Enhancement of Speech Signals Using Neural Networks with Spectral Subtraction

by

Amirarsalan DARABPOUR

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT OF A MASTER'S DEGREE
WITH THESIS IN ELECTRICAL ENGINEERING
M.A.Sc.

MONTREAL, APRIL 8, 2024

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Gheorghe Marcel Gabrea, Thesis Supervisor
Department of Electrical Engineering, École de technologie supérieure

Ms. Sylvie Ratté, Chair, Board of Examiners
Department of Software Engineering and IT, École de technologie supérieure

Mr. Jean-Marc Lina, Member of the Jury
Department of Electrical Engineering, École de technologie supérieure

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON MARCH 18, 2024

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

I would like to extend my heartfelt thanks to Professor Gheorghe Marcel Gabrea, my Thesis supervisor, for his invaluable help, support, and guidance during my academic journey and research endeavors.

I would like to express my appreciation to my friends who have consistently stood by me and provided endless encouragement throughout the course of my thesis.

To my parents, words cannot express the depth of my gratitude for their strong support and consistent faith in me. My brother and sister also deserve special mention; their perpetual encouragement and firm belief in my abilities have been a driving force.

Lastly, but by no means least, my deepest acknowledgments go to my girlfriend, Roya. Her unparalleled support and profound understanding have been the foundations of strength for me throughout this journey.

Amélioration des signaux vocaux à l'aide de réseaux neuronaux et de la Soustraction Spectrale

Amirarsalan DARABPOUR

RÉSUMÉ

Aujourd'hui, de nombreux domaines et moyens de communication tels que les télécommunications, la reconnaissance vocale et les systèmes audiovisuels utilisent l'amélioration de la parole comme moyen d'améliorer la qualité des signaux vocaux, généralement en réduisant le niveau de bruit de fond. Les signaux vocaux contiennent souvent un bruit sous-jacent, provenant soit du processus d'acquisition, soit du canal de transmission. Ces dernières années, des recherches importantes ont été menées dans le domaine de l'amélioration de la qualité de la parole à l'aide de techniques d'apprentissage automatique. Ces techniques ont été utilisées dans de nombreuses tâches de traitement de la parole car elles ont donné des résultats très satisfaisants. Par conséquent, dans cette thèse, l'objectif principal de notre projet a été d'améliorer les signaux vocaux, en utilisant un cadre basé sur l'apprentissage automatique à l'aide de réseaux neuronaux.

Les signaux vocaux sont composés de segments voisés et non voisés et dans l'amélioration de la parole, la classification des segments voisés et non voisés d'un signal vocal est une tâche importante car elle permet d'améliorer les signaux vocaux de manière ciblée. Notre modèle utilise un algorithme avec un processus de fenêtrage qui améliore sa précision par rapport aux autres méthodes. Il s'agit de diviser le signal d'entrée en trames ou fenêtres de courte durée et d'analyser chaque fenêtre séparément pour déterminer si elle contient un signal de parole ou de non-parole.

Notre cadre basé sur les réseaux neuronaux a été mis en œuvre afin de remplir deux tâches fondamentales. La première consiste à classer les signaux voisés et non voisés et la seconde à améliorer les signaux voisés et non voisés afin d'obtenir un signal vocal amélioré. Pour réaliser la classification, nous avons utilisé l'ensemble de données NOIZEUS pour entraîner nos modèles. Nous avons développé avec succès un cadre complet qui classifie les segments voisés et non voisés des signaux vocaux bruyants. L'amélioration repose sur un critère basé sur le type de chaque fenêtre. Cette approche nous permet d'appliquer des méthodes d'amélioration spécifiques à différents segments, ce qui permet d'obtenir un signal final entièrement amélioré et débruité. Nos résultats ont montré que le système de classification des signaux voisés et non voisés ainsi que notre stratégie de nettoyage des signaux corrompus par du bruit additif ont été très efficaces, permettant d'obtenir un signal vocal réellement amélioré en termes de rapport signal/bruit et de qualité d'écoute.

Mots-clés: amélioration de la parole, réseaux de neurones, soustraction spectrale, classification de la parole et de la non-parole

Enhancement of Speech Signals Using Neural Networks with Spectral Subtraction

Amirarsalan DARABPOUR

ABSTRACT

Today, many domains and communication mediums such as telecommunications, speech recognition and audio-visual systems use speech enhancement as a way of improving the quality of speech signals, typically by reducing the level of background noise. Speech signals often contain underlying noise, originating either from the acquisition process or the transmission channel. In recent years, there has been significant research in the field of speech enhancement using machine learning techniques. These techniques have been used in many speech processing tasks since they have provided very satisfactory results. Accordingly, in this thesis, the main objective of our project has been to improve speech signals, using a framework based on machine learning using neural networks.

Speech signals are composed of speech and non-speech segments and in speech enhancement, classifying speech and non-speech segments of a speech signal is an important task as it helps for targeted enhancement of speech signals. Our model uses an algorithm with a windowing process which improves its accuracy compared to other methods. It involves dividing the input signal into short-time frames or windows and analyzing each window separately to determine if it contains a speech or non-speech signal.

Our neural network-based framework has been implemented in order to fulfill two fundamental tasks. The first task is to classify speech and non-speech signals and the second task consists of enhancing the speech and non-speech signals in order to obtain an improved speech signal as a result. To achieve the classification, we have used the NOIZEUS dataset for training our models. We successfully developed a comprehensive framework that classifies speech and non-speech segments of the noisy speech signals. The enhancement relies on a criterion that is based on the type of each window. This approach allows us to apply specific enhancement methods to different segments, resulting in a fully enhanced and denoised final signal. Our results have shown that the classification scheme of speech and non-speech signals together with our cleaning strategy on signals corrupted by additive noise have been very effective, obtaining a really improved speech signal in terms of SNR and listening quality.

Keywords: speech enhancement, neural networks, spectral subtraction, speech and non-speech classification

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
DL	Deep Learning
CNN	Convolutional Neural Network
DNN	Deep Neural Network
MFCC	Mel-frequency Cepstral Coefficients
MMSE	Minimum Mean Square Error
MAP	Maximum A Posteriori
EMD	Empirical Mode Decomposition
IMF	Intrinsic Mode Functions
NMF	Non-negative Matrix Factorization
SNR	Signal-to-Noise Ratio
ANN	Artificial Neural Network
HMM	Hidden Markov Model
SND	Speech/Non-Speech Detection
VAD	Voice Activity Detection
MLP	Multilayer Perceptron
SAD	Speech Activity Detection
SVM	Support Vector Machine
LTSD	Long-term Spectral Divergence

LTSV	Long-term Signal Variability
RNN	Recurrent Neural Network
SCG	Scaled Conjugate Gradient
LSTM	Long Short-Term Memory
STFT	Short-Time Fourier Transform
FFT	Fast Fourier Transform
PSD	Power Spectral Density

LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

kHz	Kilohertz
Hz	Hertz
kbps	Kilobits per second
ms	Millisecond
dB	Decibel
GB	Gigabyte

INTRODUCTION

0.1 Context and motivation

It is always desirable for any communication medium that noise has little or no effect and that signal also remains unaffected for a medium so popular as speech. However, looking at the different types of noises and their sources, noise-free speech is difficult to acquire. It is evident that if available speech has been degraded by noise, the effect of that noise should ideally be minimized, if not eliminated. This area of research is commonly referred to as speech enhancement. Speech enhancement is used in almost all modern communication systems to improve the quality of speech. It is closely related to speech restoration because they both reconstruct and restore the signal after degradation. However, speech enhancement helps in refining the original signal to create a more favorable listening experience.

Enhancement of speech can be used in different settings, such as areas with interfering background noise, such as noisy streets, train stations, cars and noisy buildings in general. Speech enhancement systems are used in telecommunications, hearing aids and assistive listening devices, audio and video conferencing, automatic speech recognition, speaker identification and verification, speech-based biometric systems, voice conversion etc [Loizou (2013)].

In general, speech enhancement methods can be classified into four categories: Conventional methods, Adaptive filtering methods, Machine learning methods, and Multi-modal methods. Moreover, these techniques can be classified based on various factors, such as the type of algorithm, the number of input channels, and whether they are unimodal or multimodal. Algorithms can be adaptive or non-adaptive. Adaptive filters adjust their impulse response to eliminate correlated signal components in the speech input, requiring minimal prior knowledge of the signal and noise characteristics. They can adapt to non-stationary conditions. Non-adaptive, or fixed filters, require prior knowledge of both the signal and noise. These filters pass frequencies present in the signal and discard the frequency bands occupied by the noise. Input channels can be single, dual, or multiple. Single-channel enhancement, also known as monaural enhancement, is used when only one input channel is available, such as in mobile telephony. Multichannel speech enhancement involves noisy observations from two or more sensors. When there are only two channels in a multichannel system, it is called binaural enhancement. Speech enhancement techniques can also be unimodal or multimodal. Unimodal audio enhancement focuses on enhancing audio perception using only the auditory sense. In contrast, multimodal speech enhancement combines audio perception with other senses, such as auditory and visual senses, to improve performance. Multimodal speech enhancement, which involves the use of senses other than the auditory sense (e.g., speech, vision, language, and text), can significantly enhance performance compared to unimodal approaches [Taha *et al.* (2018)].

Over the years, researchers have tackled the issue of speech enhancement in different ways. Speech enhancement was primarily approached through a variety of signal processing techniques. In the early stages of this field, spectral subtraction methods, statistical-based algorithms, adaptive noise cancellation, Wiener filtering and subspace enhancement methods were used. Spectral subtraction aimed to suppress noise by estimating the noise spectrum and subtracting it from the noisy speech spectrum. Wiener filtering, on the other hand, was a linear estimation approach that applied a statistical method to minimize the mean square error between the estimated clean speech signals and the original, uncorrupted signals. It operated under the assumption that both noise and speech are stationary signals and uses prior statistical knowledge of their spectral characteristics. Adaptive noise cancellation employs adaptive filters to track changes in the noise environment and cancel out noise from the speech signal. These traditional methods, while effective in certain conditions, had some limitations. They often required extensive fine-tuning and customization to work effectively under diverse circumstances and were predominantly based on linear signal processing techniques, which limited their ability to deal with non-linear

