AN AUTOMATIC SPEECH RECOGNITION APPROACH USING MODIFIED VOICE ACTIVITY DETECTION MECHANISM

Neha Verma[1], Ajeet Singh[1], Atul Pratap Singh[2]

[1]Department of CSE, Moradabad Institute of Technology, Moradabad, UP, India
nehav0624@gmail.com
ajeetsingh252@gmail.com
[2]Department of CSE, Swami Vivekanand Subharti University Meerut, UP, India
atulpratap26385@gmail.com

ABSTRACT

The effectiveness of Automatic Speech Recognition (ASR), which can be employed in loud circumstances, is being researched. The effectiveness of common parameterization methods was evaluated in relation to lustiness and compared to the background signal. By merging the essential components of PLP and MFCC, a hybrid feature extractor is created for Mel frequency cepstral coefficients (MFCC), Perceptual linear predictive (PLP) coefficients, and their modified forms. Only the ASR method's training phase was used to apply the VAD-based frame dropping calculation. This technique has the benefit of eliminating pauses and speech segments that may be considerably distorted, It helps with more accurate phone modelling. The examination and contribution of the improved vocal activity detection approach are the main topics of the second section.

KEYWORDS: *Automatic Speech Recognition, Short-Time Fourier Transform, Perceptual Linear Prediction, Mel Cepstrum Frequency Coefficients, and Speech Presence Probability*

1. INTRODUCTION

Automatic speech recognition (ASR) is the process of using machines to modify a human speaker's string of words. It is preferable for an ASR system to be resilient to unpleasant fluctuation because the goal of ASR is to have speech as a suboptimal form of contact between a machine and a person [5]. It is frequently difficult for speech recognition systems to identify endpoints when background noise and non-speech events are present [1]. When ambient acoustic noise is present, speech recognition systems that were trained in quiet contexts sometimes perform worse. Dilapidation typically occurs as a result of the disparity between accurate acoustic models and noisy speech data. To reduce this mismatch and improve recognition accuracy in noisy environments, a lot of effort has been done [2]. It is possible to approach the subject of noise resilience in automated speech recognition (ASR) in a number of fundamentally different ways. One method is to just expose the machine to the type of noise it hears during the recognition step. Only for that particular form of noise, this type of system, known as a "matched system," is probably superior than many other noise-compensation techniques. To respond to these new forms of noises, the system must be retrained over an enormously long period of time and a huge library of new noise types. Multi condition training, which teaches the system on noisy speech heard under the loudest noise circumstances and eliminates the need to retrain the system every time the background noise changes [5], is a more practical alternative to matched training.

2. RELATED WORK

2.1 Review of existing VAD techniques and their limitations:

DOI: 10.5281/zenodo.10435452

Voice Activity Detection (VAD) techniques have been extensively studied and developed over the years. Here, we provide a review of some commonly used VAD techniques and discuss their limitations:

2.1.1 Energy-based VAD:

This technique measures the energy level of the audio signal and determines the presence of speech based on a predefined threshold.

Limitations:

It may fail in noisy environments where the background noise level is similar to or higher than the speech signal.

It may incorrectly classify non-speech sounds with high energy as speech.

2.1.2 Zero Crossing Rate (ZCR)-based VAD:

This technique analyses the number of times the audio signal crosses zero within a specific time frame.

Limitations:

It is sensitive to background noise, leading to false detections and inaccurate VAD decisions. It may not work well with non-stationary signals or when the speech signal is low in energy.

2.1.3 Statistical Model-based VAD:

This technique uses statistical models, such as Gaussian Mixture Models (GMMs) or Hidden Markov Models (HMMs), to distinguish speech from non-speech segments.

Limitations:

It requires training data for model adaptation, which can be time-consuming and resource-intensive. It may struggle with variations in speakers, languages, or acoustic conditions not covered by the training data.

