped. The overlap range is between 0 and 75% of the frame size and Frame size is 15-35ms. Figure 3 and figure 4 schematically illustrate the function window multiplication by signal and overlapping would not result in any information loss.

B. Window shifting

It is typical to overlap windows when performing a transform over time. A quarter of the window size is recommended. Windowing is necessary in order to avoid of loss of information.

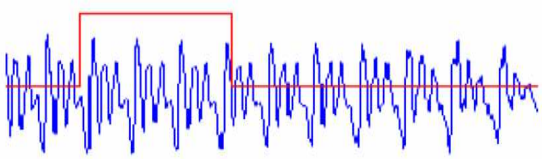

Fig. 3. Windowing cross-multiplication by signal

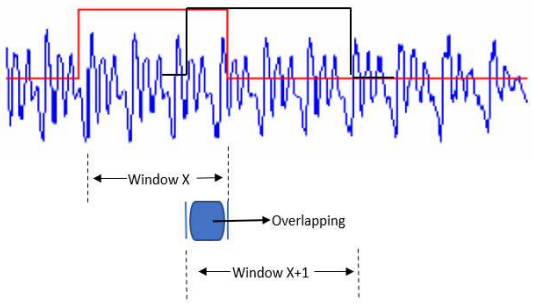

Fig. 4. Window Overlapping

Researchers have recently shown an interest in improving the unsupervised approaches, such as Gaussian Mixture

Models [37]. This approach incorporates likelihood ratio and short-time energy function detection based on expression. On the other hand, a VAD technique based on vector thresholding is suggested in [37]. Additionally, deep multi - modal end to end architectures, visual and audio network representation, and diffusion networks based system against transient noise, and others emerged as recent innovations in the domain of VAD detection. Additionally, a new K-Means VAD-based scheme has been proposed which are efficient in transient and noisy environments. In this K-Means technique, means and deviant methods have been applied.

The scheme in the figure 5 which is a proposed methodology to detect noise in a speech signal. The automation based approach will be used to categories fractal dimensions, which will be changed automatically based on the audio being tested. On a short time scale, the audio signal stays the same, so we believe it doesn't shift. To this end, this is why we divide the signal into frames which have a length between 20 and 40 milliseconds If the sample rate is too low, we will not be able to get an accurate spectral calculation, if it is too high, the signal changes drastically during the acquisition.

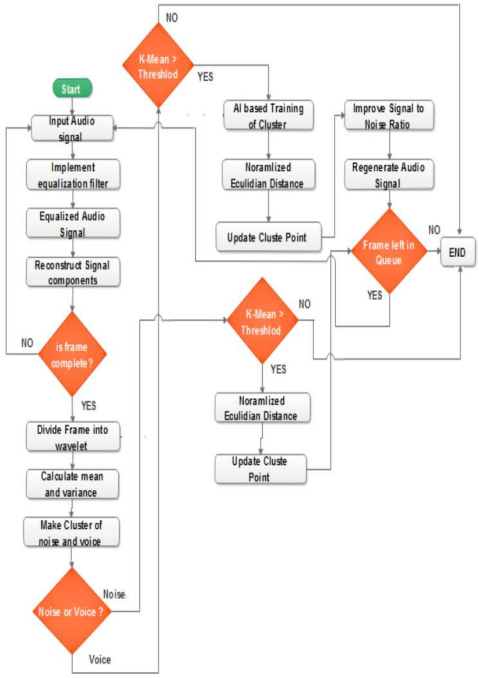

Fig. 5. VAD K-Means clustering Methodology

IV. Comparison Of Proposed Technique With Existing Techniques

In sound processing, the mel-frequency cepstrum (MFC) is a reflection of the short-term power spectrum of a logarithm spectra (sinusoidal) cosine transform. MFCCs are used in speech recognition systems, such as recognition applications. If MFCC values are above the threshold, they can be adversely affected by additive noise, and so it is best to normalize their values in speech recognition systems. Some researchers suggest various modifications to the MFCC algorithm, such as raising the log-amplitude to a power (around 2 or 3) prior to reduce the effect of low-impact components. MFCCs are focused on known frequency fluctuations of the critical bandwidths. The key thing to understand about speech is that it passes through the form of a vocal tract, including lips, etc. If the shape is known, the sound should be as well, this should be an accurate depiction of it [44]. TIMIT is used for VAD performance in the 285 and 284 speakers samples. Sadjadi algorithm is purely based on features that function as proxies for fundamental traits of speech, which are regulated by the speech development process and in dry run testing of the DARPA project's Phase I dry run SPINE data, (on the SPINE corpus) [45] for both simulation and real data.