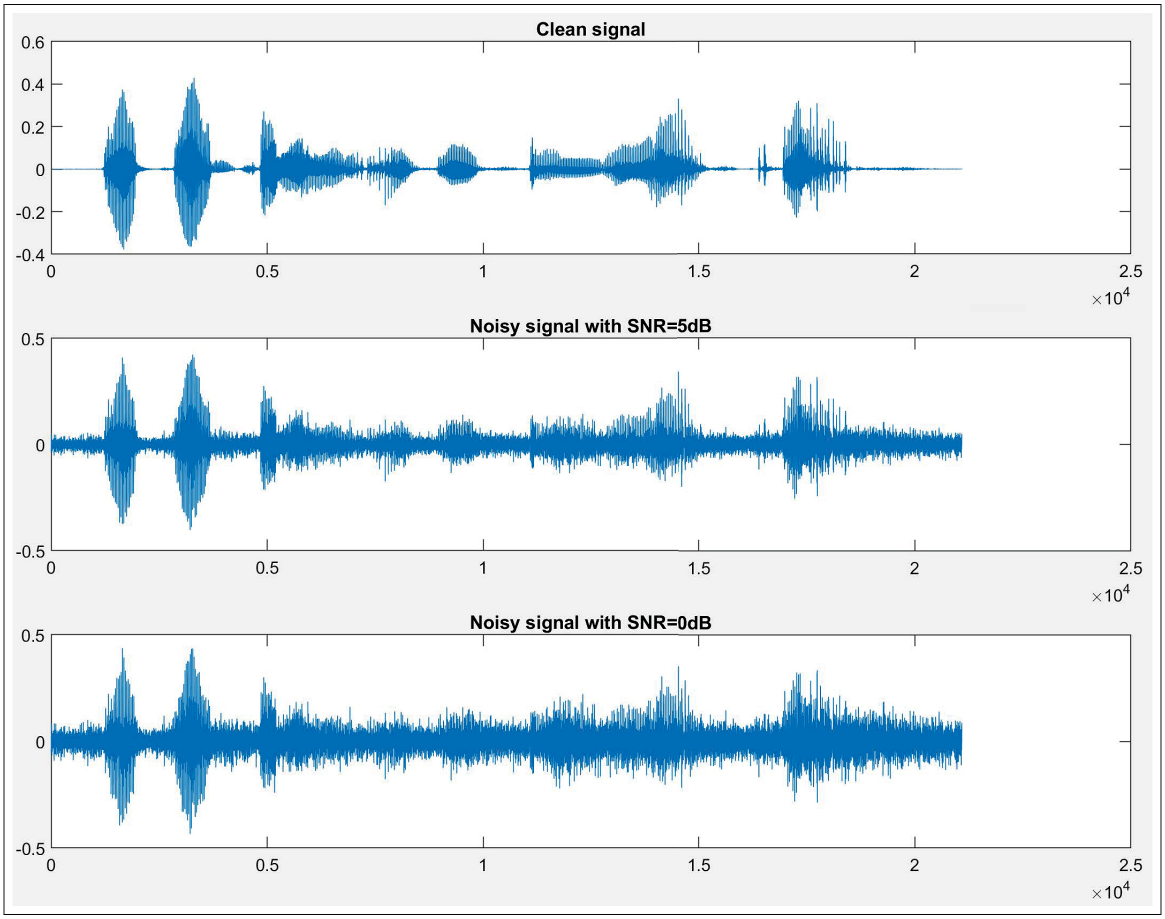

Figure 0.1 Clean and Noisy Speech Signals with Different Levels of Background Noise

distortions. Additionally, they were not as adaptable to varying conditions such as different speakers, languages, and acoustic environments [Taha *et al.* (2018)].

Due to significant advances in the field of Artificial Intelligence (AI) and especially Deep Learning (DL), deep learning-based methods have been very successful recently due to their superior performance in various speech enhancement tasks. These methods employ various neural network architectures, such as feedforward, recurrent, and convolutional neural networks (CNN), to model complex relationships between clean and noisy speech signals. Deep learning-based speech enhancement has seen significant advancements over time. Early research focused on the use of Deep Neural Networks (DNNs) to develop a regression-based speech enhancement framework. This involved employing a multi-layer deep architecture trained on a vast amount of simulated speech data.

Overall, traditional speech enhancement techniques rely on simplistic and often handcrafted assumptions about speech and noise characteristics. These methods can be sensitive to deviations from the assumed noise models, which can lead to artifacts and a reduced ability to adapt to diverse noise environments. In contrast, neural network-based methods, are capable of learning complex, non-linear relationships between clean and noisy speech signals from large-scale training data. This allows them to better capture the underlying structure of speech and noise, resulting in improved speech enhancement performance across a wide range of noise conditions. Moreover, neural networks can be designed to handle non-stationary and highly varying noise sources, further enhancing their robustness and generalization capabilities.

Despite the remarkable effectiveness of more recent neural network-based methods in speech enhancement, these approaches often do not focus on optimizing both speech and non-speech segments within speech signals. Furthermore, these existing methods rely solely on deep learning techniques, neglecting the potential advantages of incorporating classic speech enhancement techniques. In this work, we introduce a novel method to classify speech and non-speech segments of speech signals using neural networks. This method enhances the quality of speech signals by effectively reducing the background noise, leveraging the strengths of both traditional and more recent state-of-the-art neural network-based speech enhancement techniques. Specifically, our approach has significant use cases for speech recognition software that often struggles with noisy speech signals. By significantly reducing background noise and enhancing speech clarity, our method ensures that speech recognition systems can operate more accurately and efficiently, even in challenging acoustic environments. This improvement is crucial for applications where high precision voice commands or dictations are necessary, such as in voice-activated assistants, automated transcription services, and interactive voice response systems. By providing cleaner input signals, our method helps to minimize the error rates of speech recognition algorithms, thereby enhancing user experience and expanding the practical usability of speech-based applications in noisy conditions.

0.2 Problem statement

Classic speech enhancement methods, such as spectral subtraction, Wiener filtering, minimum mean-square error (MMSE) estimation and statistical model-based techniques, have faced several challenges that limited their effectiveness in real-world scenarios. One key challenge was the reliance on certain assumptions about the noise characteristics, such as stationarity or Gaussian distribution, which may not hold true in many real-world situations. This can lead to suboptimal performance when these assumptions are violated. These methods often assumed that the noise was stationary and that its characteristics could be easily estimated, which is rarely the case in real-world environments. Moreover, these methods typically operated on a short-time spectral basis, overlooking important long-term temporal dependencies in speech signals. As a result, their performance often did not reach the desired level when dealing with complex and non-stationary noise sources.

Another challenge classic speech enhancement methods faced was their inability to adapt to different conditions, such as various speakers, languages, and acoustic environments. Traditional methods required extensive fine-tuning and customization to work effectively under diverse circumstances, which made them less practical for all-purpose applications. Furthermore, these methods mainly relied on linear signal processing techniques, which limited their ability to deal with non-linear distortions and led to performance degradation. In summary, classic speech enhancement methods struggled to provide robust and adaptable solutions to noise reduction and speech quality improvement, especially in the presence of non-stationary noise and varying conditions. Therefore, a new approach is required that can overcome the limitations of traditional methods. Artificial Intelligence-based speech enhancement methods can map complex and non-linear relationships between clean speech signals and noisy signals, allowing them to reduce a wide range of noise types effectively. This approach needs to be adaptable to various noise conditions so that it can be used for real-world applications with dynamic and unpredictable noise environments. Neural network-based methods can automatically learn the essential features from the input data and eliminate the need for manually tuning the parameters. Consequently, they can develop efficient classification techniques, enabling them to adapt well to unseen noise types and conditions.

0.3 Research questions

In order to achieve enhancement of speech signals using neural networks with a classic filtering scheme we need to address the following research questions:

- **RQ1.** How can we make our speech enhancement method more adaptable to different noise types and variable levels of signal to noise ratio?
- **RQ2.** How can we make use of the conventional speech enhancement methods used in the past to further improve the performance and accuracy of our proposed method?
- **RQ3.** How can we use features of the speech signal for the classification of speech signals to enhance the accuracy and robustness of neural network-based speech enhancement?

0.4 Objectives

The main objective of the thesis is to design a framework that can fulfill two fundamental tasks of classifying speech and non-speech segments of speech signals and using enhancing techniques based on each type of these segments in order to obtain an improved speech signal as a result. This framework takes into account the different types of noise we may encounter in actual real-world situations to make this method more adaptable to real-world situations. This objective can be divided into three sub-objectives based on the research questions.

- **SO1.** To design and implement a neural network architecture that effectively incorporates the identified features of a speech signal to enhance the accuracy and robustness of the speech enhancement process.
- **SO2.** To develop a comprehensive speech enhancement framework that applies the appropriate enhancing technique to each segment of speech signals using classification with the help of conventional and modern AI-based speech enhancement techniques.
- **SO3.** To implement and test the optimized classification technique within the speech enhancement framework.

0.5 Plan

This thesis consists of an introduction, three main chapters, and a conclusion that highlights possible directions for future research. The organization of the thesis is as follows:

- The Introduction presents the background of the subject of this thesis, emphasizing the significance of speech enhancement by examining prior studies and providing an overview of its development over time. This also includes motivations for this thesis, followed by the goals we want to achieve.
- In Chapter 1, we review the literature on conventional and state-of-the-art speech enhancement methods related to the subject of this thesis.
- Chapter 2 outlines the methodology employed to accomplish the goals of this thesis. Within this chapter, we detail our proposed framework for classifying and enhancement of speech signals.
- In Chapter 3, we present the experimental setup for validating our proposed method, in addition to the evaluation process we have carried out to validate our method followed by a discussion on the results achieved.
- Conclusion summarizes the thesis findings and presents possible future work.

CHAPTER 1

BACKGROUND AND LITERATURE REVIEW

In this chapter, we give the technical background relevant to our work. We proceed with an overview of the historical progression and advancements within the field of speech enhancement. This is followed by literature on two different categories of traditional and state-of-the-art speech enhancement techniques according to three research questions. Finally, we go over techniques in literature that are a hybrid of classic and state-of-the-art techniques that are similar to our proposed method. Regarding our research question, the next section sheds light on the existing solutions for speech enhancement. We investigate the pros and cons of the existing works that lead to a pathway for our work.

1.1 Background

1.1.1 Speech Signal

The speech signal, as a complex and dynamic auditory phenomenon, serves as the basic means of human communication within the bandwidth of 0 to 4 kHz. They are defined as air pressure variations emitted from a speaker with time as its independent variable. These variations in pressure can be described as waves and correspondingly they are often called sound waves. Speech signals convey lots of information using the changes that are made in their frequency and amplitude through time. There are several ways of characterizing the communication potential of speech. Speech signal is a one-dimensional signal with time as its independent variable and it is a highly non-stationary signal since its power spectrum changes over time. However, when closely examined over a short period of time (10–30 milliseconds), its spectral properties remain mostly stationary. The signal waveform is segmented to correspond to different sounds or words spoken. The segments can display varying properties, some being quasi-periodic, others aperiodic, and certain segments even consisting of brief silence gaps. The diversity in duration, intensity, and spectrum of these segments contribute to the uniqueness of each individual's speech pattern [O'Shaughnessy (1987)].

Speech signals get transformed into electrical signals by microphones for speech processing. Analog-to-digital converters turn these analog voltages into digital signals. Bandlimited speech signals, limited to 4000 Hz, can be represented by 8000 samples per second, each quantized to 256 levels (8 bits) with minimal audible degradation. This results in a data rate of 64 kbps for a high-quality speech signal. The goal of speech compression is to reduce this data rate for efficient storage and transmission [Rabiner (1978)].

Even though speech signals can be considered as one-dimensional functions of time, their generation involves an intricate system of parallel nerve commands from the brain that control the articulatory organs, such as the vocal cords, tongue, and lips. This complexity is often ignored or simplified in speech signal compression but has a significant role in more advanced applications like speech recognition and enhancement. The understanding and modeling of these complexities is a critical research area as it contributes to the enhancement of the naturalness and intelligibility of synthesized speech [Rabiner (1978)].

