International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Leveraging Intelligent Voice Activity Detection to Elevate Speech Recognition Systems

[1]Dr. Anita Mundra, [2]Dr. N. Bargavi, [3]Dr. Ritesh Sharma, [4]Dr. Renu Vij, [5]Dr. Melanie Lourens, [6]Charanjit Singh, [7]Dr. Anoop Beri

Submitted: 10/10/2023 **Revised**: 29/11/2023 **Accepted**: 09/12/2023

Abstract: Research is conducted on Automatic Speech Recognition (ASR) that is practical for use in noisy conditions. The effectiveness of common parameterization strategies was evaluated in comparison to the background signal in terms of lustiness. For Mel frequency cepstral coefficients (MFCC), Perceptual linear predictive (PLP) coefficients, and their modified versions, a hybrid feature extractor is employed by merging the basic blocks of PLP and MFCC. Only during the training phase of the ASR method was the VAD-based frame dropping formula used. The benefit of using this technique is that it eliminates pauses and possibly severely distorted speech parts, which helps with more accurate phone modeling. The analysis and contribution of the modified vocal activity detection technique are the focus of the second part.

Keywords: *Speech Presence Probability, Mel Cepstrum Frequency Coefficients, Perceptual Linear Prediction, Short-time Fourier Transform, Automatic Speech Recognition*

1. Introduction

Automatic speech recognition (ASR) is the process of using machines to modify a human speaker's string of words. It is desired that an ASR system be resilient to unpleasant fluctuation because the goal of ASR is to have speech as a substandard form of interaction between a machine and a human [5]. In speech recognition systems, endpoint detection caused by non-speech occurrences and background noise is typically troublesome [1]. When ambient acoustic noise is present, speech recognition systems that were trained in quiet contexts sometimes perform worse. The discrepancy between clear acoustic models and noisy speech data is typically what causes dilapidation. To reduce this mismatch and restore recognition accuracy in noisy environments, a lot of work has been done [2]. Noise robustness in automated speech recognition (ASR) can be discussed in a number of ways that are fundamentally different. One method is to train the system exclusively on a certain type of noise that it encounters during the recognition phase. This type of system is known as a matched system, and it is likely to be better than many other techniques of noise compensation—but only for that particular form of noise. A large library of new noise kinds and an excruciatingly long re-training process are required to adapt the system to new sorts of noises. Multi condition training, which trains the system on noisy speech heard in the loudest noise conditions and avoids the need to retrain the system every time the background noise changes [5], is a more practical alternative to matched training.

2. Related Work

Qi Li et al. [1] discuss the endpoint issue and suggest a timeline strategy. For endpoint detection, it employs an association best filter and a three-state transition diagram. Many criteria are being used by the proposed filter to assure accuracy and strength. A noise-strong feature compensation (FC) formula supporting polynomial regression of vocalization signal-to-noise ratio (SNR) is planned by **Xiaodong Cui et al. [2]**. The expectation maximization (EM) formula, together with the most probability (ML) criterion, can be used to calculate a set of polynomials that approximate the bias between clean and noisy speech alternatives. **Kapil Sharma et al. [3]** propose a comparative examination of various feature extraction methods for isolated word end detection in noisy situations. We tested the cases of

[1]Takshila Institute of Engineering and Technology, Jabalpur (MP). anitamundra26@gmail.com

[2]Assistant Professor (Senior Grade), Faculty of Management, SRM Institute of Science and Technology, Vadapalani Campus, Chennai. divu209@gmail.com

[3]Assistant Professor, GLA University, Mathura, Uttar Pradesh. ritesh.sharma@gla.ac.in

[4]Associate Professor, University School of Business, Apex Institute of Technology (MBA), Chandigarh University, Gharun, Mohali, India. Email ID: renuvij@gmail.com

