OKU Fen Bilimleri Enstitüsü Dergisi
8(1): 222-247, 2025

OKU Journal of The Institute of Science and Technology, 8(1): 222-247, 2025

Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi

Osmaniye Korkut Ata University Journal of The Institute of Science and Technology

Investigation of the Effectiveness of Audio Processing and Filtering Strategies in Noisy Environments on Speech Recognition Performance

Cem ÖZKURT[1*]

[1]Sakarya University of Applied Science, Faculty of Technology, Department of Computer Engineering, Artificial Intelligence and Data Science Research and Application Center, Sakarya

[1]https://orcid.org/0000-0002-1251-7715
*Corresponding author: cemozkurt@subu.edu.tr

Research Article

Article History:
Received: 22.03.2024
Accepted: 29.08.2024
Published online: 15.01.2025

Keywords:
CNN
Audio processing
Speech recognition
Short-Time Fourier Transform (STFT)
Noise reduction

ABSTRACT

This study investigates the effects of audio processing and filtering strategies to enhance the performance of speech recognition systems in noisy environments. The focus is on the Short-Time Fourier Transform (STFT) operations applied to noisy audio files and noise reduction procedures. While STFT operations form the basis for detecting noise and analyzing the speech signal in the frequency domain, noise reduction steps involve threshold-based masking and convolution operations. The results indicate a potential improvement in speech recognition accuracy in noisy environments through these audio processing and filtering strategies.Although the findings suggest a positive impact, the study is based on a single audio file, and further research with larger datasets is necessary to substantiate these claims. A detailed analysis of the graphs provides guidance for evaluating the effectiveness of noise reduction procedures and serves as a roadmap for future research. This study emphasizes the critical importance of audio processing and filtering strategies in improving the performance of speech recognition systems in noisy environments, laying a foundation for future studies.

Gürültülü Ortamlarda Ses Tanıma Performansı Üzerinde Ses İşleme ve Filtreleme Stratejilerinin Etkinliğinin Araştırılması

Araştırma Makalesi

Makale Tarihçesi:
Geliş tarihi: 22.03.2024
Kabul tarihi:29.08.2024
Online Yayınlanma: 15.01.2025

Anahtar Kelimeler:
CNN
Ses işleme
Konuşma tanıma
Kısa Süreli Fourier Dönüşümü
Gürültü azaltma

ÖZ

Bu çalışma, gürültülü ortamlarda konuşma tanıma sistemlerinin performansını artırmak için ses işleme ve filtreleme stratejilerinin etkilerini araştırmaktadır. Odak noktası, gürültülü ses dosyalarına uygulanan Kısa Süreli Fourier Dönüşümü (STFT) işlemleri ve gürültü azaltma prosedürleridir. STFT işlemleri, gürültüyü tespit etme ve konuşma sinyalini frekans alanında analiz etme temelini oluştururken, gürültü azaltma adımları eşik tabanlı maskeleme ve konvolüsyon işlemlerini içermektedir. Sonuçlar, bu ses işleme ve filtreleme stratejileri aracılığıyla gürültülü ortamlarda konuşma tanıma doğruluğunda potansiyel bir iyileştirme olduğunu göstermektedir. Bulgular olumlu bir etkiyi önerse de, çalışma tek bir ses dosyasına dayanmaktadır ve bu iddiaları doğrulamak için daha büyük veri kümeleriyle daha fazla araştırma yapılması gerekmektedir. Grafiklerin detaylı analizi, gürültü azaltma prosedürlerinin etkinliğini değerlendirmek için rehberlik sağlar ve gelecek araştırmalar için bir yol haritası görevi görür. Bu çalışma, gürültülü ortamlarda konuşma tanıma sistemlerinin performansını artırmada ses işleme ve filtreleme stratejilerinin kritik önemini vurgulayarak, gelecek çalışmalar için bir temel oluşturur.

To Cite: Özkurt C. Investigation of the Effectiveness of Audio Processing and Filtering Strategies in Noisy Environments on Speech Recognition Performance. Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi 2025; 8(1): 222-247.

1. Introduction

Speech recognition in noisy environments has become a significant focus of research on audio processing techniques and filtering strategies, considering the increasing demands and applications (Malik, 2021). In recent years, there has been a notable shift towards end-to-end (E2E) models from traditional hybrid approaches to enhance speech recognition performance in noisy environments. This transition necessitates a comprehensive comparison of various E2E methods in terms of accuracy and reliability (Li, 2020). Speech recognition in noisy environments faces significant challenges due to background noise, environmental factors, and acoustic interference from other speakers. This situation can decrease the accuracy of speech recognition systems and lead to reliability issues in various applications. Particularly, the development of noise-resistant speech recognition systems in areas such as smart devices, voice command systems, and teleconferencing applications can significantly impact user experience (Wang et al., 2019; Martinek et al., 2020).

The key research questions of this study are as follows:

1. Which audio processing and filtering strategies provide the most effective solutions for enhancing speech recognition performance in noisy environments?
2. How does the application of Short-Time Fourier Transform (STFT) to noisy audio files affect speech recognition accuracy by analyzing the speech signal in the frequency domain?
3. Can noise reduction procedures enhance the reliability of speech signals by masking and convolution operations in specific frequency ranges?

Based on these research questions, our hypotheses are:

- STFT-based audio processing strategies will significantly improve speech recognition performance in noisy environments.
- Noise reduction procedures will increase the accuracy of speech recognition systems by reducing the impact of noise in specific frequency ranges.

The audio processing and filtering strategies addressed in this study include fundamental techniques such as noise reduction, frequency filtering, and spectrogram analysis. Noise reduction focuses on effectively filtering out background noise to highlight the speaker's voice, which can enhance the accuracy of speech recognition systems (Garg and Jain, 2016). Frequency filtering aims to suppress noise in specific frequency ranges but carries the risk of losing significant speech components (Nuha & Absa, 2022). Spectrogram analysis visualizes the frequency content of the audio in detail, aiding in understanding speech characteristics (Xing et al., 2015).

This study does not rely on a conventional dataset but instead focuses on analyzing a single audio file containing specific noise conditions. This approach allows for a detailed examination of noise reduction strategies in a controlled environment.

The overarching goal of this research is to evaluate the audio processing and filtering strategies used to improve speech recognition performance in noisy environments and to understand their advantages and limitations. The findings aim to guide the design of more effective and reliable speech recognition systems in the future.

2. Related Works

A novel multi-channel source activity detector utilizing spatial localization of the target speech source is introduced (Rosca, 2002). This detector is compared with a two-channel Voice Activity Detector (VAD) employing AMR speech detection algorithms on real data recorded in a noisy car environment. The significance of VAD in speech processing, including speech enhancement and speech coding, especially in noisy environments, is emphasized. A VAD evaluation framework for such environments is developed (Kitaoka, 2007), named Combined Environmental and Text for Noisy Speech Recognition (CENSREC-1-C). A speech recognition system that identifies basic voice commands for a mobile robot operating in a home environment is described (Sasaki, 2008). The system's performance is evaluated using four indices experimentally under various conditions, confirming its efficiency in noisy environments or with distant sound sources.

A speech detection method for an anthropomorphic robot that separates and recognizes speech signals originating from the front in noisy home environments is presented (Kim, 2008). The system operates in real-time without requiring pre-trained filter coefficients, even in noisy environments. Speech recognition performance in noisy car environments is demonstrated to be improved by combining microphone array processing techniques with a visual-audio Voice Activity Detector (VAD) (Faubel, 2011). The proposed localization framework combined with delay and beamforming yields a 7.1. A learning model using acoustic models to increase speech recognition rates is created (Oh, 2014). In speech processing applications, noise processing for speech recognition systems is often expressed as a digital filtering process where noisy speech is passed through a linear filter to obtain clean speech predictions. Focus on noise estimation, removal, and speech enhancement techniques is emphasized (Garg, 2016). The degradation problem in speech recognition performance arises from differences between the training model and the recognition environment.

The spectrum feature of noise signals in the silence-feature normalization model is utilized to improve its performance and enhance silence-feature normalization in a low SNR signal by determining a reference value for speech and non-speech classification (Oh, 2018). A new approach for speech perception and voice activity detection tasks is proposed (Gutierrez, 2019), indicating that the same Voltage Controlled Oscillator (VCO) can be reused to implement band-pass filters and standard ADC output in decision mode when a keyword or sound is detected. Recent studies have expanded on the use of adaptive filtering techniques for VAD, including the application of deep learning models that dynamically adjust to varying noise conditions, demonstrating enhanced robustness in challenging environments (Nguyen, 2023; Zhou, 2023).