1.1.2 Speech and Non-speech segments of speech signal

Speech signal, the essential part of human communication, consists of two main components: speech and non-speech segments. Speech and non-speech segments are fundamental elements in speech signal processing and speech enhancement. A speech signal's speech segments are the parts of the signal where a person is speaking by producing audible sounds using their vocal cords. Words, phrases, and other vocalizations fall under this category. The various phonetic components of speech are represented by distinct patterns of frequency and amplitude that are characteristic of each segment. Vowels, which frequently have a more periodic or harmonic structure as a result of vocal cord vibration, and voiced consonants, which are created by restricting the vocal tract while allowing for vocal cord vibration, could be examples of such components. On the other hand, non-speech segments represent parts of the signal without speech activity that do not directly involve vocal cord activity. Silence, pauses and background noise are all included in these segments. Silent segments are periods of time during which there is no sound, such as the gaps in sentences or words. Incidental background noise could include environmental sounds that are heard during a conversation. These non-speech segments often feature a noise-like spectral characteristic and a lower periodicity. Effectively enhancing the quality and intelligibility of noisy speech signals involves the separation of these segments, a process that also has applications in other areas such as voice recognition, telecommunications and hearing aids.

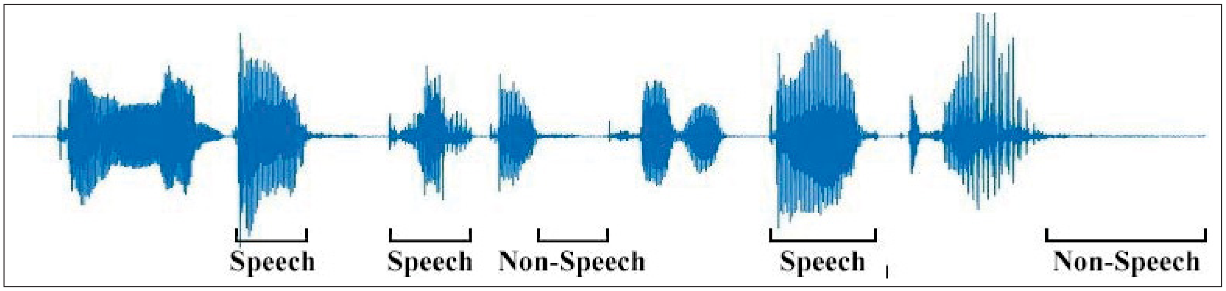

Figure 1.1 Speech and Non-Speech Segments of a Speech Signal

Understanding the distinction between speech and non-speech segments in speech signal processing is critical for effective speech enhancement. When systems can accurately distinguish between the two, they can effectively amplify the speech segments while reducing or eliminating non-speech segments such as background noise or silence. This results in enhanced audio quality and speech intelligibility, which is crucial in various applications such as telecommunications, automated transcription services, conference calls and voice-activated systems.

Similar to how speech and non-speech segments form the broader categories in a speech signal, voiced and unvoiced segments of a speech signal can be viewed as other subsets within the speech signal itself that are used in some other areas of research for speech signal processing. The distinction between voiced and unvoiced segments is determined by the manner in which the sound is produced. Voiced segments, such as vowel sounds and certain consonant sounds (like /m/, /n/, /l/, /r/, etc.), are characterized by regular periodic vibrations of the vocal cords. On the other hand, unvoiced segments, like certain consonant sounds (/s/, /f/, /t/, /k/, etc.), are produced without the vibration of the vocal cords, instead, they arise from turbulent airflow in the vocal tract. Identifying the transitions between voiced and unvoiced segments is crucial for accurate speech recognition and synthesis. The segmentation process often leverages the fundamental differences in the energy and frequency characteristics of these segments by applying different analysis and processing techniques to voiced and unvoiced portions of the signal. Voiced segments of speech are characterized by higher energy levels and a strong harmonic structure due to the periodic vibration of the vocal cords. These segments, including all vowels and some consonants, exhibit clear pitch and a stronger low-frequency emphasis, resulting from the vocal cords' modulation of the airflow from the lungs. Unvoiced segments, in contrast, demonstrate lower energy and lack a clear periodic pattern, reflecting the turbulent airflow used to generate these sounds [Atal (1976)]. To quantify this characteristic difference in energy levels between voiced and unvoiced segments, we employ the energy calculation formula, represented as:

$$E = \sum_{n=0}^{N-1} x(n)^2 \tag{1.1}$$

This equation calculates the energy E of a signal segment. Here, $x(n)$ represents the signal amplitude at a discrete time n, and N is the total number of samples in the segment. The energy is the summation of the squares of the amplitude of the signal over the segment. This can be used to distinguish between voiced and unvoiced segments as voiced segments usually have higher energy levels. Several methods for automatic segmentation have been proposed in the literature, ranging from simple energy and zero-crossing rate-based techniques to more complex methods incorporating machine learning algorithms. These segments' dynamic nature and the transitioning boundaries between them continue to challenge researchers in the field. In general, in addition to voiced and unvoiced segments, the correct identification and analysis of speech and non-speech segments are not only significant for better speech recognition and synthesis but they also offer valuable insights into the development of more advanced enhancement techniques.

1.1.3 Environmental Noise

Environmental noises, also referred to as background noise are unwanted sounds that can interfere with the clarity of speech signals and play a significant role in the deterioration of speech quality and intelligibility. The presence of environmental noises can make it difficult for both human listeners and automated systems to accurately interpret speech content and it can block, distort, modify, or interfere with the meaning of a message. These noises often occur due to external factors and can create a challenge in various communication settings, such as teleconferencing, voice recognition systems, and hearing aids. These noises arise from various sources and can be broadly classified into two categories: stationary and non-stationary noise. Stationary noise remains constant over time, such as the hum of an air conditioner or a computer fan. In contrast, non-stationary noise is characterized by its dynamic nature, with its spectral and temporal properties changing over time. Examples include overlapping conversations in a restaurant which is called babble noise, airport noise and car noise.

In speech signals, various types of noise can be introduced, each with its own unique characteristics. Background noise, for example, refers to external sounds present in the environment, such as traffic, machinery, or conversations with other people. Echo and acoustic echo, on the other hand, occur when sound waves reflect off surfaces, leading to a delayed repetition of the original signal, which can cause confusion and decreased intelligibility. Another type of noise, amplifier noise, arises from the inherent limitations of electronic devices used in the transmission and processing of audio signals. Quantization noise, a byproduct of converting continuous analog signals into digital representations, results from the finite resolution of the digitization process. Lastly, loss of signal quality can occur due to coding and speech compression techniques that aim to reduce the data size of audio files. These techniques can potentially introduce artifacts and degrade the overall clarity of the speech signal. In general, the additive noise model can be mathematically represented as:

$$y(t) = s(t) + n(t) \tag{1.2}$$

This model assumes that the noise, $n(t)$, is added to the original signal, $s(t)$, to produce the observed signal, $y(t)$. Understanding the spectral characteristics of different types of environmental noise is crucial for developing effective speech enhancement techniques. The spectral characteristics of noise provide information about the distribution of noise energy across the frequency domain. For instance, wind noise primarily concentrates in the low-frequency range, typically below 500 Hz, while restaurant noise spans a broader frequency range due to the mix of conversations and other background sounds. By accurately characterizing the spectral properties of environmental noise, we can design targeted speech enhancement algorithms that effectively suppress unwanted noise components while preserving the intelligibility of the speech signal. As the field of speech enhancement continues to evolve, there is an increasing demand for algorithms that can seamlessly adapt to diverse noise profiles. These algorithms should be capable of handling a wide array of noise types that may be encountered across various applications. The adaptability and versatility of these algorithms reinforce their robustness, promising improved performance and wide-ranging applicability in diverse sound environments.

1.1.4 Features of Speech Signal

Speech signals are composed of a wide range of features that make them distinct and capable of conveying a plethora of information. The first key feature is that they are time-varying signals, meaning they change over time in amplitude, pitch, and various other properties. This temporal variability is affected by factors such as the emotional state of the speaker, the specific words being pronounced, and more [Deller Jr (1993)]. Furthermore, speech signals possess quasi-periodic properties, the degree of which is dictated by the phonetic content. Vowels, for instance, tend to display a stronger periodicity compared to other phonetic elements [O'Shaughnessy (1987), pp. 35-37]. The second important characteristic of speech signals is formants, which are the resonant frequencies of the human vocal tract. These are particularly significant in the speech spectrum and are primarily responsible for the tonal qualities of spoken language. The first two formants, F1 and F2, are particularly critical in articulatory phonetics, heavily influencing the perception of vowel sounds. The unique positioning of these formants within the frequency spectrum aids in the identification of vowel sounds [Deller Jr (1999)].

Another critical feature of speech signals is a substantial degree of redundancy. This redundancy exists both within the signal itself and in the linguistic structure of the speech. It refers to the presence of repetitive or predictable elements within the speech, which don't necessarily contribute new information but aid in the intelligibility and robustness of the signal. This feature can be observed on several levels, from the repetition of certain phonemes and syllables to recurring patterns in pitch and rhythm. Despite this, speech signals are highly efficient, carrying extensive information about the speaker, their emotional state, and the speech content [Aylett & Turk (2004)].

Feature extraction from speech signals is a key process in understanding and analysis of speech signals. It involves selecting and extracting key attributes from raw speech data that can be used in further analysis. This helps in minimizing the computational resources needed for accurately representing the data. Therefore, when dealing with large and possibly redundant noise data, the data is transformed into a compact set of features. The selection of features aims to capture relevant information from the original data. The desired operation then uses this condensed representation rather than the complete input set. Feature extraction computes numerical attributes that describe the data, necessary for reducing the dimensionality of the data fed to the neural network. The most common features extracted from speech signals include Mel-frequency cepstral coefficients (MFCCs), pitch, energy, and formant frequencies. These extracted features capture the unique characteristics of the speaker's voice, emotional state, and the content of the speech [Shaw, Vardhan & Saxena (2016)].

$$c_i = \sum_{n=1}^{Nf} S_n \cdot \cos\left[i(n-0.5)\left(\frac{\pi}{Nf}\right)\right], \quad i = 1, 2, ..., L \tag{1.3}$$

The MFCCs, represented by equation 1.3, are a key feature in speech and audio processing, encapsulating the short-term power spectrum of a sound. In this equation, c_i denotes the i^{th}

MFCC coefficient for $i = 1, 2, ..., L$, where L is the total number of coefficients being calculated. The term S_n represents the logarithmic energy output of the nth filter in a filter bank. The equation performs a summation of the product of the logarithmic filter bank energy outputs, S_n, and a cosine term, which facilitates the transformation to the cepstral domain. Here, Nf is the number of filters in the filter bank.

One common practice in feature extraction is to analyze the speech signal in frames. Speech signals are typically divided into short frames (usually 20-30ms), which are then processed separately. This is due to the quasi-stationary nature of speech signals which means they change slowly over time, so a short frame can be treated as if it were stationary. The features of these frames can then be used for a variety of applications other than speech enhancement, including speech recognition, speaker identification, and emotion detection [Smith, Lukasiak & Burnett (2006)].

1.2 Classic Speech Enhancement Techniques

In this section of this chapter, an overview of the historical advancements of speech enhancement techniques will be presented, outlining the evolution of various methods and highlighting key milestones in the field. Spectral subtraction, developed in the 1970s, was one of the earliest speech enhancement techniques. Spectral subtraction is a simple method that involves estimating the noise spectrum and subtracting it from the noisy speech signal. It operates in the frequency domain, aiming to reduce the impact of additive noise on the speech signal. This process is typically performed using a Short-time Fourier Transform (STFT) to convert the time-domain signal into the frequency domain. The technique assumes that the noise and speech signals are statistically independent, with the noise being additive. While spectral subtraction has its limitations, such as residual noise and the potential for musical noise artifacts, it has formed the basis for many subsequent speech enhancement algorithms [Loizou (2013)].

