World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 7.029

SPEECH BACKGROUND VOICE SUPPRESSION WITH DEEP LEARNING

Krishna Swaroop A, Sindhu K S, *Prachala B C, Deepti C R, Dharanesh C, Ananya M R.

Department of Information Science & Engineering, Malnad College of Engineering, Hassan.

Article Received on 11/03/2024 **Article Revised on 01/04/2024** **Article Accepted on 21/04/2024**

***Corresponding Author**
Prachala B. C.
Department of Information Science & Engineering, Malnad College of Engineering, Hassan.

ABSTRACT

This study presents a complete examination of noise addition and evacuation processes in signal handling, significant for understanding and further developing sound sign loyalty in certifiable conditions. Through point-by- point block graphs and numerical details, the components of commotion expansion, including breath noise and Gaussian noise, are clarified, recreating complex hear-able situations. Moreover, a refined noise expulsion calculation is illustrated, including spectrogram analysis, Wiener filtering, voice action discovery, spectral gating, and non- local means denoising. These procedures on the whole mean to reestablish sound signs to their perfect state, improving sign clearness and devotion in the midst of unavoidable commotion contamination.

Index Terms: Breath noise, Gaussian noise Spectrogram analysis, Spectral subtraction, Wiener filtering, Voice activity detection, Spectral gating, non-local means denoising, Signal clarity.

1. INTRODUCTION

In the domain of sound handling, the test of upgrading signal clearness by smothering foundation discourse voices has prodded creative methodologies, utilizing headways in profound learning and sign handling. This paper presents a clever philosophy that joins spectral subtraction, Wiener filtering inside a succession to-grouping model, and spectral gating innovation. The goal is to relieve undesirable foundation discourse, in this way working on the general understandability of essential sound signs. As correspondence advancements keep on developing, the interest for compelling sound decrease in different sound conditions has become progressively articulated. Our methodology tends to this basic as well as coordinates flawlessly into commonsense applications through a deep learning-based model. This presentation gives a passage into the investigation of our proposed method, featuring its likely effect on real time communications, sound recording, and continuous correspondence frameworks. Through a union of spectral gating and deep learning, our system expects to add to the continuous talk on imaginative answers for foundation discourse voice concealment. The principal objective of discourse improvement is to limit the impacts of clamor on discourse by working on the perceptual nature of loud discourse. In a genuine setting, most frequently the discourse signal is joined by foundation commotion. The presence of more undesirable foundation clamor like vehicle commotion, and train commotion influences the nature of the discourse signal. A few discourse improvement techniques are proposed to upgrade the nature of the debased discourse. By and large, the discourse upgrade issue comprises of a lot of issues described by the sort of commotion source, so that clamor obstructs the perfect sign, the quantity of receiver yields and the quantity of voice channels accessible for improvement.[8] The Wiener filter is one of the significant time area techniques utilized for discourse improvement. This is utilized for upgrading discourse corrupted by added substance fixed foundation clamor. Weiner filter is one of the elective strategies to spectral subtraction improve the debased discourse signal. In this paper, we zeroed in on the spectral subtraction commotion evacuation approach in discourse handling alongside the Weiner separating approach. Spectral subtraction is a famous technique that is utilized to upgrade the designated discourse within the sight of foundation commotion.

2. Literature survey

The study investigates using diffusion-based generative models for improving speech quality and reducing reverberation. They focus on optimizing network architecture and demonstrate that their method competes well with other models across various datasets and real-world tests. Furthermore, it proves effective in listening experiments and is applicable for both noise removal and dereverberation.[1] This paper presents RCEN, a technique for diminishing foundation commotion in seismic records got utilizing conveyed optical fiber acoustic detecting (DAS), especially for vertical seismic profiles (VSP). RCEN combines deep iterative memory blocks (DMB) and channel aggregation blocks (CAB) to enhance noise reduction and signal preservation. It's trained using both synthetic and real DAS noise data. The results show that RCEN outperforms other methods when processing both synthetic and field DAS-VSP data.[3] The article uses supervised learning, deep learning, and different methods like speech enhancement and speaker separation to improve speech separation. It also considers multi-microphone techniques for better results in various environments.[4]