Orchid ID: https://orcid.org/0000-0001-9202-8390. Scopus ID: 58362729500

[5]Deputy Dean Faculty of Management Sciences, Durban University of Technology, South Africa. Email: melaniel@dut.ac.za. Orcid: 0000-0002-4288-8277

[6]Assistant Professor, Department of Computer Science & Engineering, MM Engineering College, Maharishi Markandeshwar (Deemed to be University), Mullana-Ambala, Haryana, India 133207. Charan2910@gmail.com

[7]Professor, School of Education, Lovely Professional University, Phagwara. email: beryanoop@rediffmail.com

colored noises, babbling noise, industrial plant noise at various SNR levels, and distortions caused by the recording media. **Tomas Dekens et al. [4]** demonstrates that in noisy circumstances, bone-conducted mics will not be able to enhance automatic speech recognition. Voice Activity Detection (VAD) was used using a throat mike signal as an input, and it was discovered that this significantly improved recognition accuracy in non-stationary noise compared to when VAD is conducted on a typical mike signal. **Sami Keronen et al. [5]** a comparison of three essentially unrelated noise-strong techniques is carried out. In an extremely large vocabulary continuous speech recognition system, the effectiveness of multi-condition training, Data-driven Parallel Model Combination (DPMC), and cluster-based missing information reconstruction methods is assessed. **M. G. Sumithra et al. [6]** the speech signal is strengthened and the background noise is removed using a Kalman filter. To ensure strong performance under shouting environment settings, the upgraded signal is integrated into the front part of the recognition system. **Lamia BOUAFIF et al. [7]** demonstrate a set of academic software programmes for signal and speech processing. This interface, which was created using Matlab, can be used for speech recognition, writing, and signal denoising.. **Md. Mahfuzur Rahman et al. [8]** Utilizing Cepstral Mean standardization (CMS) for strong feature extraction, we construct a distributed speech recognizer for noise that is robust enough for use in practical applications. The majority of the effort is devoted to managing a variety of noisy settings. By using a first-order all-pass filter rather than a unit delay, Mel-LP based speech analysis has been used in speech coding on the linear frequency scale to achieve this goal. **Stephen J. Wright et al. [9]** gives more information on specific application challenges in (machine translation) MT, speaker/language recognition, and automatic voice recognition while outlining the range of problems in which optimization formulations and algorithms play a role. **Namrata Dave et al. [10]** Speech selections are taken from male or female speakers' recorded speech and compared to templates in the database. Linear Predictive Codes (LPC), Perceptual Linear Prediction (PLP), Mel Frequency Cepstral Coefficients (MFCC), PLP-RASTA (PLP-Relative Spectra), etc. will be used to parameterize speech. **Eric W. Healy et al. [11]** Monophonic (single-microphone) algorithms that can improve speech comprehension in noisy environments have eluded researchers despite significant effort. Given their unique issue with hissing backgrounds, hard-of-hearing (HI) listeners require the no-hit construction of such an associate degree algorithmic rule. To distinguish speech from noise in the current work, binary masking backed by an algorithmic method was devised. **Deividas Eringis et al. [12]** report the findings of a study on the impact of frame shift and study window length on voice recognition rate. Analyzed three entirely separate cepstral analysis methods—mel frequency cepstral analysis (MFCC), linear prediction cepstral analysis (LPCC), and perceptual linear prediction cepstral analysis—for this goal (PLPC).**Jürgen T. Geiger et al. [13]** discuss remote voice recognition in jingling shire situations. ASR systems can be strengthened by using speech enhancement techniques such abuse of non-negative matrix resolution (NMF). **Taejin Park et al. [14]** suggest a feature extraction method that is resilient to unstable settings. The weighted bar graph of the time-frequency gradient in a very Mel picture image serves as the foundation for the anticipated theme. **Roger Hsiao et al. [15]** Creating a superior system without having access to the right training and development information is the challenge's key feature. The training information involves phone voice and near talking, as opposed to the analysis data, which are recorded using far-field microphones in noisy, bright environments. **Colleen G. Le Prell et al. [16]** Speech communication generally occurs in yelling situations; this can be an urgent problem for military personnel who must communicate in loud settings. Depending on the origins of the noise, the volume and types of talkers, and the listener's hearing capacity, there are a variety of effects of noise on speech recognition. **Ashrf Nasef et al. [17]** A challenging problem is still finding voice recognition software that can be used in noisy locations, such as workplaces, cars, planes, and other places. Even if deep learning algorithms perform better, the task of speaker recognition in noisy contexts still suffers from an oversized recognition loss. **Raviraj Joshi et al. [18]** in the context of voice search functionality on the Flipkart e-Commerce platform, suggest automated speech recognition (ASR). Used Listen-Attend-deep Spell's learning architecture (LAS)