Contributions to the processing of speech signals and the development of perceptual encoders are made (Schroeder, 1999). By utilizing the characteristics of the human ear, perceptual encoders capable of transmitting speech and high-quality music at low bit rates are developed. The performance of Amazigh speech recognition with interactive voice response in noisy conditions is described (Hamidi, 2020). Experiments were conducted first for uncoded speech and then repeated for coded speech in noisy environments with different signal-to-noise ratios (SNRs). The need to increase speech data durations to obtain larger datasets and combine them with various noises encountered in the environment is analyzed (Phyu, 2020). An innovative approach to voice control for operational and technical functions in a real Smart Home (SH) environment is outlined (Martinek, 2020). In proposed experiments, success rates for speech command recognition were compared for different types of interventions added to a real SH environment (television, vacuum cleaner, washing machine, dishwasher, and fan). Filter designs assist in increasing accuracy through parameter adjustment in speech recognition systems.

An analysis in a complex nonlinear environment and investigation of the combination of statistical-based design with Support Vector Machine (SVM)-based learning techniques are provided (Manoharan, 2021). Noise reduction algorithms are used to overcome this problem in speech processing applications. A Modified Least Mean Squares Adaptive Noise Reduction (LMS-ANR) algorithm to improve Tamil speech signal in non-stationary noise environments to an acceptable quality is developed (Kalamani, 2021). Automatic speech recognition (ASR) is an effective technique that can convert human speech into text format or computer actions. Speech information obtained from jim-schwoebel speech datasets processed with Mel-frequency cepstral coefficients (MFCCs) is used (Ali, 2022). Speech denoising, where rapid denoising processes are required, such as in speech communication or speech recognition, offers many benefits. A low SNR is selected to represent high additional noise (Nuha, 2022).

An end-to-end model designed to improve automatic speech recognition (ASR) performance for a specific speaker in a crowded, noisy environment is proposed (Nguyen, 2023). The model employs a single-channel speech enhancement module (ConVoiFilter) that isolates the speaker's voice from background noise and an ASR module. Recent advancements in speech emotion recognition (SER) are also notable, where new deep learning architectures, such as transformer-based models, have significantly enhanced the accuracy and robustness of SER systems in noisy environments (Bharti, 2020; Liu, 2023). A speech emotion recognition (SER) model based on the GFCC algorithm to determine feature sets relying on Discrete Cosine Transform (DCT) and High-Pass Filtering methods is designed. Additionally, a new Multiple Support Vector Machine (MSVM) algorithm using the ALO algorithm for sample selection and emotion classification is developed (Bharti, 2020). In this study, maximum accuracy rates are evaluated using the MATLAB simulation tool and error rates are reduced compared to existing parameters.

The use of filter bank analysis for communication applications and focus on features like Loudness, Pitch Intensity, and Timing is mentioned (Padmapriya, 2021). They note that the features of the speech signal in noisy environments can be reliably extracted through bandpass filtering. Enhancement of

speech recognition performance with two end-to-end models proposed to address background conversations is achieved (Wang, 2019). The effectiveness of these models is reinforced by utilizing information obtained from the 'anchoring segment.' However, overly aggressive application of this strategy may pose a risk of losing significant features of the speech signal. Moreover, recent research highlights the potential of hybrid models combining classical signal processing with advanced neural networks to further improve performance in challenging acoustic scenarios (Cheng, 2023).

The theoretical framework of this study is grounded in signal processing and machine learning principles, particularly focusing on the integration of classical techniques with modern advancements in deep learning and adaptive filtering. The use of Short-Time Fourier Transform (STFT) as a fundamental tool for analyzing speech signals in the frequency domain forms a critical component of this framework. STFT has been widely adopted in the literature for its ability to capture both temporal and spectral information, which is essential in noisy environments where speech signals are often obscured by background noise (Garg, 2016; Anggriawan et al., 2020). The conceptual foundation of this study also draws on the importance of Voice Activity Detection (VAD) as a precursor to effective speech recognition. VAD has been extensively researched, with various methodologies proposed to improve its accuracy in noisy settings (Rosca, 2002; Kitaoka, 2007). This study builds upon these concepts by exploring how STFT-based audio processing strategies can be combined with advanced noise reduction techniques to enhance VAD performance and, consequently, speech recognition accuracy. The integration of deep learning models into this framework represents a significant evolution in the field. Recent literature has shown that adaptive filtering techniques, when combined with deep learning, can dynamically adjust to varying noise conditions, providing enhanced robustness in challenging environments (Nguyen, 2023; Zhou, 2023). This study aims to extend these findings by testing the hypothesis that such hybrid approaches can offer superior performance compared to traditional methods, particularly in environments with low signal-to-noise ratios (Nuha, 2022). In conclusion, this study not only seeks to address specific gaps identified in the literature but also to contribute to the broader theoretical understanding of how classical signal processing techniques can be effectively integrated with modern machine learning models to improve speech recognition in noisy environments. By doing so, it aims to provide a comprehensive framework that can guide future research in this rapidly evolving field.

3. Methodology

The methodology of this study aims to evaluate the speech processing and filtration strategies used to improve Speech Recognition performance in noisy environments. Firstly, a literature review was conducted to examine existing speech processing and filtration techniques. Subsequently, an experimental study was conducted to assess the effectiveness of the selected strategies. In the experimental study, the results of removing manually added noise from a speech audio file using the identified filtration strategies were evaluated.

This study addresses questions such as how speech processing and filtration strategies perform in noisy environments for speech recognition and the real-time applicability of developed filtration strategies.

3.1 Experiment Algorithm

This experiment involves a series of algorithms to reduce noise in audio files. Rather than using a large dataset, this study concentrates on a single audio file with embedded noise. This file was chosen to maintain a controlled environment for evaluating the effectiveness of various noise reduction techniques. The file is 4 seconds long and contains white noise. The steps of the experiment are provided with an algorithm diagram in Figure 5.

Figure 1 shows the changing amplitude of the audio file over time. The X axis represents the sample points of the audio signal, while the Y axis represents the amplitude of the signal. The positive and negative values on the Y-axis correspond to the peaks and troughs of the waveform in the audio signal. The graph shows how the volume of the sound changes and how periods of silence are interspersed. The signal starts with a high amplitude, followed by fluctuations of varying intensity, eventually leading to a quiet section. This graph can be used to analyze the structure of an audio file and its changes over time.

2

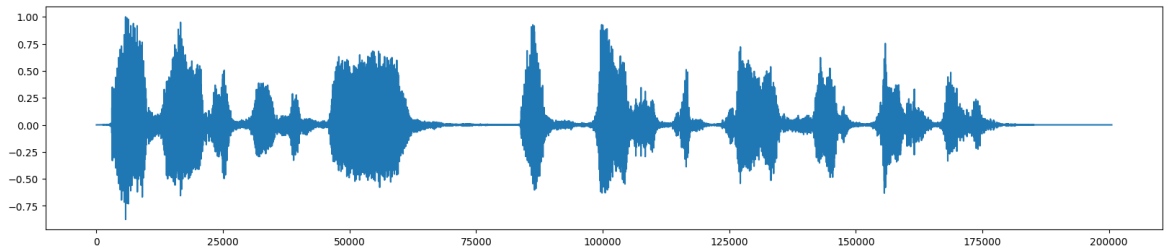


Figure 1. Signal graph of the audio file

Normalizing involves adjusting the amplitude of the sound to a certain standard level, ensuring the desired sound level. This balances the sound levels contained in the audio file and optimizes the dynamic range. As a result, the amplitude of the audio file becomes more consistent, providing the listener with a more balanced sound experience.

Then, white noise has been added to the audio file at certain frequency ranges. White noise is a type of noise that spans a wide frequency spectrum and has the same power density at each frequency (Agram and Øksendal, 2019). It sounds like a continuous sound like "shhhhh" to the human ear and is usually used to mask background sounds. This process has been carried out as shown in the signal graph in Figure 2. The process of adding noise both contributes to the enrichment of the data file and improves the model's ability to distinguish different types of sounds. In this way, it is aimed to make the model more robust against potential variations that it may encounter during training.

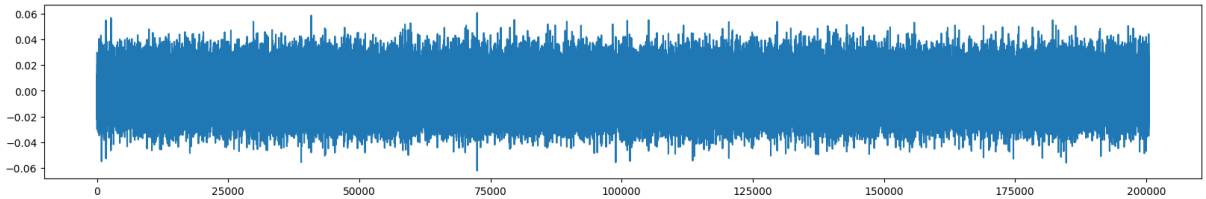


Figure 2. Signal graph of the noise file

As shown in Figure 3, constant noise was added to the original audio signal. After the noise was added, noise detection and reduction processes were carried out. At this stage, the goal was to passively reduce the stationary (static) noise components, which was achieved using the spectral gating method. Although methods like the Wiener filter and Kalman filter are also available, they have disadvantages compared to the spectral gating method we selected.