Figures 6,7 and 8 illustrates the SNR differences among the existing and proposed techniques. In certain cases, updated K-means clustering can be used to complement existing speech processing technique. Even if speech is segmented into different features, this algorithm performs well. Less computation and memory are needed.

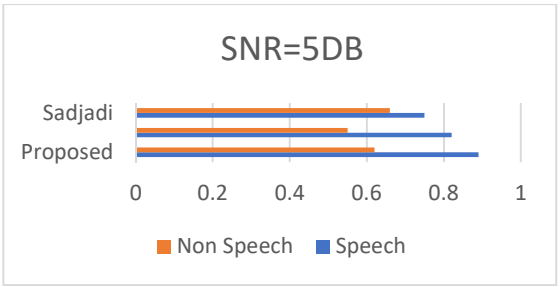

Fig. 6. Speech and Non-Speech recognition under SNR 5 DB

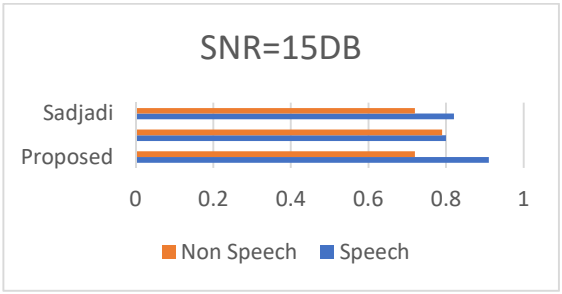

Fig. 7. Speech and Non-Speech recognition under SNR 15 DB

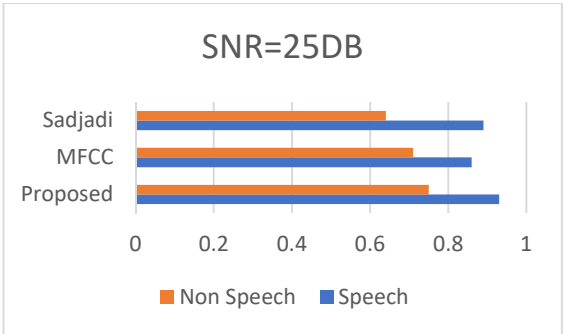

Fig. 8. Speech and Non-Speech recognition under SNR 25 DB

Speech and non- speech signals found in the different SNR ratio are more in the algorithms proposed than in the MFCC and Sadjadi VAD.

V. ISSUES AND OPPORTUNITIES

The various VAD architectures and algorithms have been developed to improve the robustness to different types of noise as it is proposed as most critical issue since there are large varieties of noises which behave differently and can significantly affect the performance of any algorithm. One of the commonly used noise estimators may not provide a clear list of what noise types are in nature. Therefore, the demand for audio classification has made it necessary to pursue this topic at present. Lastly, multimodal VAD is in its early stages, and needs to be investigated and based on specifically. The findings of the research were the most surprising. Academician and research staff still use M-Fourier Coefficients for feature extraction in deep learning environment. On the contrary, a classifier such as MFCC's is a classical classifier[30]. Thus, it is important to incorporate other feature extraction methods such as linear predictive coding (LPC). This research is most important and timely in the study of VAD. A study found that seventy-five percent of DNN models are standalone, while approximately twenty-five percent of models emphasize hybrid techniques. This creates a wide scope for researchers and experts of the VAD sector to use hybrid models for clinical trials. Hence, authors believe hybrid model methodology would increase the robustness of VAD architectures to environmental noise. However, it is often noted that none of the. researches were conducted in the direction of recurrent neural networks (RNN). Other than LSTM, Long Short-Term Memory has also proved to be an effective tool for speech recognition and voice Authentication. As computational complexity of any algorithm play an important role, hence to reduce the complexity of already available methods can be seen as future path, even though many attempts have already been made.

Today the most common applications for artificial intelligence, deep learning and machine learning are increasing rapidly. The deployment of these technologies in VAD is premature which requires more testing to evaluate the outcomes. Most of the work in VAD is focused on mathematical models and testing parameters on synthetic data, but very little VAD architectures are evaluated in actual environments. Thus, businesses can incorporate and evaluate these mathematical models in real-time and real world environments.