3. Proposed Work

The noise power spectrum estimate approach is based on Speech Presence Probability (SPP). A subpar calculation of the first 20 frames of the speech spectrum is used to approximate the sound power spectrum in this instance. The goal of this effort is to make speech recognition systems more resilient to real-time reverberant situations. Using Mel Cepstrum Frequency Coefficients (MFCCs) and Perceptual Linear Prediction (PLP) for feature extraction, a speech recognition system for the set of Hindi letters is presented. It includes a method for improving the auditory spectrum and a short-time feature standardization technique that reduces the variance of cepstral features between the training and test environments by adjusting the balance and mean of cepstral features. The first step in treating a voice signal is preprocessing (pre-emphasis, normally by a second order high-pass filter). Using a set duration, short-time Fourier transform (STFT) examination is accomplished (40 ms). To estimate the signal's power spectrum, utilize the Hamming window. In this attempt, 750 samples are evaluated one at a time after being educated on various noises. The accuracy rate will be determined based on their training and examination of these samples.

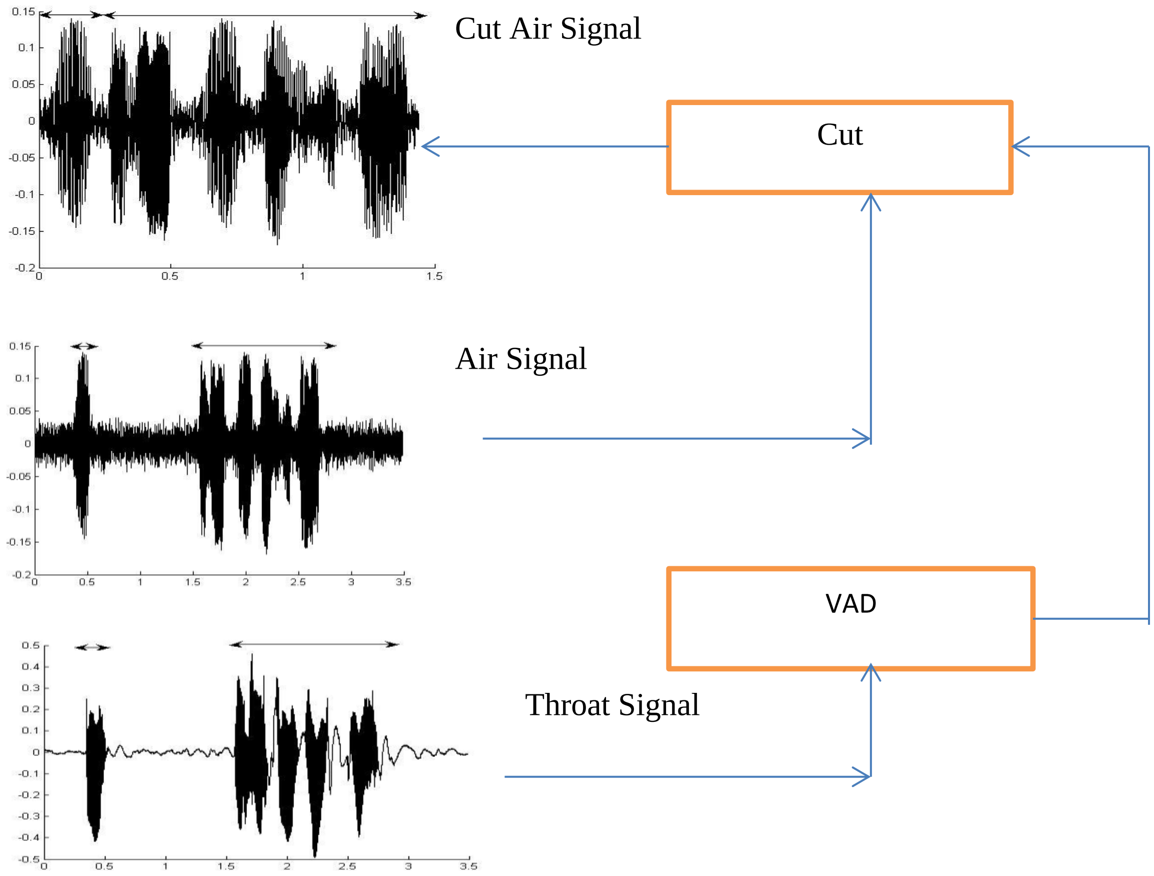

Fig 3.1 Experimental Setup

For these experiments, we used three different noise types: fan, automobile, and diesel engine noise. These noises are static in nature. In order to determine whether the event corresponds to a talking user, VAD evaluates the signal by watching at the energy levels, length, and frequency content of trials. There is some low-frequency noise present in the throat microphone's data, but there is no high-frequency speech energy. For this reason, the energy ratios for the VAD were computed using the energy in the range [250 5000] Hz. A speaker was the source of the ambient noise.

3.1 Algorithm

Step 1: Apply the silence indicator to identify the periods when the signal is inactive, and then update the noise scales during these periods.

Step 2: Use the Short-time Fourier Transform (STFT) to convert a time-domain signal to a frequency-domain signal. A magnitude operator follows the STFT.