3. Proposed system

The proposed system is an advanced solution for background speech voice suppression in audio signals, combining spectral subtraction, Wiener filtering within a sequence-to-sequence model, and spectral gating technology. Operating in the frequency domain using spectrogram representations, the system dynamically selects between spectral subtraction and Wiener filtering based on an estimated noise spectrum, enhancing adaptability to various noise characteristics. The processing pipeline involves calculating spectrograms for both the noisy and enhanced signals, with optional integration of deep learning models for further refinement. Spectral gating technology is introduced to provide real-time adjustments, ensuring optimal background speech suppression in response to varying noise levels. Key contributions include adaptability through dynamic noise reduction methods, the introduction of spectral gating for real-time adjustments, and optional deep learning integration for enhanced robustness. The desired result is an enhanced and denoised speech signal, showcasing the system's efficacy in mitigating unwanted background voices.

3.1 Wiener filter

The primary point of the Wiener Filter is to gauge genuinely an obscure sign involving a sign as an info and sifting that referred to motion toward produce the gauge as a result. The Wiener Filter is utilized to eliminate the clamor from the adulterated sign to give a gauge of the hidden sign of interest. The Wiener Filter depends on a factual methodology, and an expanded accentuation on measurable examination inside the hypothesis is given in the base mean square blunder (MMSE) assessor.

Normal deterministic filter is intended for a needed recurrence reaction. Notwithstanding, the plan of the Wiener Filter adopts an alternate strategy. One is considered to know about the spectral properties of the first sign and the clamor, and another looks for the direct time-invariant channel whose result would come as near the first sign as could be expected. Both signal and noise are fixed direct stochastic cycles with known phantom qualities. The Wiener filter should be actually feasible/causal. Least mean-square mistake (MMSE).

3.2 Spectral subtraction

Spectral subtraction is a frequency domain method widely applied in audio signal processing to reduce background noise in speech signals. Operating on the spectrogram representation of the audio signal, it estimates and subtracts the spectral profile of background noise, enhancing the clarity of speech components.

4. Implementation

The block diagram provided depicts the process of noise addition, which is a fundamental concept in signal processing. The beginning stage is a 'Clean Sign,' which is an uncorrupted and unadulterated type of the sign that contains no obstruction or twisting. This could address, for instance, an unblemished sound recording got in a controlled climate where no outside sounds are available. The process of noise addition is then illustrated, where external noise is deliberately combined with this clean signal. This process often simulates real-world conditions, where signals rarely exist in isolation and are frequently contaminated by extraneous sounds and disturbances.

This particular chart adds two unmistakable sorts of commotion to the spotless sign. The first is 'Breath noise,' normal in vocal accounts or wind instruments, starting from the entertainer's relaxing. The second is 'White or Gaussian noise,' a kind of commotion with a steady power otherworldly thickness. It’s named 'white' by similarity to white light, which contains every noticeable recurrence, and 'Gaussian' on the grounds that its power dispersion follows a Gaussian, or normal, dissemination. This commotion expansion process brings about the 'noisysignal,' a blend of the first sign with the breath and repetitive sounds. This noisy signal serves as a more realistic representation of signals found in everyday environments, and studying it can be critical for developing effective noise reduction and signal clarification technologies. This process often simulates real-world conditions, where signals rarely exist in isolation and are frequently contaminated by extraneous sounds and disturbances.

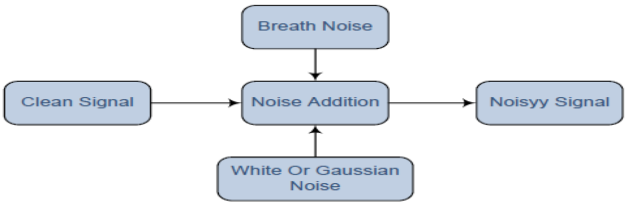

Fig. 1.1: Block diagram of noise addition.