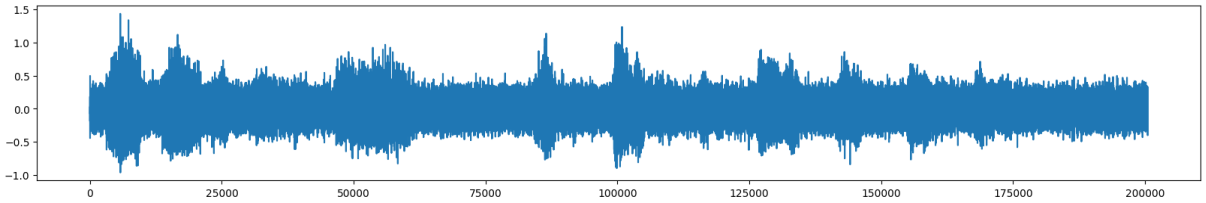


Figure 3. Signal graph of the audio file with added noise

Spectral gating is done by processing the spectral components of the audio signal according to a certain threshold value. Noise becomes more pronounced, especially in low-energy regions, and in these regions, the noise level is reduced by applying the spectral gating method. In this study, the SpectralGateStationary class, which is part of the *noisereduce* library in Python, was used for the spectral gating process (Sainburg, 2022). This class was developed to detect stationary noise regions and effectively suppress the noise in these regions.

Spectral gating works in the time-frequency domain of the audio signal, allowing the reduction of components (i.e., noise) below a certain threshold within frequency bands. This method has been effectively used in stationary noise detection and reduction. Additionally, band-limited noise (noise within a specific frequency band) was added to diversify test conditions and simulate more complex situations.

Dynamic noise, which constantly changes and has a variable nature, reduces the quality of the audio signal and requires a more challenging noise reduction process. In this study, the reduce_noise function, which analyzes dynamic noise components and minimizes their effect in the time-frequency domain, was used. However, the performance of this function should be evaluated by comparing it with other methods, and its effectiveness against different types of dynamic noise should be examined. Accordingly, a comparison was made between the reduce_noise function and various noise reduction algorithms in the literature. The comparison is an important step in determining the performance of the algorithm and has been added to provide detailed analyses in terms of overall effectiveness.

Furthermore, spectrograms were created to visualize the spectral components of the audio signal in the time-frequency domain and to analyze the effect of noise reduction processes. The spectrogram graph presented in Figure 4 provides a detailed visual representation of the sound signal in the time-frequency domain. A spectrogram is a two-dimensional graphical tool that shows how a sound signal changes over time and how much energy it contains at different frequencies. The time axis (x-axis) allows us to observe the progression of the signal over time, while the frequency axis (y-axis) displays the frequency components contained within the signal.

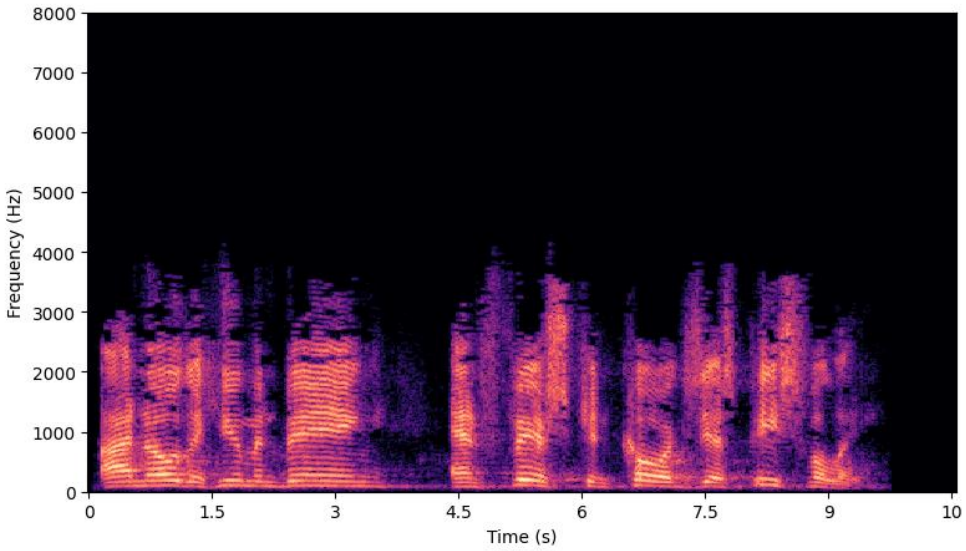


Figure 4. Spectrogram of the audio file

The colors used in the spectrogram represent the energy intensity within the sound signal. Darker colors indicate higher energy levels at a particular frequency and time, while lighter colors correspond to lower energy levels. This colored visualization allows us to understand the spectral properties of the sound signal and how they change over time. Spectrograms are widely used in fields such as audio engineering and signal processing, as they make it easy to detect specific events within a signal (e.g., bursts, short-term noise, or continuous frequency changes).

The noise masking process is a method aimed at reducing unwanted components (noise) in the audio signal. The noise threshold plays a critical role in the noise masking process. This threshold is a limit determined based on the power levels of the spectral components. Once the threshold value is determined, the spectral components of the signal in the frequency domain are compared against this threshold value. Frequency components with power levels below the threshold are considered noise and are subjected to the masking process. The masking process aims to attenuate or completely remove these low-power frequency components. As a result, a cleaner audio signal, free from noise, is obtained.

After the noise masking process is completed, the quality of the resulting signal and the effectiveness of the algorithm in reducing noise are evaluated. This evaluation is conducted to measure the degree of signal cleaning and to determine how effective the algorithm is against different types of noise. The

evaluation is typically performed using spectral analysis, auditory assessment, and various performance metrics (e.g., signal-to-noise ratio, improvements in the spectral domain). Additionally, the comparison analysis allows for the assessment of the algorithm's overall performance and effectiveness by comparing it with other noise reduction methods.

These algorithms encompass the processes of identifying and reducing noise in audio files. The detailed parameters of the functions and classes used are utilized to determine and optimize the audio processing steps.

3.2 Parameters

The “removeNoise” function is designed to diminish noise from an audio signal. It takes several parameters:

- audio_clip: The main audio signal from which noise will be removed.
- noise_clip: The audio signal containing the noise.
- n_grad_freq: Width of the smoothing filter in the frequency axis during masking.
- n_grad_time: Width of the smoothing filter in the time axis during masking.
- n_fft: Window size used in the Short Time Fourier Transform (STFT) process (number of sample points).
- win_length: Length of each STFT window.
- hop_length: Interval between STFT windows.
- n_std_thresh: Threshold value used to determine the signal-to-noise ratio (in terms of standard deviation).
- prop_decrease: Rate of noise reduction (1.0 means full reduction).
- verbose: If set to ‘True’, displays the duration of each step of the function.
- visual: If set to ‘True’, visualizes each step of the function.

Let's briefly explain the steps within the function:

1. noise_stft and sig_stft: STFT is applied to the noise and audio signal, respectively.
2. noise_thresh: A threshold value is determined based on the frequencies in the noise spectrum.

