An Approach for Speech Recognition via Voice Activity Detection

Neha Verma[1]

[1]Department of Computer Science & Engineering, MIT Moradabad

---

Abstract - Research is being done on Automatic Speech Recognition (ASR), which can be used effectively in noisy environments. In terms of lustiness, the effectiveness of popular parameterization techniques was assessed in contrast to the background signal. A hybrid feature extractor is used for Mel frequency cepstral coefficients (MFCC), Perceptual linear predictive (PLP) coefficients, and their modified forms by combining the fundamental building blocks of PLP and MFCC. The VAD-based frame dropping formula was only applied to the ASR method's training phase. This method has the advantage of removing pauses and potentially significantly distorted speech segments, which aids in more precise phone modelling. The second portion focuses on the examination and contribution of the modified vocal activity detection technique.

Key Words: optics, photonics, light, lasers, templates, journals

1. INTRODUCTION

The practice of utilizing machines to change a human speaker's string of words is called automatic speech recognition (ASR). Because the aim of ASR is to have speech as a substandard form of interaction between a machine and a person, it is desirable that an ASR system be resilient to unpleasant fluctuation [5]. Endpoint identification in speech recognition systems that is brought on by non-speech events and background noise is often problematic [1]. Speech recognition systems that were trained in quiet environments often perform worse when ambient acoustic noise is present. Usually, dilapidation results from the difference between precise acoustic models and noisy speech data. There has been a lot of work done [2] to lessen this mismatch and restore recognition accuracy in noisy conditions. The topic of noise resilience in automatic speech recognition (ASR) can be approached in a variety of fundamentally distinct ways. One approach is to just subject the system to a specific kind of noise that it encounters during the recognition stage. This kind of system is called a "matched system," and it is probably superior to many other noise-compensation strategies—but only for that specific type of noise. The system needs to be retrained over an incredibly lengthy period of time and a vast library of new noise types in order to respond to these new types of noises. A more practical alternative to matched training is multi condition training, which trains the system on noisy speech heard at the loudest noise circumstances and removes the need to retrain the system every time the background noise changes [5].

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 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 Speech Presence Probability (SPP) is the foundation of the noise power spectrum estimate method. The sound power spectrum is here approximated by a crude estimate of the first 20 frames of the speech spectrum. Making speech recognition systems more resistant to in-the-moment reverberant circumstances is the aim of this work. A speech recognition system for the set of Hindi characters is provided using Mel Cepstrum Frequency Coefficients (MFCCs) and Perceptual Linear Prediction (PLP) for feature extraction. It comprises a strategy for enhancing the auditory spectrum as well as a short-time feature standardization method that lowers the variance of cepstral features across the training and test environments by modifying the balance and mean of cepstral features. A vocal signal is first preprocessed (pre-emphasized, typically using a second order high-pass filter). Short-time Fourier transform (STFT) analysis is completed in 40 ms using a predetermined time frame. Use the Hamming window to calculate the power spectrum of the signal. This approach evaluates 750 samples one at a time after educating the user on distinct noises. They will examine these samples and their training data to estimate the accuracy rate.

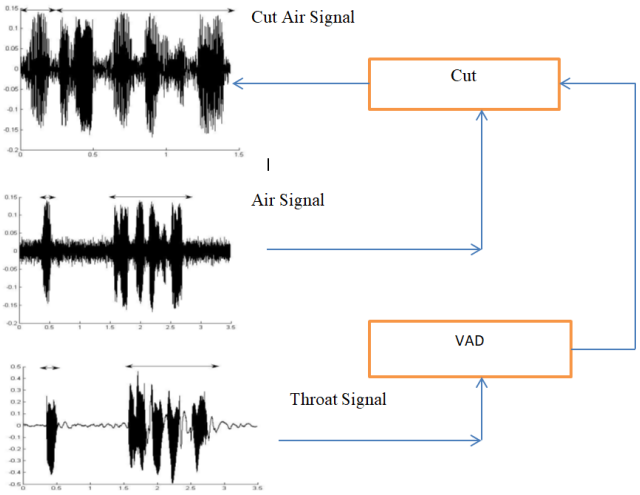

Figure-3.1: Investigational Arrangement

Three different noise types—fan, car, and diesel engine noise—were used in these experiments. These sounds have a static quality. The energy levels, lengths, and frequency contents of the trials are taken into consideration while VAD assesses the signal to determine whether the event is related to a talking user. The data from the neck microphone contains some low-frequency noise, but no high-frequency speech energy. Due to this, the energy ratios for the VAD were calculated using energy in the [250 5000] Hz frequency range. The background noise came from one of the speakers.

3.1 Algorithm

Step 1: First, use the silence indicator to pinpoint the times when the signal is not present. During these times, update the noise scales.

Step 2: To transform a time-domain signal into a frequency-domain signal, use the Short-time Fourier Transform (STFT). After the STFT is an operator for magnitude.

Step 3: Apply a high pass filter (HPF) to diminish processing errors brought on by noise variances. The HPF's objective is to reduce noise fluctuation.

Step 4: Correct any processing problems brought on by spectrum subtraction using a post-processor.