In signal processing, the exemplar or 'Clean Signal,' denoted as $S_{clean}(t)$, represents an idealized waveform devoid of distortive elements and external perturbations. This archetype signal can be equated to a high-fidelity audio capture in a serene setting, devoid of ambient interference, often represented as:

Where A is the $S_{clean}(t)=A\sin(2\pi ft+\phi)$ amplitude, f is the frequency, represents time, and ϕ is the phase of the pure tone.

Noise addition is a deliberate process wherein extraneous acoustical patterns are superposed onto to $S_{clean}(t)$ simulate the $S_{noisy}(t)=S_{clean}(t)+N_{b}(t)+N_{w}(t)$ intricacies of real-world auditory scenarios. This conflation is mathematically modelled as follows:

Where $S_{noisy}(t)$ symbolizes the composite noisy signal, $N_{b}(t)$ signifies the 'Breath Noise' component, and $N_{w}(t)$ indicates the 'White or Gaussian Noise'. The breath noise, predominant in vocal or wind instrument productions, is not uniform and can be depicted as a sporadic signal with a variable amplitude $A_{b}(t)$: $N_{b}(t)=A_{b}(t)\sin(2\pi f_{b}t+\phi_{b})$

On the other hand, 'White or Gaussian Noise,' $N_{w}(t)$, is characterized by a mean of zero and a constant power spectral density across all frequencies, akin to white light comprising all visual frequencies. The power spectral density $P_{w}(f)$ for white noise is steady, and for Gaussian noise, it is distributed normally:

$$P_{w}(f)=\frac{1}{2\pi\sigma^{2}}e^{-\frac{f^{2}}{2\sigma^{2}}}$$

Where σ^2 denotes the variance of the noise.

Consequently, $S_{noisy}(t)$ it embodies a more authentic aural signal that emulates environmental auditory experiences, enabling analysts and engineers to devise and refine noise mitigation strategies, thereby enhancing signal processing techniques for real-world application. The modelling and study of such signals are crucial in developing algorithms that can discern between signal and noise, ultimately facilitating more transparent communication and accurate information retrieval in noisy environments.

The block diagram you've provided outlines the process of Noise addition is a deliberate process wherein extraneous acoustical patterns are superposed onto$S_{clean}(t)$ to simulate the intricacies the complex procedure into individual steps that aim to filter out various types of noise and restore the signal to its clean state. This process is crucial in multiple applications, such as audio engineering, telecommunications, and signal processing for voice recognition systems.

The process begins with an 'Input Noise Signal,' the raw audio data that includes unwanted noise. The first step in the noise removal process is to 'Calculate Spectrogram,' which involves converting the time-domain signal into the frequency domain using a Fourier transform. The spectrogram visually illustrates how the frequency of spectrum of signal changes with time, which enables the identification and isolating noisecomponents. Next, the system 'Finds Frequency Range and Noise Type' by analysing the spectrogram. This step is critical as it allows the noise removal system to differentiate between the desired signal and noise.

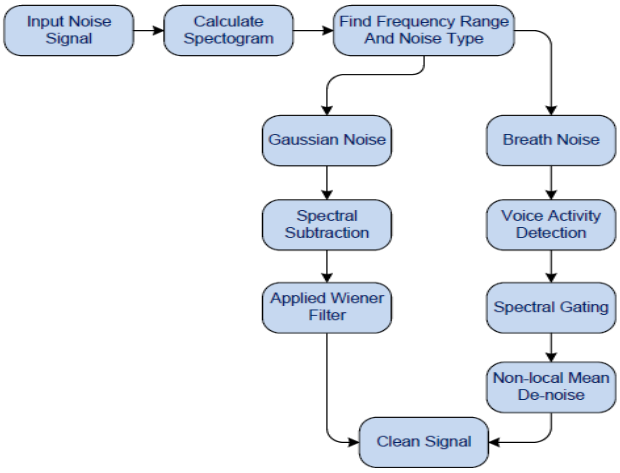

Fig. 1.2: Block diagram of Noise Removal.