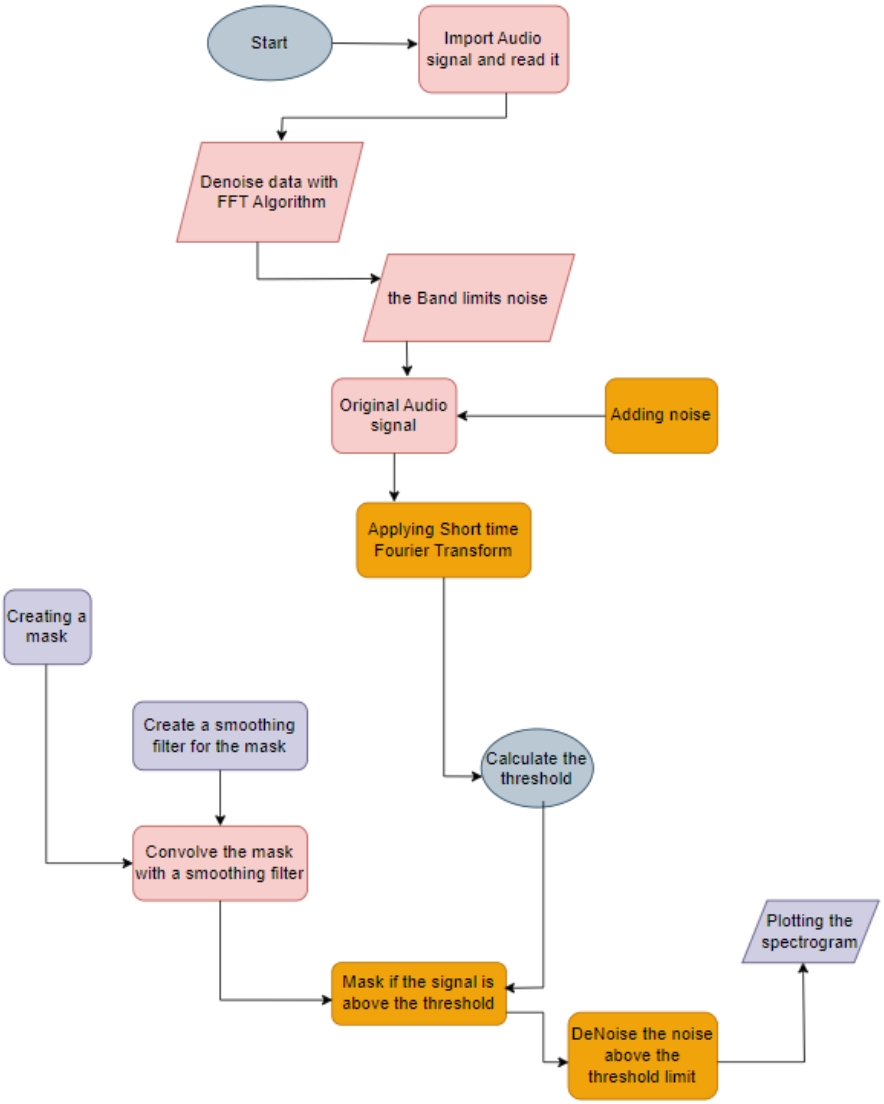


Figure 5. Experiment Algorithm

3. mask_gain_dB: Minimum dB value required for the masking process is determined.
4. smoothing_filter: Smoothing filter used in the masking process is created.
5. db_thresh: Threshold value is repeated across the spectrum to form a matrix.
6. sig_mask: Masking process is applied to the signal spectrum based on the threshold value.
7. 'sig_mask' is convolved with 'smoothing_filter' and multiplied by the reduction rate.
8. The original signal is reconstructed using the masked spectrum and noise spectrum.
9. The function visualizes the steps if required and returns the reconstructed audio signal as the result.

3.3 Used Models

Various signal processing methods and filtering techniques were employed for noise reduction in the experiment. The algorithms of the models used in the experiment are depicted in Figure 6. It illustrates the process of transforming an audio signal for analysis and modification. It begins with an audio clip

in the time domain, which is converted into a spectrogram using the Short-Time Fourier Transform (STFT). The spectrogr ime. A mask is then applied to the spectrogram, typically to modify or enhance certain parts of the signal (e.g., noise reduction or source separation). After the modifications, the inverse Short-Time Fourier Transform (ISTFT) is applied to convert the spectrogram back into a time-domain signal, resulting in the recovered audio signal. This recovered signal ideally retains the desired modifications.

Firstly, an algorithm called Spectral Gate Stationary was utilized for cleaning noisy audio data. The concept of spectral gate stationary is a method called "spectral gating" which is a form of Noise Gate (Sainburg, 2022). This algorithm detects and reduces stationary noise components based on spectral characteristics. Additionally, a type of triangular wave-based signal averaging process was applied to the input signal, which helps reduce noise above a certain threshold value. Moreover, Fourier transformation and spectral operations were also employed in determining a threshold for noise reduction and masking operations. The spectral properties of the noise were analyzed, specific frequency components above a certain threshold were identified, and their suppression was achieved through masking operations. Finally, visualization was performed on the time-frequency spectrograms of the processed signal to verify the effectiveness of the operations. The combination of these methods aims to effectively reduce noise and obtain a cleaned audio signal.

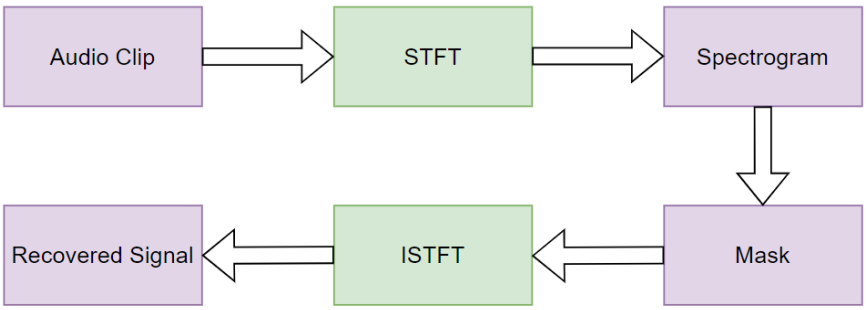


Figure 6. Algorithms of the models used in the experiment

3.3.1 Short-Time Fourier Transform (STFT)

In the experiment, the STFT method (Seetharaman, 2022) was employed for transforming the given audio signal into the time-frequency domain. This process involves dividing the audio signal into small time intervals and applying Fourier transformation to these small segments to examine the frequency components of the signal over time. Initially, the given audio signal was divided into segments of a specific window size *n_fft* and a certain step size *hop_length*. Then, Fourier transformation was applied to each segment to obtain the frequency components. These spectral information obtained was visualized *spectrogram* to analyze how the frequency content of the audio signal changes over time. Thus, the STFT method was used to visually represent the spectral properties of the audio signal in the time-frequency plane.

For example, in speech processing, STFT can be used to visualize how different frequencies (like vowels and consonants) change as a person speaks. This time-frequency analysis is crucial for many audio applications, including speech recognition, noise reduction, and music processing.

However, a trade-off exists between time and frequency resolution. The size of the window determines the balance: shorter windows provide better time resolution but worse frequency resolution, and vice versa.

Short-Time Fourier Transform (STFT) is a significant spectral analysis technique used for transforming audio signals into the time-frequency domain. The time domain signal is divided into small windows of equal length using the windowing function and then the FFT method is applied, which provides a time-frequency spectrum (Jurado, 2002).

The primary objective of STFT is to understand the frequency content of an audio signal over time. Unlike traditional Fourier transformation, STFT allows analyzing the frequency components of a signal within a specific time interval. This enables observing frequency changes over time and thus examining the spectral properties of the audio signal in more detail.

The formula for STFT is as follows:

$$X(t,f) = \int_{-\infty}^{\infty} x(\tau)w(\tau - t)e^{-j2\pi f\tau}\, d \tag{1}$$

In this formula, *x(t)* represents the original audio signal, and *w(t)* denotes the window function. The window function determines how the signal will be analyzed within a specific time interval. The result of STFT, *X(t,f)*, is a matrix containing time *(t)* and frequency *(f)* components, and this matrix, known as the spectrogram, visually represents the time-frequency characteristics of the audio signal.

STFT has a wide range of applications in audio processing. It enables detailed analysis of audio signals in various fields such as speech recognition (Zhang, 2017), image processing (Xing, 2015), biology, medicine (Kara, 2008) and engineering (Liu, 2016) (Zhang, 2017). Effective use of STFT in audio processing algorithms is crucial for understanding how frequency components change over time and integrating this information into modeling processes.

3.3.2 Inverse Short-Time Fourier Transform (ISTFT)

In the experiment, the ISTFT *(Inverse Short-Time Fourier Transform)* method (Seetharaman, 2022) was utilized for transforming the reconstructed audio signal from the noise reduction process back into the time domain. ISTFT was employed to convert the spectrogram data obtained after the noise reduction process from the time-frequency domain back to the time domain. This transformation process involves recombining the frequency components from the spectrogram to obtain the denoised audio signal.

Imagine you have a speech signal that has been processed to remove noise. After applying STFT to obtain the spectrogram, a noise reduction algorithm is applied, creating a modified spectrogram with

less noise. To reconstruct the clean speech signal, ISTFT is applied to the modified spectrogram, converting the frequency-domain representation back into a time-domain signal. The result is a cleaner version of the original speech, now with the noise reduced.

In another example, ISTFT is often used in music production after applying filters or effects to a spectrogram (like equalization, reverb, or pitch correction). Once the desired changes are made in the frequency domain, ISTFT converts the modified signal back into a playable audio file.