Step 3: Apply a high pass filter (HPF) to reduce noise variation; the HPF's goal is to lessen processing errors brought on by noise variances.

Step 4: Use a post-processor to remove any processing errors caused by spectrum subtraction.

Step 5: Apply a Short-time Fourier Transform (ISTFT) to the treated signal in to convert it to a time-domain signal.

Step 6: The section is categorized as speech or non-speech using a grouping method. When a value exceeds a threshold, this grouping rule determines it. The classifier's output is a continuous number, but it is threshold to provide a result. A lot of information about the signal that was lost due to thresholding is contained in the continuous output. When the value is high, it is almost likely that the signal is speech, but when the value is close to the threshold, it is less definite. The classifier's output is a rough estimate of the likelihood that the input signal is speech.

Step 7: Create a set of features from the signal that are intended to look at characteristics that separate speech from non-speech.

Step 8: In a classifier, combine the evidence from the attributes to determine the likelihood that the signal is speech.

4. Results and Discussion

Using Voice Activity Detection, robust feature extraction reduces discrepancy between training and test phases.

The results of the detail recognition are displayed in this section. Table 4.1 lists the word accuracy for MFCC-PLP without applying a filter and is regarded as the baseline result. For the baseline, it is discovered that the average word accuracy over all noises within the SNR range of 15 to 0 dB is 55.2. Table 4.2 provides word correctness with filter. 65.6 percent on average is discovered to be the MFCC-PLP with filter's recognition ability. Fig. 4.3 shows how the proposed filter performs when subjected to various types of noise. Additionally, it has been noted that, when compared to baseline performance, diesel engine and fan noises show the greatest increases. For all noises, the average recognition accuracy dramatically increases. It has been noted that speech sounds between 15 dB and 0 dB show the greatest improvements in recognition accuracy.