The process then bifurcates to deal with different types of noise identified—'Gaussian Noise' and 'Breath Noise.' Gaussian noise, typically characterized by a normal distribution in the frequency domain, is addressed through 'Spectral Subtraction.' This technique estimates and subtracts the noise spectrum from the noisy signal's spectrum. Subsequently, an 'Applied Wiener Filter, ' an adaptive filter that minimizes the mean square errorbetween the estimated clean signal and the original signal, is used to refine the signal further.

'Breath Noise' is handled differently on the other branch of the process. 'Voice Activity

Detection' is applied to identify the presence of $S_{clean}(f,\tau) = S_{noisy}(f,\tau) - \alpha \cdot N(f,\tau)$ speech; this is particularly useful to distinguish breath noise from the actual vocal content. Following this, 'Spectral Gating,' a process that suppresses the frequencies where noise is dominant and keeps the frequencies associated with the speech signal, is employed. In some cases, an advanced technique called 'Non-local Mean De-noise' is used. It removes noise by considering the entire signal to find repeating patterns and average them, thus preserving the signal's details while reducing noise.

The final step merges the outputs of the Wiener filter and the spectral gating, ensuring that all types of noise are appropriately addressed. The result is a 'Clean Signal,' which has been processed to remove as much noise as possible while retaining the integrity of the original signal. This clean signal is now more suitable for further processing or listening, free from the distortions and interruptions that noise can cause. The noise removal process is a testament to the sophisticated methods developed to ensure clear and accurate audio signal transmission in the face of ubiquitous noise.

The block diagram delineates a sophisticated noise removal algorithm to purify an audio signal by meticulously separating and eliminating unwanted noise components. This process begins with an input noise signal, denoted by $S_{noisy}(t)$, where represents the time domain. The initial step is to transform this signal into a frequency-time representation through a spectrogram calculation. The transformation is typically achieved using the Short-Time Fourier Transform (STFT), represented as:

$$STFT\{S_{noisy}(t)\} = S_{noisy}(f,\tau)$$

Where f is the frequency and is the time-frame index in the spectrogram.

Following the spectrogram analysis, the algorithm proceeds to identify the frequency range and typology of the noise. The approach employs spectral subtraction for Gaussian noise, which is prevalent in many real-world signals. The underlying principle of spectral subtraction is formulated as:

Where $S_{clean}(f,\tau)$ represents the estimated clean signal in the spectrogram, $N(f,\tau)$ is the noise estimate, and is the over- αsubtraction factor.

After this subtraction, the Wiener, a statistical filter, is applied. The Wiener filter aims to minimize the overall mean square error in the signal and is given by:

$$\widehat{S}(f,\tau) = \frac{P_S(f)}{P_S(f) + P_N(f)} \cdot S_{noisy}(f,\tau)$$

Where $\widehat{S}(f,\tau)$ is the filtered signal, $P_S(f)$ is the signal's powerspectral density, and $P_N(f)$ is the power spectral density of the noise. In parallel, breath noise is treated through voice activity detection (VAD), which functions as a binary classifier $VAD(\tau)$ to discriminate between noise and speech. Post- detection, spectral gating is applied, which involves a thresholding process that suppresses noise while preserving speech frequencies. This process can be illustrated as:

\[S'_{clean}(f, \tau) = \begin{cases} S_{noisy}(f, \tau) &
\text{if } S_{noisy}(f, \tau) > \Theta(f) \\0 &
\text{otherwise}\end{cases}\]

where $\Theta(f)$ is the spectral gate threshold.

Finally, a non-local means denoising technique may be incorporated to further enhance the signal by exploiting the self- similarity across the entire signal. The output from the Wiener filter and spectral gating are combined to reconstruct the clean signal, denoted by $S_{final}(t)$, offering an optimizednoise-reduced version of the original audio input.

The intricate nature of this noise removal process and its reliance on multiple advanced signal processing techniques illustrate the depth of modern digital signal processing methods and their capacity to yield high-fidelity audio signals from corrupted inputs.