2.2 Discussion of previous research on modified VAD approaches:

Several research studies have proposed modifications to traditional VAD techniques to address their limitations and improve the overall performance of ASR systems. Some notable approaches include:

2.2.1 Feature-based VAD:

This approach leverages additional features, such as Mel-frequency cepstral coefficients (MFCCs), spectral flatness, or pitch, in addition to traditional features, to enhance VAD accuracy.

Previous research has demonstrated improved performance by incorporating these additional features and using machine learning algorithms for classification.

2.2.2 Deep Learning-based VAD:

Deep learning techniques, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been applied to VAD, achieving significant improvements in accuracy.

These models can learn complex patterns and dependencies in speech signals, leading to robust VAD performance even in noisy environments.

2.2.3 Multi-modal VAD:

This approach combines audio signals with other modalities, such as visual information from video streams or linguistic context from text data, to enhance VAD accuracy.

Previous studies have shown that integrating multiple modalities can improve VAD performance, especially in challenging scenarios with overlapping speakers or low-quality audio.

2.2.4 Hybrid VAD:

Hybrid approaches combine multiple VAD techniques, such as energy-based, ZCR-based, and statistical model-based methods, to leverage their respective strengths and mitigate their limitations.

These approaches aim to achieve higher accuracy and robustness by fusing the outputs of multiple VAD algorithms.

These previous research efforts have demonstrated promising results in improving VAD performance. However, there is still room for further exploration and innovation to develop more accurate, robust, and efficient VAD mechanisms for ASR systems. In this paper, we propose a modified VAD approach that addresses the limitations of existing techniques and aims to enhance the performance of ASR systems.

The endpoint problem is discussed, and a timing strategy is offered by Qi Li et al. [1]. It uses a three-state transition diagram and an association best filter for endpoint detection. The proposed filter makes use of numerous criteria to ensure accuracy and strength. Xiaodong Cui et al. [2] intend to develop a noise-strong feature compensation (FC) method that supports polynomial regression of vocalisation signal-to-noise ratio (SNR) [2]. The bias between clean and noisy speech alternatives can be calculated using a set of polynomials using the expectation maximisation (EM) formula and the greatest probability (ML) criterion. Kapil Sharma and others. [3] suggest a comparison of different feature extraction techniques for isolated word end detection in noisy environments. At different SNR levels, we tested the cases of coloured noises, babbling noise, industrial plant noise, and distortions brought on by the recording medium. According to Tomas Dekens et al. [4], bone-conducted microphones cannot improve automatic speech identification in loud environments. When Voice Activity Detection (VAD) was applied using a throat mike signal as an input, it was shown that the recognition accuracy in non-stationary noise was greatly increased as compared to when VAD was applied using a standard mike signal. A comparison of three fundamentally independent noise-strong approaches is done by Sami Keronen et al. [5]. The performance of multi-condition training, Data-driven Parallel Model Combination (DPMC), and cluster-based missing information reconstruction techniques is evaluated in an exceptionally large vocabulary continuous speech recognition system. According to M. G. Sumithra et al. [6], a Kalman filter is used to boost the voice signal and reduce background noise. The enhanced signal has been included into the front portion of the recognition system to provide robust performance in shouting environment situations. A group of scholarly software programmes for signal and speech processing are shown by Lamia BOUAFIF et al. in their paper from 2007 (p. This MATLAB-created interface can be used for signal denoising, writing, and speech recognition. Md. Mahfuzur Rahman and others. We build a distributed noise speech recognizer using Cepstral Mean standardisation (CMS) for strong feature extraction, which is reliable enough for usage in real-world applications. The majority of the work goes towards controlling various loud environments. To do this, Mel-LP based speech analysis has been applied to speech coding on the linear frequency scale, replacing the unit delay with a first-order all-pass filter. In addition to describing the variety of issues in which optimisation formulations and algorithms play a part, Stephen J. Wright et al. [9] provide more information on specific application challenges in (machine translation) MT, speaker/language recognition, and automatic voice recognition. Namrata Dave and others, p. Speech samples are gathered from the recorded speech of male or female speakers, and they are compared to templates in the database. Speech will be parameterized using linear predictive codes (LPC), perceptual linear prediction (PLP), mel frequency cepstral coefficients (MFCC), PLP-RASTA (PLP-Relative Spectra), etc. Eric W. Healy and others [11] Despite much effort, researchers have been unable to develop monophonic (single-microphone) algorithms that can enhance speech comprehension in loud circumstances. Hard-of-hearing (HI) listeners require the no-hit construction of such an associate degree algorithmic rule due to their specific concern with hissing backgrounds. In the current study, binary masking supported by an algorithmic method was developed to separate speech from noise. The results of a study on the effects of frame shift and study window