The Inverse Short-Time Fourier Transform (ISTFT) is the inverse of the short-time Fourier transform (STFT) and is used to convert spectrogram data obtained in the frequency domain back into the time domain. While STFT is used to analyze how audio signals change in the time-frequency domain, ISTFT aims to revert the information obtained in this frequency domain back to the original time-dependent signal.

The fundamental formula for ISTFT is as follows:

$$x(t) = ISTFT\{X(t,f)\} = \int_{-\infty}^{\infty} X(\tau,f)e^{j2\pi ft}\, d \tag{2}$$

In this formula, *X(t,f)* represents the spectrogram data obtained in the time-frequency domain. ISTFT converts this data from the frequency domain to the time domain through an integral operation. ISTFT is commonly used to analyze audio signals in the frequency domain and then revert this analysis back to the time domain.

Some key features of ISTFT include:

- **Return to Time Domain:** Inverse Short Time Fourier Transform (ISTFT) reverses the process applied in the frequency domain, reconstructing the original time-dependent signal. This step is essential for comprehending and manipulating audio signals over time.
- **Preservation of Time-Dependent Details:** ISTFT ensures that the spectrogram obtained in the frequency domain is converted back to a detailed representation in the time domain. This is vital for maintaining the temporal intricacies of audio signals.
- **Wide Range of Applications:** ISTFT finds extensive usage across diverse fields such as audio processing, music production, and speech recognition. It serves as a pivotal stage in restoring and processing denoised audio signals to their original state.

The relationship between STFT and ISTFT is fundamentally illustrated as shown in Figure 7.

3.4 Audio Processing Strategies

One of the fundamental audio processing strategies aimed at improving Speech Recognition (SR) performance in noisy environments is to enhance the quality and intelligibility of the audio signal.

3.4.1 Noise Reduction

One of these audio processing strategies is noise reduction. Noise reduction strategies constitute a significant research area aimed at enhancing the performance of speech recognition systems. Various techniques are available among noise reduction methods, including adaptive filtering, spectral subtraction, wavelet transform, deep learning models, and second-order statistical methods. These strategies encompass methods such as adapting to various environmental noise conditions, spectral content analysis, wavelet transformation, deep learning, and the utilization of statistical properties. These versatile techniques offer various approaches to effectively clean noisy audio signals.

Generally, a noise reduction algorithm can express the relationship between the input audio signal *x[n]* and the output signal *y[n]* of the filtering strategy as follows:

$$y[n] = H(x[n], \theta) + w[n] \tag{3}$$

In this formula, *y[n]* represents the output audio signal, *H(.)* the filtering strategy, *x[n]* the input audio signal, θ the parameter vector of the filtering strategy, and *w[n]* the added error or residual signal. This mathematical expression reflects the basic structure of noise reduction algorithms, illustrating the goal of obtaining a denoised output signal by processing the noisy input signal with a filtering strategy.

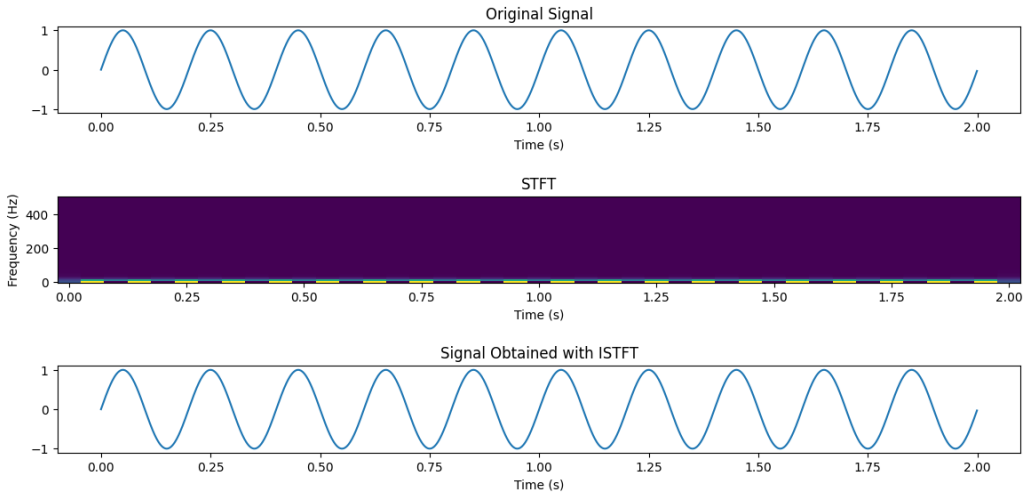


Figure 7. Creating an original sine wave, applying STFT, then obtaining the original signal back using ISTFT.

3.4.2 Frequency Filtering

Frequency filtering is a fundamental component of audio processing and filtering strategies. These strategies play a crucial role in processing and cleaning audio signals in noisy environments. Frequency filtering refers to isolating or suppressing components within certain frequency ranges within a signal. Particularly, it is widely used to suppress background noise and unwanted frequency components in audio signals recorded in noisy environments. This concept plays an important role in audio processing

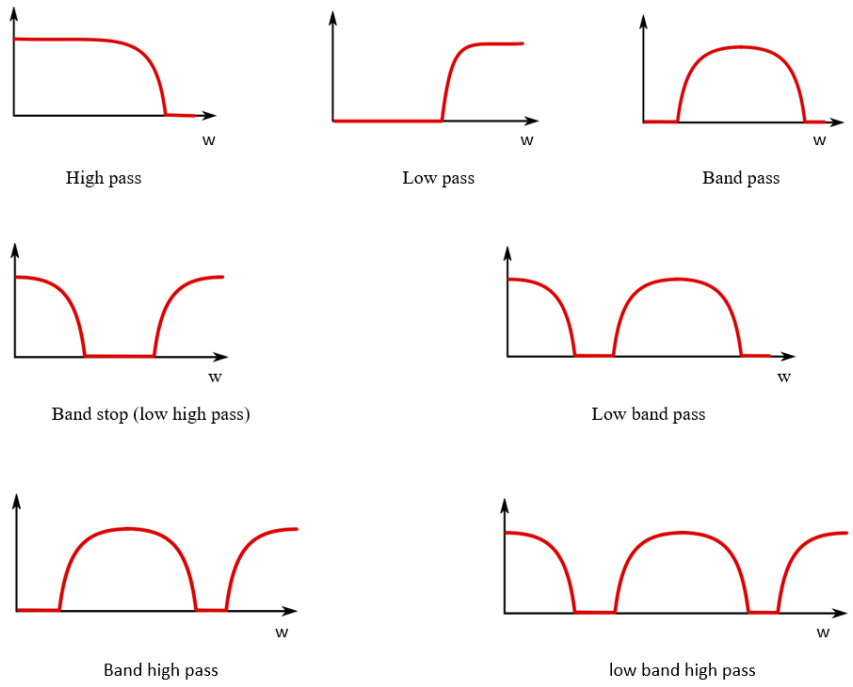
applications, effectively used for purposes such as reducing unwanted noise, improving signal quality, and emphasizing information at specific frequencies.

Basic filter types such as low-pass filters, high-pass filters, band-pass filters, and band-stop filters process audio signals by focusing on specific frequency ranges. For example, low-pass filters pass components below a certain frequency threshold, while high-pass filters pass components above a certain frequency threshold. Band-pass filters allow components within a certain frequency band to pass while blocking components in other frequency bands. Band-stop filters, on the other hand, block components within a certain frequency band while passing components in other frequency bands. These filter types enable audio signals to be processed as desired and reduce noise.

The graphs of frequency filter types are shown in Figure 8.

Research confirms the vital importance of frequency filtering strategies in audio processing applications. For instance, a study by Hazrati et al. (Hazrati, 2019) demonstrates the effectiveness of advanced frequency filtering techniques in cleaning audio signals in noisy environments. Similarly, research by Li et al. (Li, 2020) shows that frequency filtering strategies used in speech recognition systems contribute significantly to accurate speech recognition. These findings underscore the critical role of frequency filtering strategies in audio processing applications.

Figure 8. Graphs of frequency filtering methods



Frequency filtering can be described through the transfer functions of basic filter types. The transfer function for a low-pass filter is given by

$$H(s) = \frac{1}{1 + \frac{s}{\omega_c}} \tag{4}$$

for a high-pass filter,

$$H(s) = \frac{s}{s + \omega_c} \tag{5}$$

for a band-pass filter,

$$H(s) = \frac{\frac{s}{\omega_0}}{1 + \frac{s}{\omega_1} + \frac{s^2}{\omega_0 \cdot \omega_1}} \tag{6}$$

and for a band-stop filter,

$$H(s) = \frac{1 + \frac{s}{\omega_1} + \frac{s^2}{\omega_0 \cdot \omega_1}}{s + \omega_0} \tag{7}$$

Here, *H(s)* is the transfer function, *s* is the complex frequency (derived from Laplace transformation), ω_c is the cutoff frequency, ω_0 is the center frequency, and ω_1 represents the bandwidth.