In the 1980s, researchers developed more advanced methods, such as Wiener filtering which Norbert Wiener introduced in the 1940s. It is an optimal linear filtering method in the context of minimum mean square error (MMSE) estimation. The Wiener filter aims to estimate the clean speech signal from the noisy input by minimizing the error between the clean speech and the filtered output. The method depends on the statistical properties of both the speech and noise signals, presuming that they are stationary and that their power spectra can be estimated.

$$H(f) = \frac{P_s(f)}{P_s(f) + P_n(f)} \tag{1.4}$$

Equation 1.4 illustrates the Wiener filter where $H(f)$ is the filter gain, $P_s(f)$ is the power spectral density of the clean speech signal, and $P_n(f)$ is the power spectral density of the noise signal. The emergence of statistical-based methods, such as the MMSE and maximum a posteriori (MAP) estimators started in the 1990s. These methods extend the Wiener filter concept by focusing on estimating the speech signal by leveraging statistical properties, which enable improved noise suppression and speech quality but with increased computational complexity. This approach became influential in the field of speech enhancement as it provided a robust technique for enhancing noisy speech signals in the spectral domain. On the other hand, the MAP technique looks for the clean speech signal that is most likely to be present given the observed noisy speech and any prior knowledge of the clean speech and noise distributions. The equation below is an example of an MMSE estimator in the spectral domain, which is often used in speech enhancement.

$$\hat{X}_k = \frac{\xi_k}{1 + \xi_k} \cdot Y_k \tag{1.5}$$

MMSE estimator of the kth signal spectral component, i.e., the Wiener estimator is represented by $\hat{X}_k$ in the equation 1.5. Here, ξ_k is the a priori SNR at frequency bin k, signifying the ratio of the power of the clean signal to the power of the noise at that frequency bin. Finally, Y_k is the noisy speech spectral component at frequency bin k, representing the combination of the clean speech spectral component and the noise spectral component at that frequency bin [Ephraim & Malah (1984)].

In the early 2000s, researchers began exploring techniques based on empirical mode decomposition (EMD) for speech enhancement. It is a data-driven, adaptive, and non-parametric approach to decomposing a signal into its intrinsic mode functions (IMFs). The EMD technique has been particularly useful in the context of non-stationary and nonlinear signal processing, such as speech signals. The EMD process begins with the identification of the local extrema (maxima and minima) in the speech signal. Then, an upper and lower envelope is formed by connecting the local maxima and minima using a cubic spline interpolation. The mean of the upper and lower envelopes is subtracted from the original signal, resulting in the first IMF. This procedure is repeated on the residual signal until a predetermined stopping criterion is met or the residual signal becomes a monotonic function. Each IMF represents a specific frequency band of the original signal, and these IMFs can be analyzed and processed individually. In the context of speech enhancement, noise can be suppressed by modifying the IMFs, and the enhanced speech signal is obtained by summing the processed IMFs. EMD offers a flexible and adaptive approach that can handle non-stationary and nonlinear speech signals [Khaldi, Boudraa & Turki (2016)].

During the early 1990s and later in 2000s, researchers also investigated the use of wavelet-based techniques for speech enhancement. These techniques are founded on wavelet theory, which provides a versatile and powerful mathematical tool for the analysis and processing of non-stationary signals, such as speech. Wavelet-based methods involve decomposing a noisy speech signal into a set of wavelet coefficients using a wavelet transform, such as the discrete wavelet transform (DWT) or the continuous wavelet transform (CWT). The wavelet transform facilitates the analysis of signals at different frequency bands with different resolutions, enabling a more precise representation of the signal's characteristics. Once the noisy speech signal has been transformed into wavelet coefficients, noise reduction techniques are applied to these coefficients. This may involve thresholding, where coefficients with magnitudes below a certain threshold are set to zero, effectively removing noise components. Alternatively, more advanced techniques like wavelet-domain filtering can be used, in which wavelet coefficients are adjusted based on their statistical properties or a particular noise model. After the noise reduction process, the enhanced speech signal is reconstructed by applying the inverse wavelet transform to the modified wavelet coefficients. Wavelet-based speech enhancement methods have been shown to effectively suppress various types of noise while maintaining the original speech signal's quality [Mohammadi, Zamani, Nasersharif, Rahmani & Akbari (2008)].

In the late 2000s and early 2010s, non-negative matrix factorization (NMF) emerged as a promising speech enhancement technique using machine learning. In this technique, a non-negative input data matrix is split into two non-negative matrices with lower dimensions. To effectively separate speech and noise components from a noisy speech signal, it is important to achieve a sparse and meaningful representation of the original data. In this technique, the noisy speech signal is first represented as a non-negative matrix, often in the time-frequency domain. Then, NMF divides this matrix into two lower-dimensional non-negative matrices: a basis matrix that represents the speech components and a coefficient matrix that represents their activations.

$$V \approx WH \tag{1.6}$$

The equation 1.6 is used in NMF where V is the original non-negative data matrix, W is the basis matrix, and H is the coefficient matrix. The decomposition process minimizes the difference between the original matrix and the product of the two lower-dimensional matrices. The speech and noise components can be separated after collecting the basis and coefficient matrices by applying constraints or by adding additional information. By inverting the time-frequency representation of the estimated speech matrix, this separation enables the reconstruction of the enhanced speech signal. NMF-based speech enhancement techniques have gained popularity due to their ability to capture the structure of both speech and noise signals effectively, resulting in improved performance compared to traditional methods. However, these techniques may require additional tuning and assumptions, depending on the specific application and the characteristics of the noise to be removed [Mohammadiha, Smaragdis & Leijon (2013)].

In addition to matrix factorization methods, researchers have also explored the use of Sparse representation in the 2010s to separate clean speech from noise by exploiting the sparse nature of speech signals in a transformed domain. Using an overcomplete dictionary with more basis elements than input signal dimensions, the goal is to find the sparsest representation of the

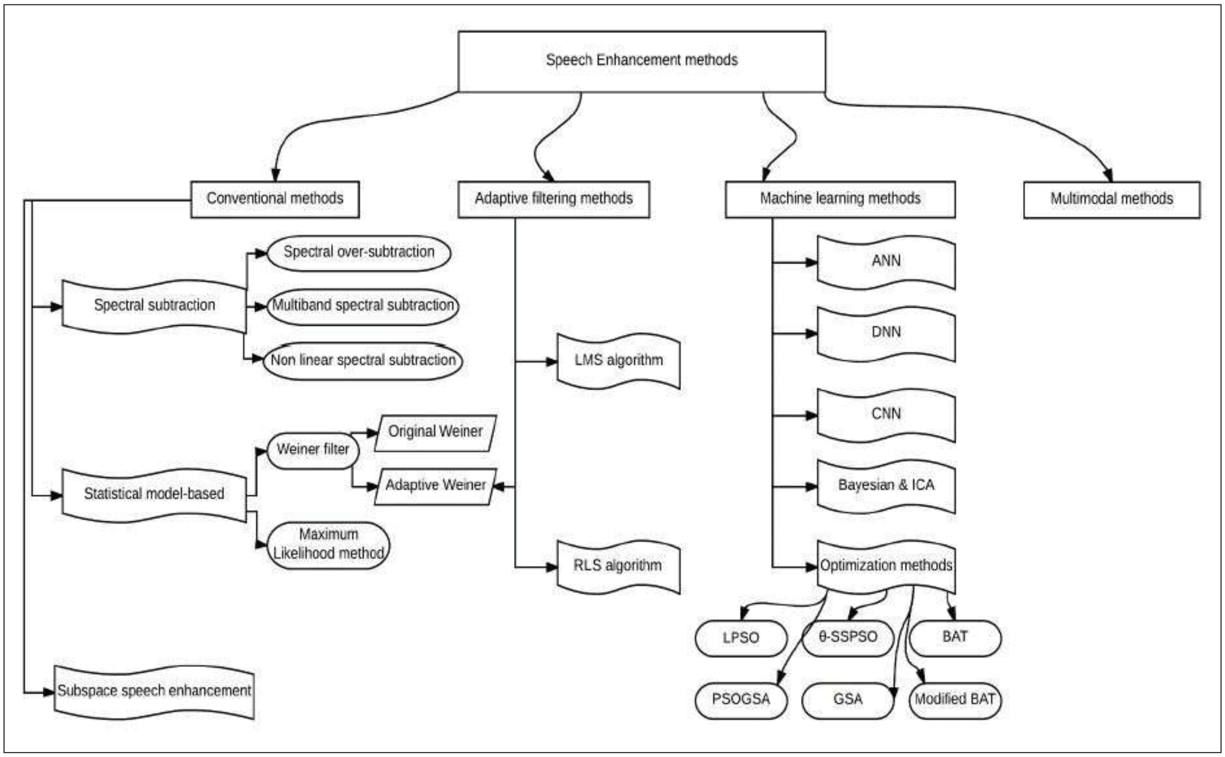

Figure 1.2 An Overview of Speech Enhancement Methods. This figure has been modified from the original to correct a typographical error in the labeling.
Taken from Taha *et al.* (2018)

noisy speech signal. Popular methods include Matching Pursuit (MP), Orthogonal Matching Pursuit (OMP), Basis Pursuit (BP), and K-SVD. These techniques improve speech enhancement performance by effectively separating speech and noise components while preserving essential speech characteristics [Zhao, Xu & Yang (2011)].

1.2.1 Spectral Subtraction

Spectral subtraction is a popular classic technique utilized for enhancing the quality of speech signals by minimizing the effects of additive noise. This technique is based on the principle that the power spectrum of a noisy speech signal is equivalent to the sum of the power spectrums of the noise-free speech and the noise. The method involves estimating the spectrum of the noise and subtracting it from the spectrum of the noisy signal to recover the spectrum of the original, noise-free speech signal. The performance of spectral subtraction relies heavily on an accurate estimation of the noise spectrum [Boll (1979)].

$$Y(\omega_k) = X(\omega_k) + D(\omega_k) \tag{1.7}$$

$$X_e(\omega_k) = \sqrt{|Y(\omega_k)|^2 - |D(\omega_k)|^2}\, e^{j\theta_y} \tag{1.8}$$

The equation 1.7 represents the Short-time Fourier Transform (STFT) of the noisy speech signal. In this equation, $Y(\omega_k)$ denotes the STFT of the noisy speech, $X(\omega_k)$ is the STFT of the clean speech, and $D(\omega_k)$ represents the STFT of the noise component. The variable ω_k represents the frequency component at the k-th index. The equation 1.8 provides the method for estimating the clean speech signal, $X_e(\omega_k)$, by subtracting the noise spectrum from the noisy speech spectrum and retaining the phase of the noisy speech signal, denoted by θ_y. This approach is fundamental in the spectral subtraction method, aiming to isolate the clean speech signal from the background noise [Anusha, Indira & Maheshwari (2022)]. The spectral subtraction approach often yields better results when accurate noise estimation is employed and when phase information is taken into consideration. Typically, a transformation like the Fourier Transform is applied to convert the signal to the frequency domain before applying spectral subtraction, and an inverse transformation is used to revert the processed signal back to the time domain.