VI. CONCLUSION

The information will assist the specialist in choosing a research field to enhance robustness against noisy environments and VAD. VAD is an important part of speech communication; thus, the accuracy of VAD selection is most challenging in terms of sophistication, feature extraction, and threshold selection. The experimental result has been computed and drawn from it using MATLAB concludes that we can have high speech and non-related recognition accuracy in the proposed K-means method. The authors believe that this modified K-means clustering approach can complement other speech detection methods.

REFERENCES

[1] Mc Cowan, D. Dean, M. McLaren, R. Vogt, S. Sridharan: The Delta Phase Spectrum With Application to Voice Activity Detection and Speaker Recognition, IEEE. trans. Audio Speech Lang. Proc., vol. 19, pp. 2026- 2038, 2011.

[2] D. Valj, B. Kotnik, B. Horvat, Z. Kacic: A Computationally Efficient Mel Filter Bank VAD Algorithm for Distributed Speech Recognition Systems, Eurasip J. Appl. Signal Processing, no. 4, pp. 487-497, 2005.

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

[5] J. Haigh, J. Mason: A voice activity detector based on cepstral analysis, Proc. Eurospeech, pp. 1103-1106, 2003

[6] S.O. Sadjadi, J. Hansen: Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux, IEEE Sig. Pro. Letters, vol. 20, pp. 197-200, 2013.

[7] M. Marzinzik, B. Kollmeier: Speech pause detection for noise spectrum estimation by tracking power envelope dynamics, IEEE Trans. Speech Audio Process., vol. 10, pp. 109-118, 2002

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

[9] K. Ishizuka, T. Nakatani: Study of Noise Robust Voice Activity Detection Based on Periodic Component to Aperiodic Component Ratio, in Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition, pp. 6570, 2006.

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

[11] P. Ghosh, A. Tsiartas, S. Narayanan: Robust voice activity detection using long-term signal variability, IEEE Trans. Audio Speech Lang. Process., vol 19, pp. 600-613, 2011.

[12] Y. Kida, T. Kawahara: Voice Activity Detection based on Optimally Weighted Combination of Multiple Features, in Proc. Interspeech, pp. 2621-2624, 2005.

[13] S.Soleimani, S. Ahadi: Voice Activity Detection based on Combination of Multiple Features using Linear/Kernel Discriminant Analyses, in Proc. Information and Communication Technologies: From Theory to Applications, pp. 1-5, 2008.

[14] H. Singh and A. K. Bathla, ''A survey on speech recognition,'' Int. J. Adv. Res. Comput. Eng. Technol., no. 2, no. 6, pp. 2186–2189, 2013.

[15] M. A. Anusuya and S. K. Katti, ''Speech recognition by machine: A review,'' Int. J. Comput. Sci. Inf. Secur., vol. 6, no. 3, pp. 181–205, 2009.

[16] J. Padmanabhan and M. J. J. Premkumar, ''Machine learning in automatic speech recognition: A survey,'' IETE Tech. Rev., vol. 32, no. 4, pp. 240–251, 2015.

[17] Chung-Ching Shen, William Plishker, and Shuvra S. Bhattacharyya, "Design and Optimization of a Distributed, Embedded Speech Recognition System", In Proceedings of the International Workshop on Parallel and Distributed Real-Time Systems, Miami, Florida, April 2008.

[18] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, ''Nonlinear feature based classification of speech under stress,'' IEEE Trans. Speech Audio Process., vol. 9, no. 3, pp. 201–216, Mar. 2001.

[19] C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni, and A. Ghio, ''Application of Automatic Speaker Recognition techniques to pathological voice assessment (dysphonia),'' in Proc. Eur. Conf. Speech Commun. Technol. (Eurospeech), 2005, pp. 149–152.

[20] V. A. Petrushin, ''Emotion recognition in speech signal: Experimental study, development, and application,'' in Proc. 6th Int. Conf. Spoken Lang. Process. (ICSLP), 2000, p. 5

[21] N. Fragopanagos and J. G. Taylor, ''Emotion recognition in human– computer interaction,'' Neural Netw., vol. 18, no. 4, pp. 389–405, 2005.

[22] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, ''Emotional speech: Towards a new generation of databases,'' Speech Commun., vol. 40, nos. 1–2, pp. 33–60, 2003

[23] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan and R. Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM SPINE evaluation system," Proc. ICASSP, 1, pp. 53-56, 2002.