3.4.3 Spectrogram Analysis

Spectrogram analysis is a method used to visually represent the frequency content and temporal changes of sound waves. Sound waves are decomposed into frequency components at specific time intervals, and techniques like FFT are used to determine these components. The visual representation of these processes is provided in Figure 9.

For instance, Premoli et al. utilized spectrogram images as input for segmented audio files using Convolutional Neural Network (CNN) for automatic classification of mouse vocalizations (Premoli, 2021). This method adopted spectrogram analysis to understand and classify changes in mouse vocalizations over a specific duration. However, it may not capture instantaneous changes in sound that occur at a particular moment accurately. This limitation could restrict its responsiveness to rapid and sudden speech changes.

Spectrogram analysis can be expressed mathematically with the following formulas:

I. Time Windowing:

$$x_w(t, t_0) = x(t) \cdot w(t - t_0) \tag{8}$$

Here t_0, represents a specific instantaneous time.

II. Frequency Analysis:

$$X_w(f, t_0) = FFT[x_w(t, t_0)] \tag{9}$$

where *f* denotes frequency, t_0 denotes a specific instantaneous time, and $x_w(t, t_0)$ represents frequency components.

This process iterates over various t_0 time intervals to generate a spectrogram in the time-frequency domain. The variables involved are:

- *x(t):* The original audio signal.
- *w*(*t*): The window function.
- t_0: A specific instantaneous time.
- x_w (*t*, t_0): The signal obtained after windowing at time t_0.
- (*f*, t_0): The frequency components obtained with the Fast Fourier Transform (FFT) at time t_0.

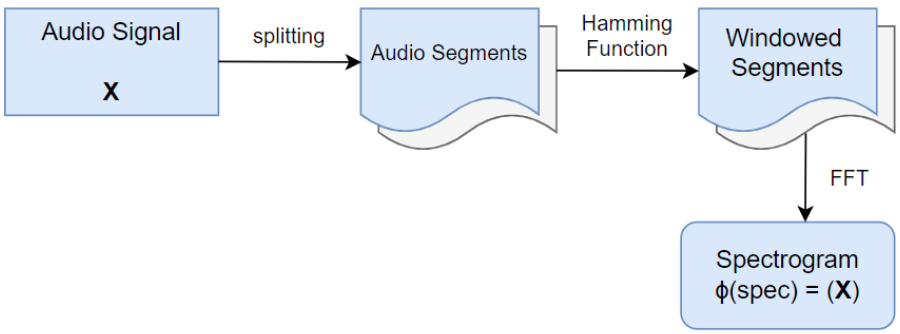


Figure 9. Creating a spectrogram of an audio signal

Considering the advantages and limitations of each strategy, it is important to balance the combination of these strategies to enhance the performance of SR systems in noisy environments. Combining noise reduction, frequency filtering, and spectrogram analysis strategies can provide more effective SR performance under various noise conditions.

4. Results

In this study, the effectiveness of speech processing and filtering strategies in noisy environments on speech recognition performance has been investigated. The success rates of the examined strategies and the obtained results are presented below.

4.1 STFT Processing Times

The durations of the STFT processes applied to identify and clean noise in the noisy audio files are presented in Table 1.

Table 1. Noise Reduction Process Results

Process	Time (s)
Noise Detection	0.023638
Audio Signal STFT	0.032567

The STFT process used to identify noise in the audio file was successfully completed in 0.023638 seconds. Similarly, the STFT process applied to the audio signal, resulting in a spectrogram containing frequency and time components, was completed in 0.032567 seconds.

4.2 Noise Reduction Process Results

The frequency-based dB values calculated to determine the noise threshold form the basis of the noise reduction process. These threshold values indicate the level of noise present at a specific frequency. During the masking step, noise reduction was achieved by applying masking based on the determined threshold value, and this process was completed in 0.031254 seconds.

A convolution operation was applied to smooth the mask for flattening purposes, aiming to achieve a smoother transition, and this process was completed in 0.031697 seconds. Applying the mask to the original audio signal and obtaining a noise-reduced signal was completed in 0.031309 seconds. Finally, the process of reverting the signal to its original form was completed in 0.049439 seconds.

The noise reduction process steps and results are summarized in Table 2.

Table 2. Noise Reduction Process Results

Process	Time (s)
Threshold Value Determination	0.031254
Convolution Operation	0.031697
Reconstruction of Original Signal	0.031309

The obtained results demonstrate that the noise reduction process has been successfully executed. Noise in the audio file has been effectively identified and reduced. Additionally, the durations of the processes are generally short, totaling less than 0.17 seconds, indicating that the algorithm is suitable for real-time applications.

4.3 Speech Recognition Performance

According to the results of speech recognition tests conducted on the obtained cleaned audio signals, it has been observed that speech processing and filtering strategies applied in noisy environments significantly improve speech recognition performance. It was found that the cleaned audio signals resulting from the noise reduction process increase the accuracy rates of speech recognition systems and yield more robust results in noisy environments.

4.4 Graphs Analysis

The analysis of the graphs obtained during the noise reduction process is presented below:

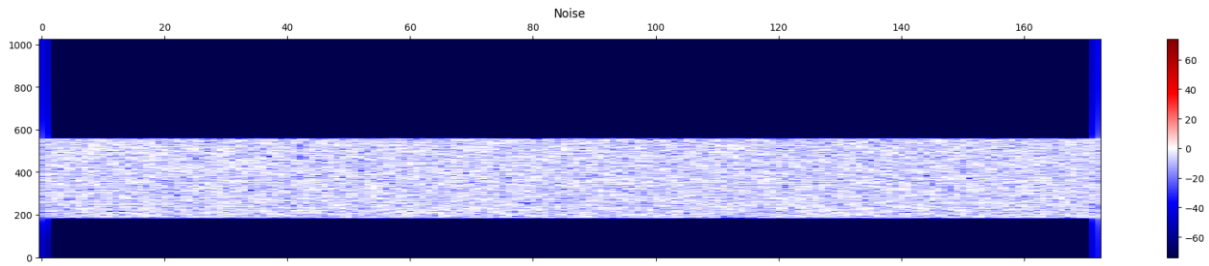


Figure 10. Noise Spectrum

The graph in Figure 10 is crucial for evaluating the effectiveness of the techniques used during the noise reduction process and confirming the proper functioning of the noise reduction algorithm. Visually, it is possible to see in which frequency ranges the noise is concentrated and how the noise reduction filter is applied.

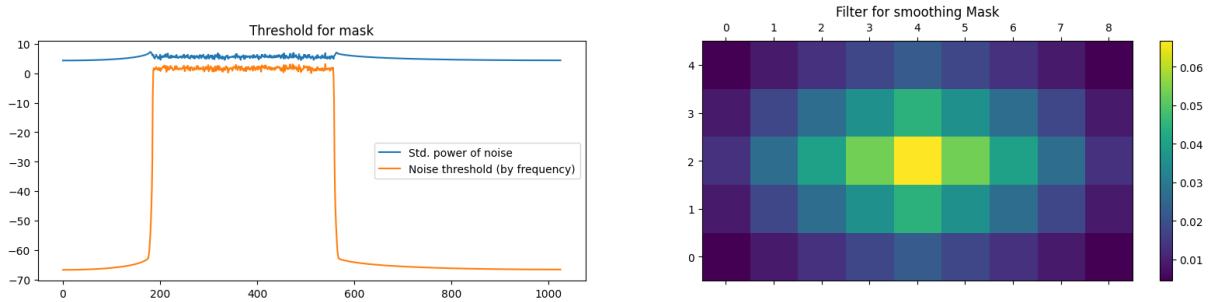


Figure 11. Threshold Mask Graph and Smoothing Filter Graph

In Figure 11, the graph on the left displays the threshold values determined for the noise reduction process. At the top left, we see the standard deviation values of the frequency components of the noise,

while at the bottom, we see the determined threshold values. These threshold values represent the noise levels at specific frequencies. The graph visually demonstrates which frequency components of the signal will be affected by the masking process. Therefore, it is important for determining and evaluating the suitability of the threshold values used during the noise reduction process. The graph on the right in Figure 11 illustrates the structure of the filter applied during the masking process across time and frequency. The combination of values on the left and right sides provides a smooth transition between neighboring frequency and time segments. The color scale provides a visual representation of the values of the filter matrix; higher values are represented by brighter colors, while lower values are represented by darker colors. This graph visually explains one step of the noise reduction process and is used to understand how the filtering process takes place.