In [Scalart *et al.* (1996)] the authors improved upon the traditional spectral subtraction method by proposing a technique known as Ephraim and Malah's MMSE estimator. They formulated a spectral gain function that minimizes the MMSE of the speech spectral amplitude estimator, resulting in a more robust and reliable speech enhancement.

Another methodological advancement is the improved spectral subtraction method proposed by [Kamath, Loizou *et al.* (2002)]. They introduced a generalized spectral subtraction method that incorporates a noise estimation algorithm to reduce musical noise, a common artifact in enhanced speech using spectral subtraction. Colored noise does not affect the speech signal uniformly over the entire spectrum, as some frequencies are affected more than others. Traditional methods of spectral subtraction, which subtract the magnitude spectrum of noise from that of the noisy speech, often result in an annoying distortion in the speech signal called musical noise. To overcome this, the authors propose a multi-band approach that divides the speech spectrum into multiple non-overlapping bands and performs spectral subtraction independently in each band. This takes into account the fact that colored noise affects the speech spectrum differently at various frequencies. Their method has shown improved performance, particularly in non-stationary noise environments.

The spectral subtraction technique, despite being a long-established method, continues to inspire newer research for noise reduction and speech enhancement. The recent advancements in this domain, as seen in these studies, prove that the integration of traditional techniques like spectral subtraction with modern, sophisticated methodologies can yield significant improvements in the quality of speech signals.

1.2.2 Speech and Non-speech classification

In the field of speech processing, speech and non-speech classification refers to the distinction between segments of audio that contain human speech and those that do not. The primary objective is to distinguish between speech and non-speech signals, such as music, environmental sounds, or silence. This task is particularly critical in various applications, including Automatic Speech Recognition (ASR), speech enhancement, and audio indexing. However, it poses a challenging problem due to the wide range of non-speech sounds and the overlap in their acoustic characteristics with speech signals. Different approaches have been proposed to tackle this problem effectively. Traditional speech and non-speech classification techniques for speech enhancement often rely on energy-based measures, which serve as the primary parameter for differentiating between speech and non-speech segments. The energy threshold is frequently calculated using Signal-to-Noise Ratio (SNR), where transitions between different states are guided by the energy of individual frames in conjunction with specific duration constraints

[Huang, Acero, Hon & Reddy (2001)].

$$\text{SNR (dB)} = 10\log_{10}\left(\frac{P_s}{P_n}\right) \tag{1.9}$$

Equation 1.9 represents the Signal-to-Noise Ratio in decibel (dB) scale. P_s is the power of the signal, and P_n is the power of the noise. The SNR in dB is obtained by taking ten times the logarithm to the base 10 of the SNR in linear scale.

Many systems add variables like pitch or entropy to the energy parameter to increase robustness. Different additional parameters that can be used in conjunction with or without energy have been tested by researchers. From distance measurements to data fusion methods like classification and regression trees, there are many ways to combine these factors [Martin, Charlet & Mauuary (2001)]. [Ajmera, McCowan & Bourlard (2003)] showcased the incorporation of entropy as a feature for speech/non-speech discrimination. Their approach employs an Artificial Neural Network (ANN) specifically trained on clean speech. The output of this ANN, when processed, provides entropy and dynamism measurements at regular intervals. These features are then incorporated into a 2-state Hidden Markov Model (HMM) that differentiates between speech and non-speech segments. The novelty lies in using posterior probability-based features (entropy and dynamism) from the ANN to guide the HMM.

[Martin *et al.* (2001)] present a new method for detecting speech vs non-speech for speech recognition systems. The technique involves using Linear Discriminant Analysis (LDA) applied to MFCC to enhance the detection process. This approach aims to improve the typical energy-based detection systems, which often falsely identify too many noise segments as speech.

In [Shin, Lee, Lee & Lee (2000)], the authors introduced a novel speech/non-speech classification methodology aimed at enhancing the endpoint detection performance in noisy environments for speech recognition, specifically for applications like voice dialing in cellular phones. This methodology utilizes multiple features including full-band energy, band energy in different frequency ranges, peakyness, LPC residual energy, and noise-filtered energy to improve robustness in noisy environments. The classification and regression tree (CART) technique is applied to efficiently combine these various features for the classification of each frame.

[Kwon & Lee (2003)] present a novel method for designing speech/non-speech classifiers using the adaptive boosting (AdaBoost) algorithm, primarily for voice activity detection and robust endpoint detection. The approach integrates simple base classifiers through the AdaBoost algorithm and optimizes speech features in conjunction with spectral subtraction, maintaining simplicity in implementation and low computational complexity. The AdaBoost classifier and spectral subtraction significantly improve the receiver operating characteristic curves of the G.729 voice activity detector, and for speech recognition purposes, the method reduces miss errors by 20-50% for the same false alarm rate.

[Shafiee, Almasganj & Jafari (2008)] introduced a novel speech/non-speech detection system that incorporates fractal dimension and prosodic features along with commonly used MFCC. By applying a Voice Activity Detector (VAD) to segment the audio signals, the system evaluates the periodicity of speech signals (prosodic features) and the level of distortion (fractal dimension). Prosodic features are captured by applying the autocorrelation function to high-energy segments, while fractal dimension features are extracted using the Petrosian method and Normalized Attractor Fractal Dimension technique. The extracted features are further optimized through a feature selection process using a genetic algorithm. System performance was assessed using neural network and Support Vector Machine (SVM) classifiers on the TIMIT speech database.

[Thambi, Sreekumar, Kumar & Raj (2014)] proposed an enhanced Speech/Non-Speech Detection (SND) system using the Random Forest decision tree algorithm to better discriminate between speech and non-speech segments in audio and video documents. The SND system relies on features extracted from the time, frequency, and cepstral domains for 20 ms frames, along with their mean and standard deviation for segments of 200 ms. The system's performance was improved by refining the selection of features using correlation-based feature selection and by smoothing the decisions over five 200 ms segments. The proposed approach resulted in an increased classification accuracy of 97.80% using only eight features, up from a baseline accuracy of 94.45% using 272 features, suggesting a highly effective method for SND.

It's evident that the application of machine learning techniques has greatly enhanced the ability to discriminate between speech and non-speech segments, paving the way for improvements in the overall performance of speech enhancement systems. [Maganti, Motlicek & Gatica-Perez (2007)] presents a novel method for unsupervised speech/non-speech detection critical for Automatic Speech Recognition (ASR) in meeting rooms. It proposes an algorithm that analyzes the long-term modulation spectrum and inspects specific frequency ranges for dominant speech components to classify speech and non-speech signals. The method works on a short-segment basis, thus delivering near real-time performance. It has been tested against other techniques, such as manual segmentation, short-term energy analysis, zero-crossing based segmentation, and Multilayer Perceptron (MLP) classifier systems. The results indicate that the proposed approach is not only robust and accurate but also less sensitive to the mode of signal acquisition and varying signal-to-noise ratios.

Furthermore, CNNs, renowned for their success in image classification, have also been adapted for speech and non-speech classification, with models such as the one proposed in [Thomas, Ganapathy, Saon & Soltau (2014)] showing promising results. They proposed a method for improving Speech Activity Detection (SAD) in mismatched acoustic conditions using CNNs. The CNN models were trained on known radio channels and then adapted with supervised data from unseen channels to tackle performance degradation. The adapted models showed significant improvements over conventional Deep Neural Networks (DNNs), demonstrating the potential of CNNs to quickly adapt to novel acoustic conditions.

1.2.2.1 Voice Activity Detection

The Speech and Non-Speech Classification is a broader category that includes Voice Activity Detection (VAD) which is a technology used in speech processing where the presence or absence of human speech is detected in an audio segment. While speech/non-speech classification can also identify other types of sounds and it discriminates between speech and all other types of non-speech audio such as silence, music, environmental noise, etc., VAD specifically identifies whether human speech is present or not. This separation can be achieved using various features of the speech signal and various statistical and machine learning methods. Moreover, in many cases names of these two techniques are used interchangeably based on the application and context of the research done. The VAD can benefit many speech-based applications such as voice-controlled assistants, telecommunication systems, automated transcription services, speech and speaker recognition, speech enhancement, emotion recognition, and dominant speaker identification.

One approach that has been effective in voice activity detection is the use of statistical features of the speech signal. In [Ramırez, Segura, Benıtez, De La Torre & Rubio (2004)] an algorithm is designed to measure the Long-term Spectral Divergence (LTSD) between speech and noise, comparing the long-term spectral envelope to the average noise spectrum to inform the speech/non-speech decision rule. The decision threshold adjusts to the measured noise energy, and a controlled hang-over is activated only when the observed signal-to-noise ratio is low.

Another statistical approach for VAD is the work done by [Ghosh, Tsiartas & Narayanan (2010)] which proposes a novel Long-term Signal Variability (LTSV) measure for VAD. Unlike traditional VAD methods which rely on short-term analysis, this new approach utilizes long-term non-stationarity characteristics of the speech signal to discern speech from noise. The LTSV measure captures the degree of signal non-stationarity over an extended period, differentiating between speech, which is highly variable, and noise, which is less so. The proposed method evaluates its performance under different noise types and SNR conditions, demonstrating superior accuracy compared to traditional methods, particularly under low SNR conditions. It's worth noting that the LTSV-based VAD doesn't require explicit SNR estimation, which simplifies the process, but might introduce a delay equal to the duration of the longer window length, depending on its implementation in specific applications.

Machine learning techniques have also been introduced into VAD for improved performance. [Obuchi (2016)] proposed a new VAD algorithm that leverages both augmented statistical noise suppression (ASNS) and a CNN for improved accuracy, particularly in noisy environments. Unlike traditional VAD methods that combine feature extraction and classification, the proposed algorithm prioritizes a noise suppression scheme, then finds an optimal classifier for the noise-suppressed signal. The process involves a sequence of ASNS, framewise speech/non-speech classification using CNNs, and inter-frame smoothing. The algorithm is further enhanced through noise adaptive training, where the classifier is trained on noisy speech data and then subjected to the noise suppression module.

Deep learning techniques have also been utilized for VAD to improve performance. [Zhang & Wu (2013)] developed a DNN based VAD that uses a denoising deep neural network (DDNN)-based voice activity detection that seeks to improve on the deep-belief-network (DBN)-based VAD's lack of distinct benefits from deeper layers. The DDNN-based VAD involves pre-training a deep neural network in a unique, unsupervised, denoising layer-by-layer manner and then fine-tuning the entire network using a common back-propagation algorithm. This process helps to discover the manifold of clean speech while mitigating background noise interference. In comparison to the DBN-based VAD, the DDNN approach demonstrates greater performance and effectiveness of the deeper layers.

The work done by [Eyben, Weninger, Squartini & Schuller (2013)] proposes a novel data-driven approach to VAD using Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) trained on standard RASTA-PLP frontend features. To create a real-life simulation, the system is trained with noisy speech instances from the TIMIT and Buckeye corpora and diverse types of real long-term noise recordings. The performance of this approach is evaluated with synthetically mixed test data and a real-life test set consisting of four full-length Hollywood movies, outperforming three state-of-the-art reference algorithms under the same conditions. The study emphasizes the advantage of LSTM-RNNs due to their ability to model long-range dependencies, learning when to access which parts of past context, thus improving the robustness in real-life and noisy settings.