Table 4.1: Word accuracy [%] without Filter

Noise	**SNR (db)**					**Average (15 db to 0 db)**
	Clean	15	10	5	0	
Car	97.5	86.4	67.3	43.5	23.6	55.2
Fan	96.3	85.2	66.5	41.7	21.8	53.8
Diesel Engine	95.4	84.5	61.3	37.6	12.2	48.9

Table 4.2: Word accuracy [%] with Filter

Noise	**SNR (db)**					**Average (15 db to 0 db)**
	Clean	15	10	5	0	
Car	99.1	90.9	78.5	55.8	37.2	65.6
Fan	98.2	92.7	76.5	57.1	35.3	65.4
Diesel Engine	96.9	93.1	76.4	53.5	30.7	63.425

Table 4.3: Performance of Proposed Filter

Noise	**Accuracy Without Filter**	**Accuracy With Filter**
Car	55.2	65.6
Fan	53.8	65.4
Diesel Engine	48.9	63.425

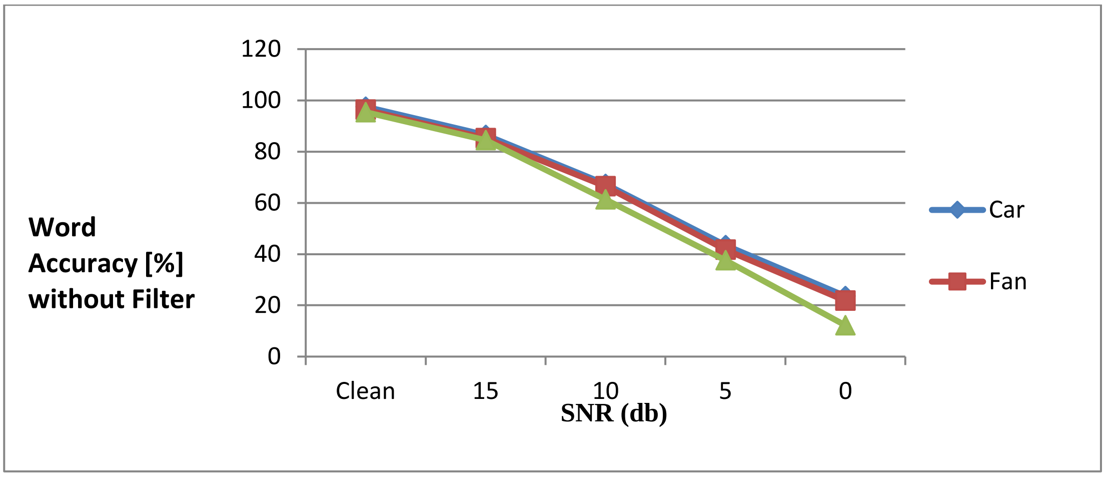

Fig 4.1: Word accuracy without filter

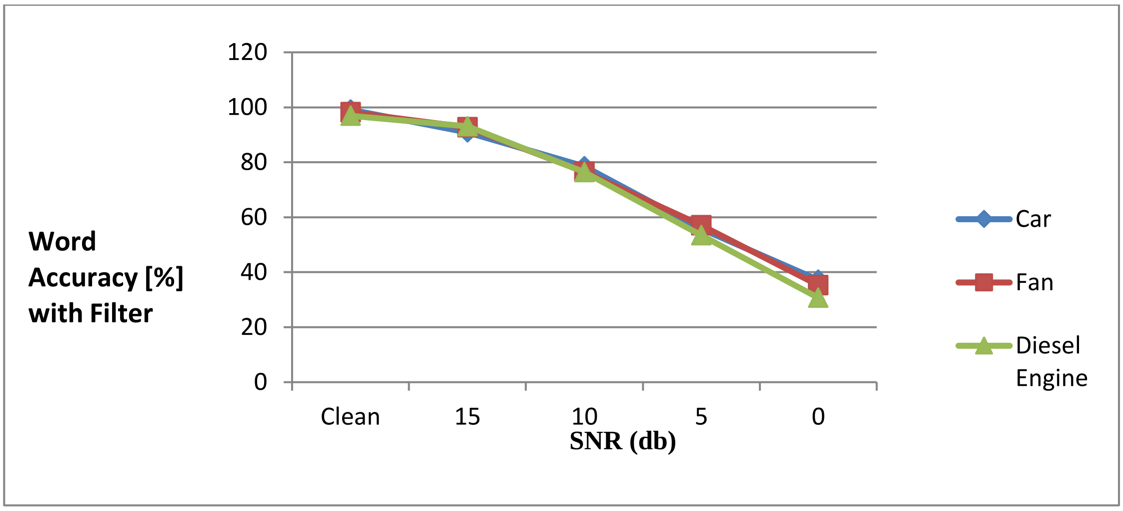

Fig 4.2: Word accuracy with filter

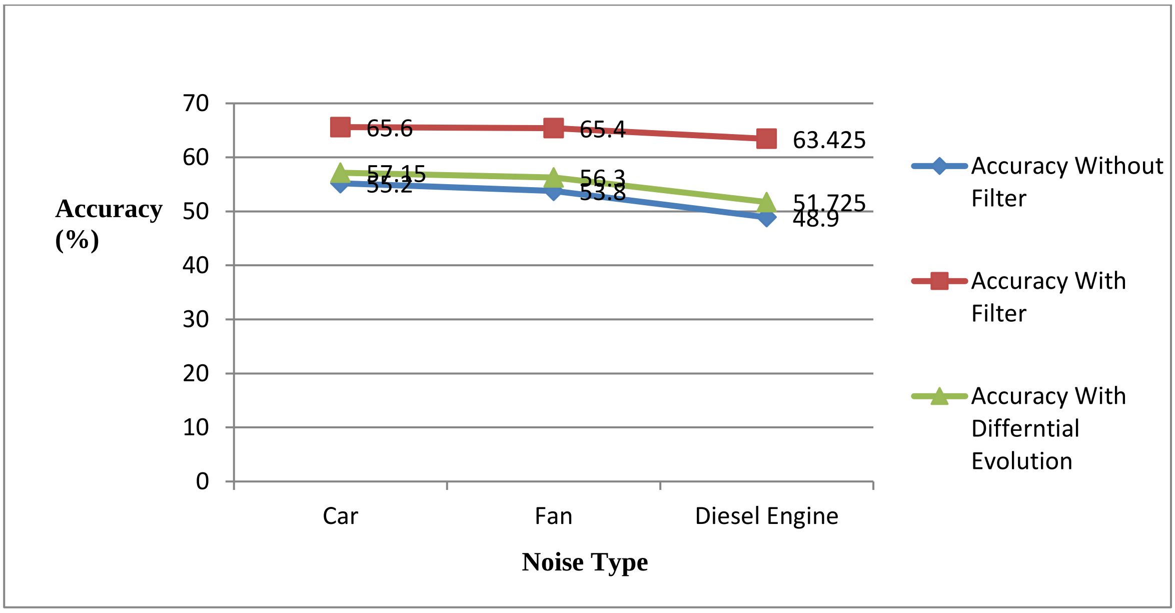

Fig 4.3: Word accuracy with and without filter and DE

5. Conclusion

Non-speech components of a signal are affected more strongly when speech is impacted by a corresponding environmental background. Even if completely different noise suppression techniques may increase the accuracy of the target ASR, such distortion is the cause of numerous flaws in the outcomes of ASR systems.