DOI: 10.5281/zenodo.10435452

length on voice recognition rate are reported by Deividas Eringis et al. in [12]. For this purpose, the Mel Frequency Cepstral Analysis (MFCC), the Linear Prediction Cepstral Analysis (LPCC), and the Perceptual Linear Prediction Cepstral Analysis (PLPC) were each examined. In jingling shire scenarios, distant voice recognition is discussed by Jürgen T. Geiger et al. [13]. Using speech augmentation methods like abusing non-negative matrix resolution (NMF), ASR systems can be upgraded. A feature extraction technique that is resistant to unstable environments is recommended by Taejin Park et al. [14]. The anticipated theme is supported by the weighted bar graph of the time-frequency gradient in a very Mel picture image. Roger Hsiao and others, [15] The challenge's main characteristic is developing a superior system without having access to the appropriate training and development information. The analysis data are recorded using far-field microphones in noisy, bright surroundings, whereas the training data uses phone voice and near talking. Le Prell, Colleen G., and others [16] Speech communication typically happens when people are shouting, which can be a serious issue for military troops who must communicate in noisy environments. There are numerous consequences of noise on speech recognition, depending on the sources of the noise, the loudness and types of talkers, and the listener's hearing capacity. Asherf Nasef and others [17] Finding voice recognition software that can be utilised in noisy environments, such as offices, cars, aeroplanes, and other places, is still a difficult subject. Deep learning algorithms still struggle with an excessive identification loss when trying to identify speakers in noisy environments. Automated speech recognition (ASR) is recommended by Raviraj Joshi et al. [18] in the context of voice search feature on the Flipkart e-Commerce platform. Utilise the LAS (Listen-Attend-Deep-Spell) framework.

3. Proposed Work

The noise power spectrum estimate approach is built on the Speech Presence Probability (SPP). Here, a rough estimation of the first 20 frames of the speech spectrum is used to approximate the sound power spectrum. The goal of this effort is to improve speech recognition systems' resistance to in-the-moment reverberant situations. With the help of Mel Cepstrum Frequency Coefficients (MFCCs) and Perceptual Linear Prediction (PLP) for feature extraction, a speech recognition system for the set of Hindi letters is offered. It includes a method for improving the auditory spectrum and a short-time feature standardisation technique that reduces the variance of cepstral features across the training and test environments by altering the balance and mean of cepstral features. The initial step in preprocessing a vocal signal (pre-emphasized, typically utilising a high-pass filter of second order). Using a predefined time frame, short-time Fourier transform (STFT) analysis is finished in 40 ms. Calculate the signal's power spectrum using the Hamming window. This method educates the user about different noises before evaluating each of the 750 samples one by one. To determine the accuracy rate, they will look at these samples and their training data.

These trials included three different types of noise: fan, automobile, and diesel engine sounds. There is a static aspect to these noises. When evaluating the signal to identify whether the event is associated to a talking user, VAD takes into account the energy levels, durations, and frequency contents of the trials. There is some low-frequency noise in the data from the neck microphone, but no high-frequency speech energy. Because of this, energy in the [250 5000] Hz frequency range was used to calculate the energy ratios for the VAD. One of the speakers was the source of the ambient noise.