As one of the recent advancements in VAD, the method proposed by [Zhang & Wang (2015)] presents a significant step forward. By leveraging the use of DNNs and contextual information, the method can provide more accurate and reliable detection of voice activity. They propose a new method to improve VAD in low SNR environments, using machine learning techniques to support contextual information across three levels. At the top level, a multi-resolution stacking (MRS) approach, which is a stack of ensemble classifiers, is used. Each classifier takes the combination of predictions from its preceding blocks and the expansion of the raw acoustic feature by a given window. A boosted deep neural network (bDNN) serves as the base classifier for MRS at the middle level. bDNN generates multiple base predictions from different contexts of a single frame using one DNN, which are then aggregated for a better prediction. At the bottom level, a multi-resolution cochleagram feature is employed to incorporate contextual information at multiple spectrotemporal resolutions.

1.3 Artificial Neural Networks

The Artificial Neural Network (ANN) is also called a Multilayer Perceptron (MLP). The multilayer perceptron consists of three primary components: the input layer, the hidden layer, and the output layer. The input layer, also known as the visible layer, receives data from the dataset and passes it on to the next layer without any modifications. Following the input layer are one or more hidden layers, which are not directly exposed to the input data. The final hidden layer serves as the output layer, responsible for generating a vector of values specific to the task at hand. Each layer in an MLP is followed by a nonlinear activation function, which sets it apart from linear models and enables it to process input data in a nonlinear manner. Every neuron in a higher layer connects to all neurons in the lower layer.

$$net = \sum_{i=1}^{n} (w_i \cdot x_i) + b \tag{1.10}$$

In equation 1.10, the weighted sum, denoted as net, in an artificial neuron is calculated as the summation of the product of each input value, x_i, and its corresponding weight, w_i, plus a bias, b, for all inputs from 1 to n. This equation quantifies the combined influence of the input values and their respective weights, adjusted by the bias, on the neuron before the activation function is applied. The neuron's responsibility is to aggregate the incoming information, and the bias allows the neuron to have some flexibility in fitting the target output during the training process.

$$\text{output} = f(\text{net}) \tag{1.11}$$

Once the weighted sum, net, is computed, it is passed through an activation function, denoted as f, to compute the output of the neuron. The activation function introduces non-linearity to the model, allowing the neural network to learn complex patterns. The type of activation function depends on the problem at hand, and it influences the network's ability to converge and learn the underlying patterns in the data. The output, denoted as output, is the final value emitted by the neuron, and it serves as the input to the subsequent layer in a multilayer neural network. It is this output that is used for computing the error and updating the weights during the training process, optimizing the network to make accurate predictions.

The core functionality of an MLP lies in its ability to learn complex patterns and representations from input data through a process known as forward propagation. During forward propagation, the input data is passed through the network and transformed at each layer by applying an activation function to the weighted sum of the previous layer's outputs. The activation function, such as the sigmoid or rectifier function, introduces non-linearity into the model, allowing it to learn intricate relationships within the data. Once the data has traversed the entire network, the output layer generates the final predictions. During the training process, the network employs a technique called backpropagation, which calculates the gradient of the error with respect to each weight by moving backward through the network. This gradient information is then used to update the weights, optimizing the network's performance and reducing prediction errors.

The training of MLP can be performed using various algorithms, each having its nuances, advantages, and specific applications. Gradient Descent and its variants are fundamental techniques where Batch Gradient Descent uses the whole dataset to compute the gradient of the cost function, stochastic gradient descent updates the weights after processing each training example, and Mini-Batch Gradient Descent is a compromise between the two. Momentum is a refinement of these techniques, incorporating past gradients in weight updates to accelerate convergence. Adaptive learning rate methods like Adagrad, RMSprop, and Adam modify learning rates during training, adapting differently for distinct features and combining the principles of Momentum and RMSprop to maintain adaptive learning rates. Second-order Optimization Methods like Newton's Method and BFGS use second-order derivatives or approximations to find the minima.

Delving deeper into optimization strategies, the Scaled Conjugate Gradient (SCG) Backpropagation is an advanced training algorithm that converges faster and is more efficient compared to standard gradient descent algorithms. The SCG algorithm combines the conjugate gradient method with a scaling factor to optimize the performance of the neural network by accelerating the convergence and reducing the risk of overshooting the global minimum. The training of ANNs often involves the optimization of a cost function to tune the model parameters or weights. Among various optimization algorithms, SCG Backpropagation algorithm is recognized for its efficiency and effectiveness in training ANNs, particularly when dealing with large-scale problems. The SCG algorithm is based on the Conjugate Gradient method but incorporates enhancements to avoid the explicit computation of the Hessian matrix, thus offering improved performance. By employing SCG backpropagation, the network efficiently computes the error gradients and updates the weights, leading to a more accurate and reliable model. This advanced optimization technique ensures that the MLP not only learns intricate representations and complex relationships within the data but also does so with enhanced computational efficiency and stability during the training phase [Møller (1993)].

1.3.1 Supervised Learning Techniques

Supervised learning techniques for speech enhancement aim to improve the quality and intelligibility of speech signals using a dataset of clean and noisy speech pairs for training. These techniques train models using labeled data which means they require both noisy speech samples and clean speech samples. By learning the mappings from noisy to clean speech, these models can subsequently enhance the quality of unseen noisy speech signals. Supervised learning can leverage a wide array of models, including linear methods such as least mean squares (LMS) and non-linear methods such as DNNs, CNNs and Recurrent Neural Networks (RNNs). Deep learning-based supervised methods have shown particularly impressive results in speech enhancement. DNNs, for instance, can model complex, non-linear relationships between noisy and clean speech, and can be trained to minimize a variety of objective functions, allowing for task-specific optimization. Similarly, CNNs have demonstrated efficacy in leveraging local correlations in spectral features, enabling effective enhancement in the time-frequency domain. RNNs, on the other hand, are particularly suitable for handling sequential data, making them capable of capturing temporal dependencies in speech signals over time. These deep learning architectures can be trained end-to-end, effectively learning the underlying data distributions and achieving state-of-the-art results in many speech enhancement tasks.

[Xu, Du, Dai & Lee (2014)] proposed a supervised approach to enhance speech quality in noisy conditions by establishing a mapping function between noisy and clean speech signals using DNNs. Rather than relying on traditional MMSE-based noise reduction techniques, this approach employs a DNN as a nonlinear regression model, trained with a large dataset encompassing a wide range of speech and noise combinations. This method is further improved by applying global variance equalization to reduce the over-smoothing problem often encountered in regression models, and incorporating dropout and noise-aware training strategies to enhance the DNNs' generalization to unencountered noise conditions. The proposed approach demonstrated superior performance over the MMSE-based method, effectively suppressing non-stationary noise, and proved effective in handling real-world noisy data without producing musical artifacts typical in conventional enhancement methods.

[Park & Lee (2016)] proposed a novel method to improve speech intelligibility in the presence of babble noise, particularly in the context of hearing aids. The authors develop a 'mapping' between noisy speech spectra and clean speech spectra through supervised learning. The proposed network, named Redundant Convolutional Encoder Decoder (R-CED), leverages the advantages of Fully Convolutional Neural Networks which have fewer parameters than fully connected networks, hence reducing the model size. The R-CED architecture operates by extracting redundant representations of a noisy spectrum in the encoder phase and maps it back to a clean spectrum in the decoder phase.

One such technique that involves DNNs is proposed by [Xu, Du, Dai & Lee (2013)]. This technique uses regression models within the DNN to learn the intricate mapping function from noisy to clean speech. To do this, it takes into account various key factors in noisy speech, such as speakers, noise types, and SNRs. A collection of stereo data is used to train the model, in which noisy and clean speech pairs are represented by log-power spectral features. The DNN, once trained, is used to enhance noisy speech signals. The paper also employs a pre-training method using restricted Boltzmann machines (RBMs) to mitigate the issues of poor local minima found in randomly initialized networks. Furthermore, the method employs a MMSE objective function to fine-tune the DNN and improve the model's ability to reduce noise in speech signals.

[Zhao, Xu, Giri & Zhang (2018)] present a novel method of speech enhancement designed to suppress noise and improve speech intelligibility. The proposed method differs from existing techniques by incorporating a perceptual speech intelligibility metric, the Short-Time Objective Intelligibility (STOI) measure into the loss function, in addition to the traditional Mean Squared Error. This perceptually guided approach is anticipated to align more closely with the purpose of speech enhancement, which is improving comprehension in noisy environments. The research systematically demonstrates improvements in speech intelligibility across various noise types and signal-to-noise ratios, maintaining speech quality.

[Liang, Kong, Xie, Tang & Cheng (2020)] proposed a real-time speech enhancement algorithm that utilizes an attention-gated Long Short-Term Memory (LSTM) model. This algorithm emulates human auditory characteristics by segmenting the frequency band based on the Bark scale, from which Bark Frequency Cepstral Coefficients (BFCCs), their derivative features, and pitch-based features are extracted. The algorithm incorporates an attention mechanism to filter out information that is less contaminated by noise, aiding in clean speech reconstruction. Additionally, the ideal ratio mask (IRM) with inter-channel correlation (ICC) is used as the learning target, enabling adaptive reallocation of the power ratio between speech and noise. The algorithm also employs a multiobjective learning strategy to optimize the networks jointly using a VAD.

Neural networks are essential in improving the efficiency and accuracy of supervised learning techniques for speech and non-speech classification. The performance of supervised learning techniques for speech and non-speech classification often depends on the features extracted from the audio data. Audio features such as Mel-frequency cepstral coefficients (MFCCs), zero crossing rate (ZCR), pitch and energy, among others, have been widely used. In addition, advanced techniques like deep learning have enabled the automatic extraction and learning of relevant features directly from raw audio data. Despite the significant progress in this area, challenges remain due to the variability in speech patterns, dialects, and noise. Continued research is focused on improving the robustness and versatility of these supervised learning models to handle diverse real-world scenarios.

The proposed approach in [Shafiee *et al.* (2008)] combines fractal dimension features, prosodic features, and commonly used MFCC features to distinguish between speech and non-speech signals. A VAD is initially applied to segment high-energy portions in the input signal. These segments are further processed using two types of prosodic features and two types of fractal dimension features. The performance of the system was evaluated using neural network and SVM classifiers on the TIMIT speech database.

In [Ryant, Liberman & Yuan (2013)], the authors investigated a novel approach for Speech Activity Detection (SAD) using DNNs, specifically for complex environments like YouTube where diverse environmental conditions are present. Traditional methods, such as those using Gaussian Mixture Models (GMMs), have limitations in these conditions. To address this, the authors explore the potential of DNNs to learn robust features during training, with multiple frames of MFCCs serving as the input. This approach drastically reduces frame-wise error rates (down to 19.6%) when applied to YouTube videos compared to conventional GMMbased systems (with 40% error rates), demonstrating the potential of DNNs in improving the performance of SAD, especially in diverse and noisy environments.