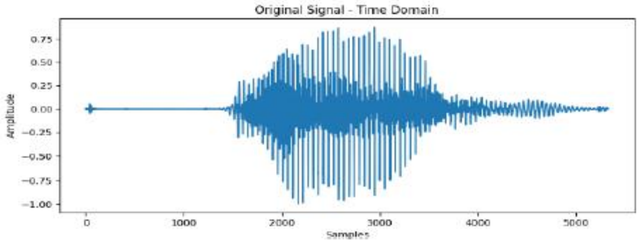

Fig. 2.1: Original signal (Clean signal) – Time domain plot.

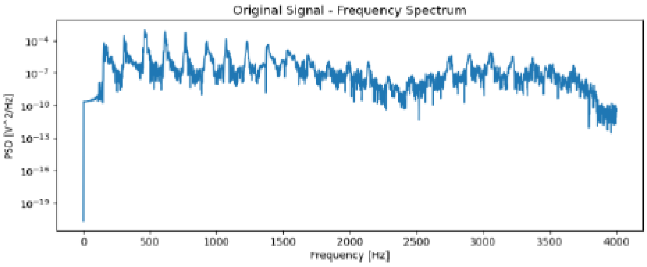

Fig. 2.2: Original signal (Clean signal) – Frequency domain (Spectrum) plot.

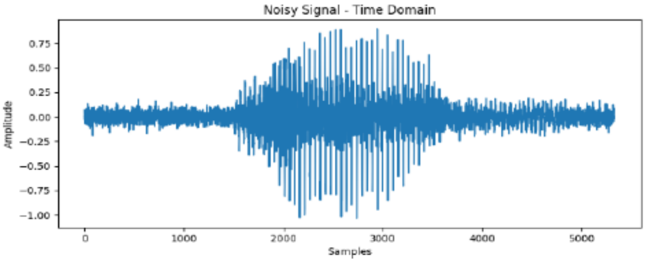

Fig. 2.3: Noisy Signal – Time domain plot.

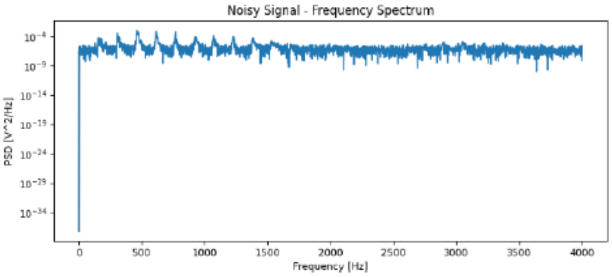

Fig. 2.4: Noisy Signal – Frequency domain (Spectrum) plot.

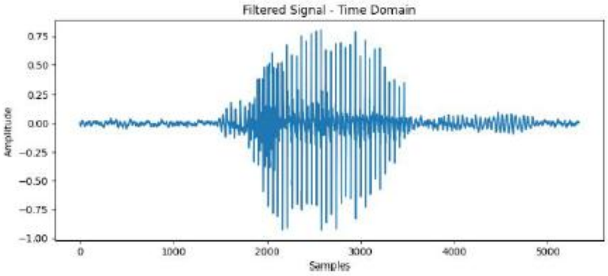

Fig. 2.5: Filtered Signal – Time domain plot.

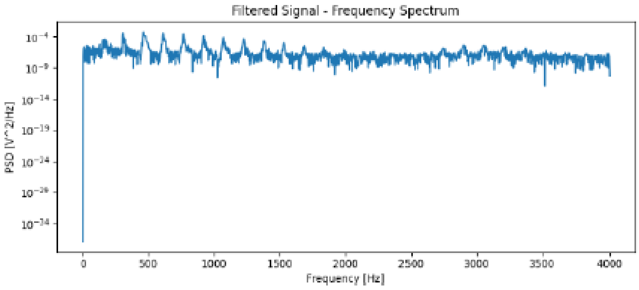

Fig. 2.6: Filtered Signal – Frequency domain (Spectrum) plot.