Typically, disrupted non-speech segments will be detected as speech, and this leads to an increase in WER as a result of the dangers of acoustic model standardization in the training section. This fact leads to the employment of the VAD rule as a frame dropping approach to remove potentially harmful non-speech parts from the processed signal. Depending on the circumstances, the VAD rule will eliminate the non-speech components of the signal, but frequently even a few frames that contained speech activity. This significant flaw will have a significant impact on how well targets are identified. However, the proposed VAD rule provides accuracy that is 16% higher.

References

[1] Qi Li, "Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition", IEEE Transactions on Speech And Audio Processing, Vol. 10, No. 3, March 2002, pp. 146-157

[2] Xiaodong Cui, "Noise Robust Speech Recognition Using Feature Compensation Based on Polynomial Regression of Utterance SNR", IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 6, November 2005, pp. 1161-1172

[3] Kapil Sharma, "Comparative Study of Speech Recognition System Using Various Feature Extraction Techniques", International Journal of Information Technology and Knowledge Management July-December 2010, Volume 3, No. 2, pp. 695-698

[4] Tomas Dekens, "Improved Speech Recognition In Noisy Environments By Using A Throat Microphone For Accurate Voicing Detection", 18th European Signal Processing Conference (EUSIPCO-2010), pp. 1978-1982

[5] Sami Keronen, "Comparison of Noise Robust Methods In Large Vocabulary Speech Recognition", 18th European Signal Processing Conference (EUSIPCO-2010), pp. 1973-1977

[6] M. G. Sumithra, "Speech Recognition In Noisy Environment Using Different Feature Extraction Techniques", International Journal of Computational Intelligence & Telecommunication Systems, 2(1), 2011, pp. 57-62

[7] Lamia BOUAFIF, "A Speech Tool Software for Signal Processing Applications", 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012, IEEE, pp. 788-791

[8] Md. Mahfuzur Rahman, "Performance Evaluation of CMN for Mel-LPC based Speech Recognition in Different Noisy Environments", International Journal of Computer Applications (0975 – 8887) Volume 58– No.10, November 2012, pp. 6-10

[9] Stephen J. Wright, "Optimization Algorithms and Applications for Speech and Language Processing", IEEE Transactions on Audio, Speech, And Language Processing, Vol. 21, No. 11, November 2013, pp. 2231-2243

[10] Namrata Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition", International Journal For Advance Research In Engineering And Technology, Volume 1, Issue VI, July 2013, pp. 1-5

[11] Eric W. Healy, "An algorithm to improve speech recognition in noise for hearing-impaired listeners", J. Acoust. Soc. Am. 134 (4), October 2013, pp. 3029-3038

[12] Deividas Eringis, "Improving Speech Recognition Rate through Analysis Parameters", doi: 10.2478/ecce-2014-0009, pp. 61-66

[13] Jürgen T. Geiger, "Memory-Enhanced Neural Networks and NMF for Robust ASR", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 22, No. 6, June 2014, Pp. 1037-1046

[14] Taejin Park, "Noise robust feature for automatic speech recognition based on Mel-spectrogram gradient histogram", 2nd Workshop on Speech, Language and Audio in Multimedia (SLAM 2014) Penang, Malaysia September 11-12, 2014, pp. 67-71

[15] Roger Hsiao, "Robust Speech Recognition In Unknown Reverberant And Noisy Conditions", 2015 IEEE, pp. 533-538

[16] Colleen G. Le Prell, "Effects of noise on speech recognition: Challenges for communication by service members", www.elsevier.com/locate/heares, Hearing Research 349 (2017), pp. 76-89

[17] Ashrf Nasef , "Optimization Of The Speaker Recognition In Noisy Environments Using A Stochastic Gradient Descent", International Scientific Conference On Information Technology And Data Related Research, Sinteza 2017, pp. 369-373

[18] Raviraj Joshi, Venkateshan Kannan, "Attention based end to end Speech Recognition for Voice Search in Hindi and English", ACM ISBN 978-1-4503, https://doi.org/10.1145/nnnnnnn.nnnnnnn, 2021