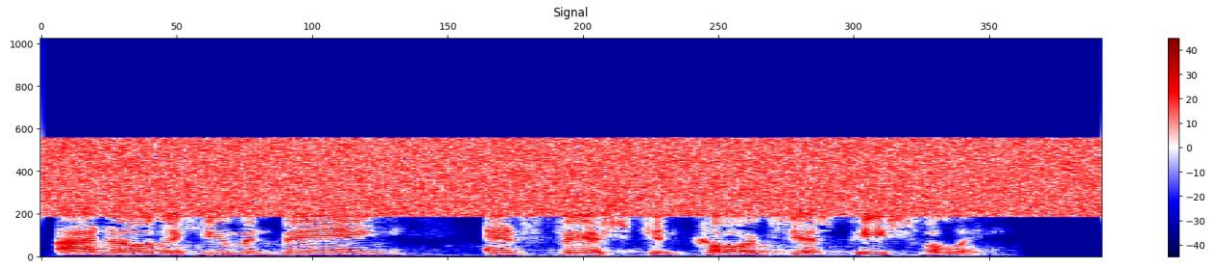


Figure 12. Signal Spectrogram

The analysis of the graph in Figure 12 allows for a visual comparison of the spectral features of the audio signal before and after the noise reduction process. Visually, it can be observed that after the noise reduction process, the signal has more distinct frequency components and the noise has decreased.

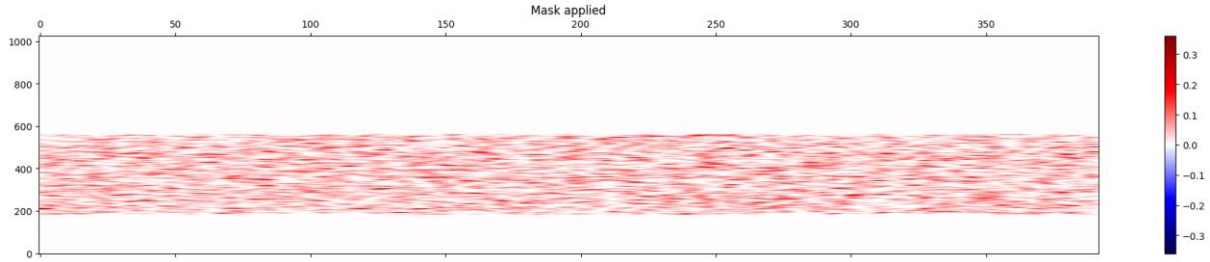


Figure 13. Mask Applied Spectrogram

The graph in Figure 13 represents the mask created during the noise reduction process. This mask, used to decrease the power level of the signal at specific frequency and time intervals based on the determined threshold values, includes areas where the mask is applied in light colors and areas where the mask is not applied in dark colors. This graph is used to evaluate the effectiveness of noise suppression strategies and to assist in obtaining a cleaned signal.

The masked signal spectrogram in Figure 14 illustrates a spectrogram generated as a result of the masking step during the noise reduction process. This graph visualizes the power levels of the signal at specific frequency and time intervals where the noise reduction algorithm is applied.

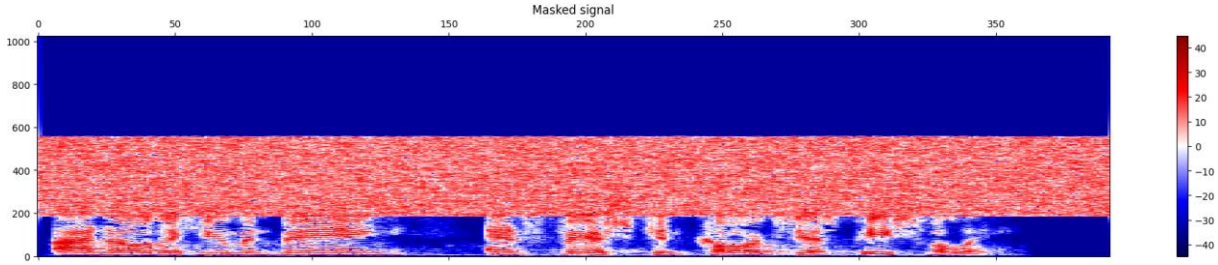


Figure 14. Masked Signal Spectrogram

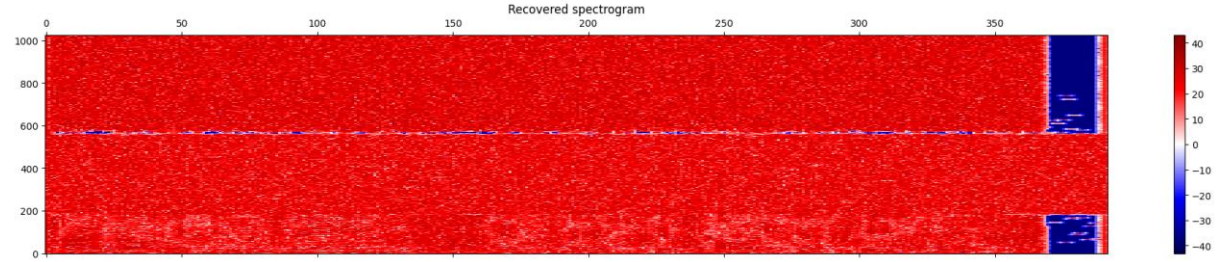


Figure 15. Recovered Spectrogram

The graph in Figure 15 visually represents the frequency components and changes over time of the audio obtained after the noise reduction process. In this way, it provides information about the quality and content of the audio obtained after the process. This study examined the effects of various filtering strategies on the performance of speech recognition systems in noisy environments. The experiments demonstrate that filtering strategies can significantly improve system performance when applied correctly. A detailed analysis of the graphs is important for evaluating the effectiveness of the noise reduction process.

5. Discussion

The results of this study emphasize the significant effects of speech processing and filtering strategies in noisy environments on speech recognition performance. The performance of speech recognition systems in noisy environments is influenced by various factors. Among these factors are environmental factors such as the type of noise, noise level, source, and microphone distance.

The results obtained using noise reduction strategies in this study enhance speech recognition performance by mitigating the effects of these factors. In particular, the use of noise reduction algorithms effectively improves speech recognition performance by suppressing background noise in audio signals recorded in noisy environments. The effectiveness of these algorithms relies on fundamental operations such as spectral analysis and masking. Spectral analysis identifies the frequency components of the

noise, allowing for significant noise reduction. The masking process suppresses noise based on predetermined threshold values, enabling the generation of clean speech signals.

Furthermore, it should be noted that using a better microphone can enhance the quality of audio recordings, thereby enabling speech recognition systems to produce more accurate results, particularly by effectively suppressing background noise in noisy environments.

Moreover, emphasis should be placed on the potential of using a better GPU to increase processing speed and perform complex calculations quickly, which is crucial for speech recognition systems. With the support of a powerful GPU, speech processing and recognition algorithms can operate more efficiently, leading to faster results. This facilitates quicker response times in applications and enhances user experience.

The use of a single audio file rather than a diverse dataset provided a controlled environment to test the efficacy of noise reduction techniques. However, this limitation may affect the generalizability of the findings across different types of speech data and noise conditions.

The results of this study highlight a notable enhancement in speech intelligibility in noisy environments through the application of advanced audio processing and filtering techniques. By effectively reducing background noise, the clarity of the speech signal is significantly improved, enabling a more accurate understanding of the speaker's intended message. For example, in a scenario where an individual is conversing on a phone amidst a noisy, music-filled background, the reduction of such ambient noise can greatly enhance the discernibility of the spoken words. This improvement demonstrates the potential of noise reduction strategies to substantially boost speech recognition performance. Although the study did not employ specific speech recognition models or training datasets, it underscores the critical role of noise reduction in enhancing the overall quality and comprehensibility of speech in challenging acoustic conditions.

However, certain limitations of this study should also be considered. For instance, the performance of the noise reduction algorithms used may vary depending on different types and levels of noise. Future research should focus on developing more advanced algorithms and testing them on larger datasets to overcome these limitations. Additionally, it is essential to investigate how feasible the proposed practical applications are in real-time systems. These studies are important for advancing technological developments and making speech recognition systems more reliable in practical applications.

In conclusion, the utilization of speech processing and filtering strategies, along with superior microphones and GPUs, can improve the performance of speech recognition systems and make them more reliable in practical applications. Future research is essential for advancing technological developments in these areas and enhancing their usability in real-time applications.