3.1.1 Method

First, identify the moments when the signal is not present by using the quiet indicator. Update the noise scales during these times.

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

Applying a high pass filter (HPF) in step three will reduce processing errors brought on by variations in the noise. The reduction of noise fluctuation is the HPF's main goal.

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

DOI: 10.5281/zenodo.10435452

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

Step 6: The section is categorised as talking or non-speaking using a grouping algorithm. If a value exceeds a threshold, this grouping rule determines it. The classifier's output is a continuous number that can be used as a threshold to draw a judgement. Numerous details about the signal that was lost due to thresholding are present in the continuous output. When the value is large, 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 will be used to analyse the characteristics that differentiate speech apart from other sounds.

Step 8: By combining the evidence from the attributes, use a classifier to assess the chance that the signal is speech.

Using Voice Activity Detection, robust feature extraction reduces the gap between the training and testing stages.

4. RESULT AND DISCUSSION

The results of the detail recognition are displayed in this section. Table 4.1, which shows the word accuracy for MFCC-PLP without applying a filter, is taken as the standard result. Over all noises in the SNR range of 15 to 0 dB, an average word accuracy for the baseline is found to be 55.2. . Table 4.2 displays word accuracy with filter. It is discovered that the MFCC-PLP with filter has an average recognition rate of 65.6%. Fig. 4.3 shows how well the suggested filter performs in the presence of various types of noise. Additionally, it has been shown that when compared to baseline performance, the diesel engine and fan noises show the biggest improvements. For all noises, the average recognition accuracy dramatically increases. The highest improvements in recognition accuracy have been reported to occur for speech sounds between 15 dB and 0 db.

5. CONCLUSION

The non-speech components of a signal are more severely impacted when speech is influenced by a similar environmental background. Although completely different noise suppression techniques can boost the target ASR's accuracy, this distortion is the cause of a lot of inaccuracies in the ASR system's output. Because interrupted non-speech segments are commonly wrongly classified as speech, the hazards of acoustic model standardisation in the training stage lead to an increase in WER. The VAD rule is employed as a frame dropping technique because of this feature to remove potentially harmful non-speech components from the processed signal. Depending on the circumstances, the VAD rule might even disqualify a few of frames that included speech activity in addition to the signal's non-speech components. This critical flaw will have a substantial impact on the selection of targets. However, the precision provided by the proposed VAD rule 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

DOI: 10.5281/zenodo.10435452

[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

AUTHORS

Neha Verma
She is working as an Assistant Professor in Department of Computer Science and Engineering at Moradabad Institute of technology, Moradabad, UP, India.. SHe did his B.tech and M.tech in Computer Science. His Research area is Software Engineering and AI.

Ajeet Singh
I Ajeet Singh, working as Assistant Professor in Department of Computer Science and Engineering at Moradabad Institute of technology , Moradabad, UP, India. He is pursuing his Phd from Faculty of Engineering and Technology at Rama University Uttar Pradesh, Kanpur (India) in Computer Science and Engineering. He did his B.tech and M.tech in Computer Science. His Research area is AI Algorithms and Deep Learning.

DOI: 10.5281/zenodo.10435452

Atul Pratap Singh

I am Atul Pratap Singh, a passionate and enthusiastic Ph.D. research student, dedicated to exploring the frontiers of knowledge in Deep Learning and Medical Imaging. I have always been curious about the complexities of the artificial intelligence world and driven by a desire to unravel its mysteries through scientific inquiry. After completing my undergraduate studies in 2015, I knew that pursuing a Ph.D. was my true calling. I was accepted into the Kalinga University where I am currently immersed in the pursuit of my doctoral degree in Computer Science and Engineering. This exciting and challenging journey has allowed me to delve deep into Medical Imaging and Deep Learning.

DOI: 10.5281/zenodo.10435452