The provided figures and performance evaluation data detail an experiment in signal processing. Figures 2.1 and 2.2 display the original clean signal in both time and frequency domains, establishing a baseline for comparison. Figures 2.3 and 2.4 present the same signal disrupted by noise, showing alterations in the signal's clarity and frequency composition. Figures 2.5 and 2.6 depict the signal after a separating cycle has been applied, illustrating how much of the noise has been removed and how closely the filtered signal matches the original clean signal's characteristics. values offer insight into the levels of noise before and after filtering. An SNR of 12.50 dB from clean to noisy indicates thelevel of commotion expansion, whereas the SNR of 8.71 dB from clean to filtered reveals the effectiveness of the noise removal process. The lower SNR in the latter suggests some residual noise or signal distortion remains.

5. CONCLUSION

In outline, this examination digs into the intricacies of noise expansion and expulsion in signal handling, vital for further developing sound sign devotion. By utilizing methods like spectrogram analysis, spectral subtraction, and progressed denoising calculations, the review plans to alleviate the effect of commotion, upgrading the clearness and respectability of audio signal in true conditions. This work adds to the improvement of sound decrease advances, at last working withmore clear correspondence and more precise data recovery in noisy settings.

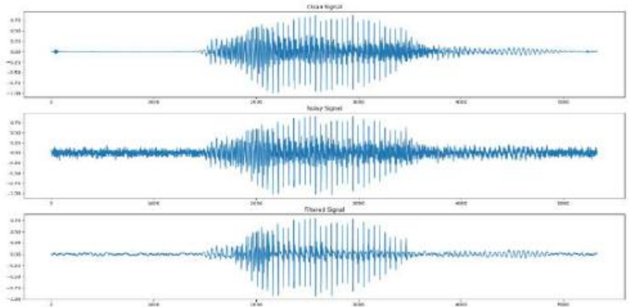

Fig. 2.7: Comparison of 3 signals (Clean, Noisy and Filtered) – Time Domain Plot.

Figure 2.7 provides a direct comparative visualization of the three states of the signal, allowing for an assessment of the noise removal's impact in the time domain, likely showing theextent to which the original signal's shape is restored. The consistency in the fundamental frequency across the figures indicates that the filter effectively retains the signal's original pitch, which is a critical component of perceived audio quality.

Performance evaluation

Fundamental Frequency of Clean Signal: 320.27 Hz Fundamental Frequency of Noisy Signal: 320.27 Hz Fundamental Frequency of Filtered Signal: 320.27 Hz
SNR (Clean to Noisy): 12.50 dBSNR (Clean to Filtered): 8.71 dB

The functionality of the noise removal process is quantitatively assessed by the fundamental frequency, which stays steady across all three conditions (clean, noisy, filtered) at 320.27 Hz, suggesting that the filtering process preserves the signal's primary characteristics. The Signal-to-Noise Ratio (SNR).

REFERENCES

1. Towards efficient models for real-time deep noise suppression Sebastian Braun, Hannes Gamper, Chandan K.A. Reddy, Ivan Tashev.
2. Supervised speech separation based on deep learning, Dealing Wang, fellow, ieee, and Jing Chen.
3. DNN-Supported speech enhancement with a cepstral estimation of both excitation and envelope Samy Elshamy, Nilesh Madhu, Wouter and Tim Fingscheidt, senior member, ieee.
4. Noise suppression with similarity-based self-supervised deep learning Chuang Niu, member, ieee, mengzhou li, member, ieee, Fenglei Fan, member, ieee, Weiwen Wu, Xiaodong Guo, Qing Lyu, member, ieee, and Ge Wang, fellow, ieee.
5. RECN: a deep-learning-based background noise suppression method for das-vsp records tie Zhong, Ming Cheng, graduate student member, ieee, Shaoping Lu, Xintong Dong, and Yue Li, member, ieee.
6. Speech enhancement by lstm-based noise suppression followed by CNN-based speech restoration Maximilian Strake1, Bruno Defraene2, Kristoff Fluyt2, Wouter Tirry2 and Tim Fingscheidt1.
7. S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Transactions on Acoustics, Speech, Signal Processing.