1.3.2 Hybrid Methods

Hybrid methods that merge traditional speech enhancement techniques with advanced neural network architectures are gaining prominence in the field of speech enhancement. By capitalizing on the strengths of both traditional methods and modern AI-based techniques, these hybrid methods aim to deliver more robust and effective solutions for enhancing the quality of noisy speech signals. Traditional methods provide well-understood, interpretable behavior that handles specific scenarios or noise types effectively, while ANN-based methods excel at learning complex patterns from data and generalizing across a broader range of noise types. The integration of traditional methods can enhance the system's robustness, offering more predictable and stable behavior, particularly in situations where ANN models might struggle with unseen data. Furthermore, the explicit modeling of noise or speech signals in traditional methods aids interpretability, beneficial for system understanding, debugging, or improvement. Efficiency can also be improved, as traditional methods may require less computational resources or data for training compared to ANN-based methods. Lastly, the inclusion of traditional methods, which often do not require labeled data, can potentially reduce the amount of labeled data needed for the system, addressing a common challenge in machine learning. However, integrating these two types of methods also poses challenges, such as determining the most effective ways to combine and train these systems. Despite the remarkable advancements in hybrid speech enhancement methods, there are still open research opportunities in this field, particularly concerning the integration of more sophisticated neural network architectures with classical speech enhancement techniques.

The authors of [Yang & Bao (2018)] proposed a novel approach for speech enhancement that combines DNN and AR-Wiener filtering. The DNN is trained to predict Autoregressive (AR) model parameters of both speech and noise, simultaneously. These predictions are then used to construct a Wiener filter for speech enhancement. The DNN is trained by minimizing the Euclidean distance between its output and the AR model parameters of clean speech and noise. During operation, acoustic features are extracted from the noisy speech and input to the DNN, which then estimates the AR model parameters for both speech and noise. The method also introduces the use of speech-presence probability (SPP) to address residual noise issues. Similarly, my proposed method also employs a classification task to identify regions of the signal that need to be enhanced. However, there is a key difference: our proposed method focuses on speech/non-speech classification and applies enhancement techniques based on this categorization. In contrast, the authors' DNN method predicts AR model parameters for both speech and noise, using these estimates to construct an AR-Wiener filter for the final speech enhancement.

In the paper written by [Yu, Ouyang, Zhu, Champagne & Ji (2019)], the authors propose a novel algorithm for speech enhancement that combines the use of DNN with Kalman Filter (KF). The method, referred to as DNN-KF, employs a DNN trained on a large database to estimate the Linear Prediction Coefficients (LPCs) from noisy speech signals. These LPCs, critical for the implementation of the KF, are then utilized in a KF for time-domain speech enhancement. The proposed algorithm operates by enhancing both the magnitude and phase of the speech, and due to the robust learning ability of DNNs, it provides a more accurate estimation of clean speech's LPCs. Similar to the authors' method, our proposed approach also combines classic speech enhancement techniques with state-of-the-art neural network-based methodologies to improve speech signal quality. However, there's a significant difference in the sequential arrangement of these techniques. In the authors' method, the neural network is applied first for speech enhancement, followed by the utilization of a classic speech enhancement technique, namely the Kalman Filter, in the subsequent stage. In contrast, our proposed method applies both the classic and state-of-the-art techniques in parallel at the same time to the speech signal in order to get the final enhanced signal. In speech enhancement, the Kalman filter can be used to estimate the clean speech signal from the noisy speech signal.

$$\hat{x}_{k|k-1} = F_k \hat{x}_{k-1|k-1} + B_k u_k \tag{1.12}$$

$$P_{k|k-1} = F_k P_{k-1|k-1} F_k^T + Q_k \tag{1.13}$$

$$K_k = P_{k|k-1} H_k^T (H_k P_{k|k-1} H_k^T + R_k)^{-1} \tag{1.14}$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k (z_k - H_k \hat{x}_{k|k-1}) \tag{1.15}$$

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} \tag{1.16}$$

The Kalman filter consists of prediction in equations 1.12 and 1.13, and update phases in equations 1.14, 1.15, and 1.16. In the prediction phase, the state estimate $\hat{x}_{k|k-1}$ is computed using the state transition model F_k, the control input u_k with the control-input model B_k, and the previous state estimate $\hat{x}_{k-1|k-1}$. The prediction error covariance $P_{k|k-1}$ is also computed using the previous error covariance $P_{k-1|k-1}$, the process noise covariance Q_k, and the state transition model F_k.

In the update phase, the Kalman gain K_k is computed using the predicted error covariance $P_{k|k-1}$, the measurement model H_k, and the measurement noise covariance R_k. The state estimate is then updated with the Kalman gain weighted difference between the actual measurement z_k and the predicted measurement $H_k\hat{x}_{k|k-1}$. Finally, the error covariance $P_{k|k}$ is updated using the Kalman gain K_k and the measurement model H_k. In the context of speech enhancement, the state can represent features of the speech signal, the control input can represent any interventions or manipulations applied to the signal, and the measurements represent the observed noisy signal. The Kalman filter then aims to estimate the true speech signal features from the noisy observations, providing an enhanced speech signal.

[Lin *et al.* (2019)] present a novel speech enhancement method, S-ForkGAN, that combines Generative Adversarial Networks (GANs) and spectral subtraction techniques. Unlike traditional GAN-based methods that operate on time-domain waveforms, the proposed method works with

Log-Power Spectra (LPS), reducing computational complexity. This method uses a forked GAN structure to simultaneously extract both speech and noise information from the input signal. The resultant data is then processed through spectral subtraction to reconstruct the enhanced speech signal. The S-ForkGAN performance is evaluated through automatic speech recognition using the TIMIT dataset under various noise conditions, demonstrating superior performance to existing GAN-based methods while maintaining lower computational complexity. Traditional speech enhancement methods have long-lasting value and can be effectively integrated with newer techniques to improve performance and robustness. These foundational elements often add stability to more recent, computationally intensive methods. In the authors' S-ForkGAN method, After extracting LPS features from a noisy speech input and using two decoders to estimate the speech and noise patterns, spectral subtraction is used to subtract the estimated noise from the original noisy signal to recover the speech. This process matches our own proposed method, particularly in the application of spectral subtraction, emphasizing the relevance of traditional techniques in modern algorithms.

In [Dash & Solanki (2019)], they propose a novel hybrid method to improve speech quality and intelligibility, especially in adverse environments with high ambient noise. This approach leverages a modified DNN trained specifically for the speech signal at hand, combined with Adaptive Multi-band Spectral Subtraction (AdMBSS) to enhance the intelligibility of the speech signal. Additional phase information calculation is utilized within the AdMBSS process, and a hybrid DNN alongside the Nelder Mead optimization technique are employed to boost signal quality. Their experimental results suggest that this innovative framework offers improved performance across different measures, including signal-to-noise ratio, perceptual evaluation of signal quality, and minimum mean square error, even under varied noise conditions like bus, train, babble, airport, station, and exhibition noise.

In the work by [He, Tian, Yu, Chang & Xiong (2022)] a novel speech enhancement methodology named PHASEN-SS is proposed. The method combines a DNN for initial speech enhancement with traditional spectral subtraction for data post-processing. The DNN operates using a two-branch communication system that separately predicts amplitude masks and phase, improving the prediction accuracy by exchanging information between the two branches. Post-DNN processing then denoises the residual noise using spectral subtraction. In a more recent paper, the authors of [Salehi & Mirzakuchaki (2022)] proposed an innovative hybrid method combining a MMSE approach with a long short-term memory fully convolutional network (LSTM-FCN). The MMSE approach has been acknowledged for its high performance in speech enhancement tasks, but it fails to accurately estimate non-stationary noise sources. This problem is tackled by utilizing an LSTM-FCN to accurately estimate a priori SNR, thus improving the performance of the MMSE approach. The proposed MMSE approach, aided by the LSTM-FCN estimator, does not make assumptions about the characteristics of the noise or the speech. Similarly, our proposed method also aims to enhance speech signal quality by distinguishing clean speech segments from noisy ones. However, there are some crucial differences. Our method employs a classification scheme focused on speech and non-speech segments and applies enhancement techniques based on this categorization, specifically using conventional techniques like spectral subtraction. In contrast, their method leverages the power of LSTM-FCN to estimate a priori SNR, which in turn improves the performance of its conventional MMSE approach.

Hybrid methods in speech enhancement, like the one presented in this thesis, are proving to be a promising avenue for future research. The combination of state-of-the-art neural network models with traditional speech enhancement techniques allows for more accurate estimation and better performance. The continuous developments in these areas are creating a solid foundation for advanced speech enhancement techniques that can handle a wider range of noisy environments with better results. These methods bring together the reliability of classical approaches and the flexibility of modern techniques, resulting in a noticeable improvement in noise reduction and speech clarity. By integrating these methods, the occurrence of processing artifacts is often reduced, leading to a cleaner listening experience. However, the complexity of implementing and optimizing these systems can be a challenge. They tend to require more computational power and involve detailed parameter tuning to achieve the desired results. Despite these challenges, the benefits they offer in creating more intelligible and natural-sounding speech make them a valuable focus for ongoing research and development.

CHAPTER 2

METHODOLOGY

2.1 Overview

Usually, speech signals have underlying noise, which may be due to the signal acquisition process or due to the nature of the channel through which they are sent or transmitted. In this sense, the main objective of our project has been to improve speech signals, using a framework based on machine learning using artificial neural networks.

According to the previous chapters, we know that speech signals are composed of speech and non-speech segments, Accordingly, we have developed a framework designed to accomplish two primary objectives. The first task is the classification of signal segments into speech and non-speech categories. The second task is the enhancement of speech signals, the ultimate goal being to generate an improved speech signal in terms of perceptive quality and intelligibility. Thus, our first task has been addressed through the training of a neural network (classifier), for the classification of the speech and non-speech signals present in the speech signals. To achieve this, we have implemented MATLAB functions for dataset preparation, feature extraction, labeling of the speech and non-speech signals, and training and evaluating our neural networks. As a result, we successfully implemented a comprehensive framework for the classification of speech and non-speech signals.

On the other hand, the second task consisted of improving the speech signal, starting from the classified speech and non-speech segments (windowed signals). Then, we performed a specific treatment depending on the type of classified window (speech and non-speech). When a window is classified as a non-speech window, the enhancement process lowers the entire window's amplitude to a minimal value. This acts as an indirect denoising step. Non-noisy non-speech signals typically have amplitudes near zero. Consequently, setting the amplitude of these signals close to zero effectively cleans the signal.

As for speech windows, the treatment is different, since they carry relevant information that we need. Consequently, when a window signal is classified as a speech signal we apply the spectral subtraction speech enhancement technique, which eliminates the underlying noise and improves the quality of the speech signals. The above represents a classical signal enhancement scheme that has proven itself to be a reliable method that has been continuously used and improved over the years. Our results have shown that the classification scheme of speech and non-speech signals together with our cleaning strategy of these signals have been successful, obtaining a really improved speech signal in terms of SNR and speech intelligibility.

2.2 Preparation of the Training Data

In principle, before performing the training of our classifier, we must prepare the training data, i.e. we must define the ground truth with which our neural networks classifier will learn to classify speech and non-speech signal segments. In this sense, the preparation of the training dataset will consist of systematically applying the following steps:

- Windowing of the signals used for training
- Extraction of features on which the neural networks classifier will be trained
- Labeling of windowed signals as speech and non-speech

It is important to highlight that from our dataset we have used a set of noisy and clean speech signals as training signals since for obvious reasons, in practical situations the input signal to our classifier will be noisy audio signals. Therefore, the classifier must be trained from noisy audio signals, so that it learns to classify segments (windowed signals) in situations where the audio signal carries underlying noise. Our preparation step starts by reading all the clean speech signal files and concatenating all of them to one single clean signal. Furthermore, we do the same for associated noisy versions of our clean speech signals and after specifying our desired SNR in the code, the concatenation is done resulting in a single noisy signal which is now ready for our further training steps.