Step 5: In order to transform the treated signal into a time-domain signal, perform a Short-time Fourier Transform (ISTFT) on it in step five.

Step 6: Using a grouping method, the portion is classified as speech or non-speaking. This grouping rule decides whether a value is greater than a threshold. The output of the classifier is a continuous number, but it is threshold to produce a conclusion. The continuous output contains a lot of data about the signal that was lost owing to thresholding. It is almost certain that the signal is speech when the value is high, but it is less certain when the value is close to the threshold. The output of the classifier is an approximation of the chance that the input signal is speech.

Step 7: Make a set of features out of the signal that will be used to examine the traits that distinguish speech from non-speech.

Step 8: Use a classifier to determine the likelihood that the signal is speech by combining the evidence from the attributes.

4. RESULTS AND DISCUSSION

Robust feature extraction lowers disparity between training and testing phases using Voice Activity Detection.

In this part, the outcomes of the detail recognition are shown. Without applying a filter, Table 4.1 displays the word accuracy for MFCC-PLP and is regarded as the benchmark result. The average word accuracy for the baseline is found to be 55.2 on average over all noises in the SNR range of 15 to 0 dB. The word accuracy with filter is shown in Table 4.2. The MFCC-PLP with filter is found to have an average recognition rate of 65.6%. The suggested filter's performance against different kinds of noise is depicted in Fig. 4.3. Additionally, it has been discovered that the diesel engine and fan noises exhibit the largest gains in performance when compared to baseline performance. The average recognition accuracy significantly improves for all noises. Speech sounds between 15 dB and 0 dB have been found to exhibit the largest improvements in recognition accuracy.

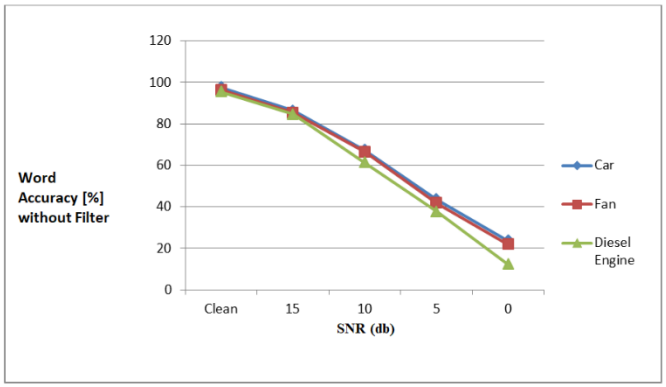

Figure-4.1: Word accuracy without filter

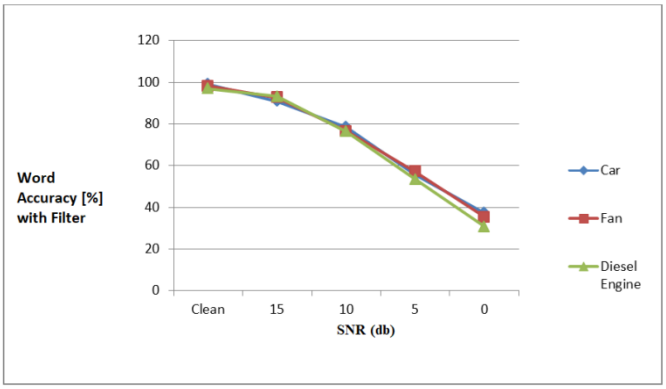

Figure-4.2: Word accuracy with filter

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

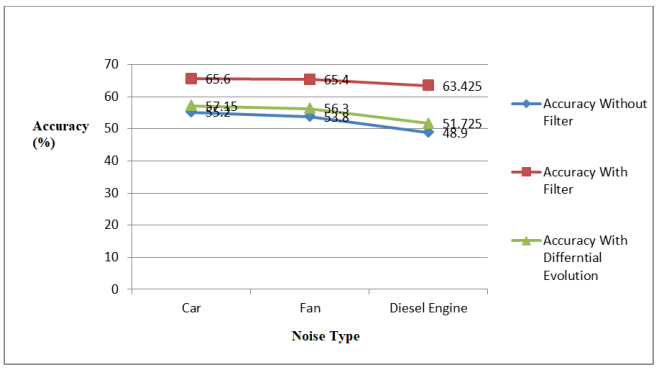

Figure-4.3: Word accuracy with and without filter and DE

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

5. CONCLUSIONS

When speech is influenced by a comparable environmental background, the non-speech components of a signal are affected more severely. Although entirely alternative noise suppression methods can improve the accuracy of the target ASR, this distortion is the root of many errors in the results of ASR systems. The risks of acoustic model standardization in the training portion result in an increase in WER because interrupted non-speech segments are frequently mistakenly identified as speech. Due to this feature, the VAD rule is used as a frame dropping technique to eliminate possibly

detrimental non-speech components from the processed signal. Depending on the situation, the VAD rule may even exclude a few frames that contained speech activity in addition to the non-speech components of the signal. The degree to which targets are selected will be significantly impacted by this serious defect. The proposed VAD rule, however, offers accuracy that is 16% better.

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