[24] T. Kristjansson, S. Deligne and P. Olsen, "Voicing features for robust speech detection," Proc. Interspeech, pp. 369-372, 2005.

[25] ETSI standard document, ETSI ES 202 050 V 1.1.3., 2003.

[26] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," IEEE Trans. Speech Audio Process, 10, pp. 109- 118, 2002.

[27] K. Li, N. S. Swamy and M. O. Ahmad, "An improved voice activity detection using higher order statistics," IEEE Trans. Speech Audio Process., 13, pp. 965-974, 2005.

[28] G. D. Wuand and C. T. Lin, "Word boundary detection with mel scale frequency bank in noisy environment," IEEE Trans. Speech and Audio Processing, 2000.

[29] 29 A. Lee, K. Nakamura, R. Nisimura, H. Saruwatari and K. Shikano, "Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs," Interspeech, pp. 173-176, 2004.

[30] B. Lee and M. Hasegawa-Johnson, "Minimum Mean Squared Error A Posteriori Estimation of High Variance Vehicular Noise," in Proc. Biennial on DSP for In-Vehicle and Mobile Systems, Istanbul, Turkey, June 2007.

[31] ETSI ES 202 050 recommendation, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," 2002.

[32] Juang, C.F., Cheng, C.N. and Chen, T.M., 2009. Speech detection in noisy environments by wavelet energy-based recurrent neural fuzzy network. Expert Systems with Applications, 36(1), pp.321-332.

[33] Wang, K.C. and Tasi, Y.H., 2008, December. Voice activity detection algorithm with low signal-to-noise ratios based on spectrum entropy. In Universal Communication, 2008. ISUC'08. Second International Symposium on (pp. 423-428).

[34] Kim, S.K., Kang, S.I., Park, Y.J., Lee, S. and Lee, S., 2016. Power Spectral Deviation-Based Voice Activity Detection Incorporating Teager Energy for Speech Enhancement. Symmetry, 8(7), p.58.

[35] Germain, F.G., Sun, D.L. and Mysore, G.J., 2013, August. Speaker and noise independent voice activity detection. In Interspeech (pp. 732-736).

[36] Kinnunen, T. and Rajan, P., 2013, May. A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data. In ICASSP (pp. 7229-7233).

[37] I. Ariav and I. Cohen, "An end-to-end multimodal voice activity detection using wavenet encoder and residual networks," IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 265–274, 2019.

[38] A. Ivry, B. Berdugo, and I. Cohen, "Voice activity detection for transient noisy environment based on diffusion nets," IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 254–264, 2019.

[39] H. Sharma, N. Kanwal and R.S. Batth, "An Ontology of Digital Video Forensics: Classification, Research Gaps & Datasets," 2019 in International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dubai, United Arab Emirates, 2019, pp. 485-491.

[40] H. Dubey, A. Sangwan, and J. H. Hansen, "Leveraging frequency dependent kernel and dip-based clustering for robust speech activity detection in naturalistic audio streams," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 11, pp. 2056– 2071, 2018.

[41] Wang, G.-B., & Zhang, W.-Q. (2019). An RNN and CRNN Based Approach to Robust Voice Activity Detection. 2019. doi:10.1109/apsipaasc47483.2019.9023320

[42] Sandhu AK, Batth RS. Software reuse analytics using integrated random forest and gradient boosting machine learning algorithm. Softw: Pract Exper. 2020;1–13

[43] Hamid, O. K. (2018). Frame blocking and windowing speech signal. Journal of Information, Communication, and Intelligence Systems (JICIS), 4, 87-94.

[44] Lokhande, N. N., Nehe, N. S., & Vikhe, P. S. (2012, March). Voice activity detection algorithm for speech recognition applications. In IJCA Proceedings on International Conference in Computational Intelligence (ICCIA2012), vol. iccia (No. 6, pp. 1-4).

[45] Nayyar, A., Batth, R.S., Ha, D.B., Sussendran, G., 2018, Opportunistic Networks: Present Scenario- A Mirror Review, International Journal of Communication Networks and Information Security 10.

[46] Pakyurek, M., Atmis, M., Kulac, S., & Uludag, U. (2020). Extraction of Novel Features Based on Histograms of MFCCs Used in Emotion Classification from Generated Original Speech Dataset. Elektronika ir Elektrotechnika, 26(1), 46-51.

[47] Sadjadi, S. O., & Hansen, J. H. (2013). Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Processing Letters, 20(3), 197-200.