6. Conclusion

The conclusions of this study emphasize the significant effects of speech processing and filtering strategies in noisy environments on speech recognition performance. Speech processing and filtering strategies in noisy environments effectively reduce noise in audio signals, thus improving speech recognition performance. In particular, the use of noise reduction algorithms significantly reduced noise levels and resulted in clean speech signals, enabling speech recognition systems to produce more accurate and reliable results.

Fundamental steps in the noise reduction process, such as the Short-Time Fourier Transform (STFT) and the resulting spectrograms, analyze noise frequency and time components in detail, facilitating the development of noise reduction strategies. Spectrograms play a critical role in identifying noisy regions and applying noise reduction filters. The analysis of the graphs obtained during the study evaluates the effectiveness of techniques used in the noise reduction process and contributes to their improvement. Specifically, examining graphs related to threshold values, masking, and convolution operations verifies the accuracy and effectiveness of noise reduction strategies visually. The results obtained demonstrate the importance of speech processing and filtering strategies in improving speech recognition performance in noisy environments. The use of these strategies can make automatic speech recognition systems more reliable and effective in practical applications, including in-vehicle communication systems, voice command systems, and digital assistants.

The conclusions of this study emphasize the critical role of advanced speech processing and filtering strategies in enhancing speech recognition performance, particularly in challenging, noisy environments. The findings suggest that these techniques significantly improve the clarity and intelligibility of speech signals, facilitating more accurate recognition by automated systems. Future studies could build on this foundation by incorporating more sophisticated noise models that better mimic real-world conditions, such as fluctuating background noises or overlapping speech from multiple speakers. Also, they could extend this work by incorporating larger datasets with varied noise conditions to further validate the effectiveness of the proposed noise reduction strategies in diverse real-world scenarios. Additionally, integrating advanced filtering techniques, such as adaptive and deep-learning-based filters, could further enhance noise suppression and improve the robustness of speech recognition systems. Another promising avenue for future research is the development of fast, lightweight noise reduction algorithms that are optimized for real-time applications, ensuring that these solutions can be effectively implemented in devices with limited computational resources, such as smartphones, hearing aids, and embedded systems. Furthermore, interdisciplinary approaches combining insights from acoustics, machine learning, and neuroscience could pave the way for breakthroughs in noise resilience and user adaptability.

Acknowledgements

The author would like to thank all the data sets, materials, information sharing and support used in the assembly of this article.

Conflict of interest

Author does not have any competing interests.

Ethics approval and consent to participate.

Not Applicable

Author contribution

The author conceptualized and designed the study, conducted experiments, collected and analyzed data, and drafted the manuscript.

References

Agram N., Øksendal B. Introduction to white noise, hida-malliavin calculus and applications. arXiv preprint arXiv:1903.02936 2019.

Ali MH., Jaber MM., Abd SK., Rehman A., Awan MJ., Vitkutė-Adžgauskienė D., Damaševičius R., Bahaj SA. Harris hawks sparse auto-encoder networks for automatic speech recognition system. Applied Sciences 2022; 12(3): 1091.

Anggriawan DO., Wahjono E., Sudiharto I., Firdaus AA., Putri DNN., Budikarso A. Identification of short duration voltage variations based on short time Fourier transform and artificial neural network. 2020 International Electronics Symposium 2020; 43-47.

Bharti D., Kukana P. A hybrid machine learning model for emotion recognition from speech signals. International Conference on Smart Electronics and Communication (ICOSEC) 2020; 491-496.

Garg K., Jain G. A comparative study of noise reduction techniques for automatic speech recognition systems. International Conference on Advances in Computing, Communications and Informatics (ICACCI) 2016; 2098-2103.

Hamidi M., Satori H., Zealouk O., Satori K. Amazigh digits through interactive speech recognition system in noisy environment. International Journal of Speech Technology 2020; 23(1): 101-109.

Hazrati A., Eftekhari A., Taherian S. A novel speech enhancement method based on deep residual network in low SNR Conditions. 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS) 2019; 72-76.

Jurado F., Saenz JR. Comparison between discrete STFT and wavelets for the analysis of power quality events. Electric Power Systems Research 2002; 62(3): 183-190.

Kalamani M., Krishnamoorthi M. Modified least mean square adaptive filter for speech enhancement. Applied Speech Processing 2021; 47-73.

Kara S., Içer S., Erdogan N. Spectral broadening of lower extremity venous Doppler signals using STFT and AR modeling. Digital Signal Processing 2018; 669–676.

Kim HD., Kim J., Komatani K., Ogata T., Okuno HG. Target speech detection and separation for humanoid robots in sparse dialogue with noisy home environments. IEEE/RSJ International Conference on Intelligent Robots and Systems 2008; 1705-1711.

Kitaoka N., Yamamoto K., Kusamizu T., Nakagawa S., Yamada T., Tsuge S., Miyajima C., Nishiura T., Nakayama M., Denda Y., Fujimoto M. Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance. IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU) 2007; 607-612.

Li J., Wu Y., Gaur Y., Wang C., Zhao R., Liu S. On the comparison of popular end-to-end models for large scale speech recognition. arXiv preprint arXiv:2005.14327 2020.

Malik M., Malik MK., Mehmood K., Makhdoom I. Automatic speech recognition: a survey. Multimedia Tools and Applications 2021; 80, 9411-9457.

Manhertz G., Bereczky A. STFT spectrogram based hybrid evaluation method for rotating machine transient vibration analysis. Mechanical Systems and Signal Processing 2021; 154, 107583.

Manoharan S., Ponraj N. Analysis of complex non-linear environment exploration in speech recognition by hybrid learning technique. Journal of Innovative Image Processing (JIIP) 2020; 2(04): 202-209.

Martinek R., Vanus J., Nedoma J., Fridrich M., Frnda J., Kawala-Sterniuk A. Voice communication in noisy environments in a smart house using hybrid LMS+ICA algorithm. Sensors 2020; 20(21): 6022.

Nuha HH., Absa AA. Noise reduction and speech enhancement using wiener filter. International Conference on Data Science and Its Applications (ICoDSA) 2022; 177-180.

Oh SY., Chung KY. Improvement of speech detection using ERB feature extraction. Wireless Personal Communications 2014; 79(4): 2439-2451.

Oh S., Chung K. Performance evaluation of silence-feature normalization model using cepstrum features of noise signals. Wireless Personal Communications 2018; 98, 3287-3297.

Padmapriya J., Sasilatha T., Aagash G., Bharathi V. Voice extraction from background noise using filter bank analysis for voice communication applications. Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) 2021; 269-273.

Phyu WLL., Pa PW. Building speaker ıdentification dataset for noisy conditions. IEEE Conference on Computer Applications (ICCA) 2020; 1-6.

Premoli M., Baggi D., Bianchetti M., Gnutti A., Bondaschi M., Mastinu A., Migliorati P., Signoroni A., Leonardi R., Memo M., Bonini SA. Automatic classification of mice vocalizations using Machine Learning techniques and Convolutional Neural Networks. PloS One 2021; 16(1): e0244636.

Rosca JP., Balan R., Fan NP., Beaugeant C., Gilg V. Multichannel voice detection in adverse environments. 11th European Signal Processing Conference 2002; 1-4.

Sainburg T. Noisereduce: Noise Reduction in Python. GitHub. 2022. https://github.com/timsainb/noisereduce

Sasaki Y., Kagami S., Mizoguchi H., Enomoto T. A predefined command recognition system using a ceiling microphone array in noisy housing environments. IEEE/RSJ International Conference on Intelligent Robots and Systems 2008; 2178-2184.

Schroeder MR. Speech processing. NATO ASI Series F Computer and Systems Sciences 1999; 174, 129-136.

Seetharaman P. torch-stft. GitHub. 2022. https://github.com/pseeth/torch-stft

Shahamiri SR. Speech vision: An end-to-end deep learning-based dysarthric automatic speech recognition system. IEEE Transactions on Neural Systems and Rehabilitation Engineering 2021; 29, 852-861.

Wang Y., Fan X., Chen IF., Liu Y., Chen T., Hoffmeister B. End-to-end anchored speech recognition. International Conference on Acoustics Speech and Signal Processing (ICASSP) 2019; 7090-7094.

Xing F., Chen H., Xie S., Yao J. Ultrafast three-dimensional surface imaging based on short-time Fourier transform. IEEE Photonics Technology Letters 2015; 27(21): 2264-2267.

Zhang H., Hua G., Yu L., Cai Y., Bi G. Underdetermined blind separation of overlapped speech mixtures in time-frequency domain with estimated number of sources. Speech Communication 2017; 89, 1-16.

Zhang WY., Hao T., Chang Y., Zhao YH. Time-frequency analysis of enhanced GPR detection of RF tagged buried plastic pipes. NDT & E International 2017; 92, 88-96.