2.2.1 Dataset

The NOIZEUS dataset was used in our work, serving as the training set for our neural network models. Created with the purpose of facilitating the comparison of speech enhancement algorithms among research teams, NOIZEUS comprises a collection of noisy speech corpus that has proven to be valuable in many research done in the field of speech recognition and speech enhancement. [Loizou (2007)]

NOIZEUS is characterized by its versatility and comprehensiveness. It contains 30 IEEE sentences spoken by three male and three female speakers, tainted by eight different real-world noises at varying SNRs. The noise signals were added to the speech signals at SNRs of 0dB, 5dB, 10dB, and 15dB. The recorded sentences were originally sampled at 25 kHz and downsampled to 8 kHz. The noise elements which are extracted from the AURORA database range consist of Babble (crowd of people), Car, Exhibition hall, Restaurant, Street, Airport, Train station and Train, offering a wide variety of sound conditions. By training our model on this extensive variety of conditions, we can significantly improve its robustness and adaptability. The varying SNRs also allow for a thorough examination of our model's performance under different noise levels. [Hu (2007)]

An additional noteworthy feature of the NOIZEUS dataset is its strict adherence to real-world conditions. The speech and noise signals have been meticulously filtered by the modified Intermediate Reference System (IRS) filters, simulating the receiving frequency characteristics of telephone handsets. Furthermore, the noise is not just arbitrarily added; the dataset creators have taken careful steps to ensure the realism of the noise application, such as scaling the noise segment to match the desired SNR level before adding it to the filtered clean speech signal. This attention to detail ensures the practical applicability of the findings derived from the use of this dataset in the current study. [Hu (2007)]

2.2.2 Windowing

The concept of windowing is an essential technique in digital signal processing, often employed when analyzing and processing finite segments of an infinite signal. Essentially, it involves taking a small window of the signal for individual examination and processing. By dividing the infinite signal into manageable segments or windows, one can analyze specific characteristics within that window. This methodology is advantageous when processing signals that are too large to be analyzed as a whole or when the signal characteristics are expected to vary over time.

In our work, the windowing process plays a crucial role in preparing the training data. We apply this process to noisy speech signals, much like the method explained in the previous sections. The application of windowing in our context allows us to handle large speech signals and extract critical features by focusing on smaller, more manageable segments.

Our windowing process involves the use of the function ’clip signal’. As part of this function, we specify the noisy speech signal and the window size (referred to as the ’clip’) as input arguments. In this particular case, the clip size is set to 500 samples. Executing the clip signal function with these parameters generates what we term a ’clipped matrix’. In this matrix, each column corresponds to a window signal comprising 500 samples. This way, our infinite signal gets segmented into individual windows, each carrying unique information, ready for subsequent processing stages.

2.2.3 Feature Extraction

After we have segmented the combined audio and noise signals into window signals, these signals are neatly arranged in the clipped matrix. The next step in our process involves the extraction of features from these window signals. Each of these window signals is subjected to the computation of its features, and the features obtained will be used to train our ANN.

The method employed for this purpose is a function that has the duty of getting the features of windowed signals. When this function is executed on our data, it results in a feature matrix. The dimensions of this matrix are 5×(number of windows), meaning that we have five features for each window signal. This rich, feature-oriented representation of our window signals forms the basis for our subsequent ANN training process.

The features that we have used for training our neural network model are zero-crossing rate, power, maximum value, standard deviation and the Root Mean Square (RMS) of the signal. These elements are critical for understanding and extracting more information from speech signals. The zero-crossing rate of a signal is a significant feature, frequently utilized in pattern recognition and signal processing. This feature refers to the number of times the signal changes from positive to negative or vice versa. This feature is significant in distinguishing between different types of audio segments. For instance, in speech signals, the number of zero crossings can distinguish between voiced and unvoiced speech segments, as voiced segments tend to have fewer zero crossings than unvoiced ones.

$$Z_c = \frac{1}{N-1} \sum_{n=1}^{N-1} |\operatorname{sgn}(x[n]) - \operatorname{sgn}(x[n-1])| \tag{2.1}$$

This equation computes the zero-crossing rate, Z_c, of a signal, where N is the total number of samples in a window, and $x[n]$ is the signal value at sample n. The power of a signal is another vital feature that is often associated with the overall energy of the signal. Power, in this context, is typically computed as the squared value of the signal amplitude averaged over time. This feature provides a measure of the energy content within the signal and is particularly valuable in telecommunications and signal processing for efficient transmission and noise reduction.

$$P_{\text{avg}} = \frac{1}{N} \sum_{n=1}^{N} |x(n)|^2 \tag{2.2}$$

The equation 2.2 represents the average power P_{avg} of a signal for a discrete signal $x(n)$ with N samples. This formula denotes that the average power is the mean of the squared magnitude of the signal over all its samples. Another feature to consider is the maximum value of the signal. It provides information about the peak amplitude of the signal, where the maximum amplitude can provide insights into the volume and intensity of the sound. For a discrete signal $x(n)$ with

N samples, the maximum value $x_{\max}$ can be mathematically represented as:

$$x_{\max} = \max_{1 \le n \le N} x(n) \tag{2.3}$$

In equation 2.3 $x_{\max}$ seeks the highest amplitude value within the entire discrete signal and n ranges over all indices from 1 to N. The standard deviation of a signal is an additional essential feature, providing insights into the variability of a signal. As an indication of the dispersion or spread of signal values, the standard deviation can help identify the stability and consistency of a signal. In audio signal processing, a high standard deviation could indicate a diverse range of frequencies present, implying a complex piece of music or noise. For the same signal $x(n)$, the standard deviation σ can be defined as:

$$\sigma = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x(n) - \mu)^2} \tag{2.4}$$

Where μ represents the mean of the signal and n ranges from 1 to N. The standard deviation, as shown in equation 2.4, quantifies the amount of variation or dispersion of the signal values around the mean.

Lastly, the Root Mean Square value of a signal is a mathematical tool that provides a measure of the signal's magnitude independent of its polarity. It is a crucial metric as it considers both the signal's duration and its amplitude. It is especially relevant in the analysis of alternating current (AC) signals and vibration analysis in the field of condition monitoring and fault detection. For the discrete signal $x(n)$, RMS value RMS_x can be defined as:

$$\mathrm{RMS}_x = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x(n)^2} \tag{2.5}$$

Where n ranges from 1 to N. The RMS value, given by equation 2.5, provides a measure of the magnitude of the signal. It represents the square root of the arithmetic mean of the squares of the signal values, offering a useful description of the signal's overall energy.

2.2.4 Labeling of windowed signals

Labeling is an integral part of training data preparation when developing machine learning models, especially in the case of supervised learning. It involves identifying and marking data based on its characteristics or properties. Accurate labeling of data is crucial as it allows the model to learn from these labels, enabling it to make accurate predictions or classifications when exposed to new, unseen data. In the context of our study, labeling helps differentiate between speech and non-speech window signals.

The final phase of our training data preparation involves this labeling process. We categorize the window signals into speech and non-speech signals. The basis for this categorization is the clean, original speech signals. We chose this reference point because it is devoid of noise-induced alterations or variability, making it straightforward to assign speech or non-speech labels to the window signals.

We have designed a function specifically for this labeling task. The function accepts the clipped window signal matrix as input. It then labels each window signal as either speech or non-speech based on a comparison with a predetermined threshold. In line with this function, we assign speech signals with a label ’0’ and non-speech signals with a label ’1’.

The effectiveness of the labeling function depends heavily on the selection of an appropriate threshold. In our case, we have established the threshold, denoted as ’th’, through a tuning process. Initially, we calculate the standard deviation (std) of the samples from a non-speech window, which typically has very small amplitudes. We then compute the std of a window signal that resembles speech but lacks the high amplitudes characteristic of an authentic speech window signal. Taking an average of these std values, we arrived at our threshold ’th’ value, which stands at 0.01. The process underscores the importance of accurate threshold setting in ensuring the correct labeling of window signals.

2.3 Speech and Non-speech signal classifier

In principle, we will use a classifier based on Machine Learning, which will have the function of classifying the different segments of the audio signal as speech and non-speech. In this sense, our structure for the ANN classifier is shown in Fig. 2.1.

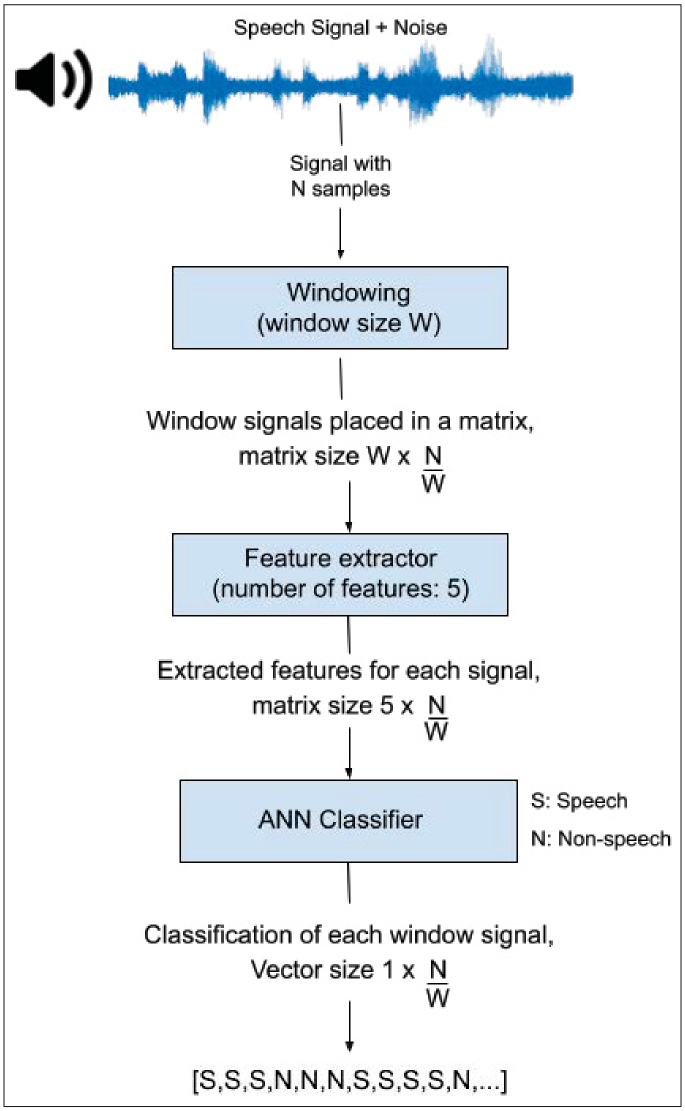
Speech Signal + Noise
Signal with N samples
Windowing (window size W)
Window signals placed in a matrix, matrix size W x $\frac{N}{W}$
Feature extractor (number of features: 5)
Extracted features for each signal, matrix size 5 x $\frac{N}{W}$
ANN Classifier
S: Speech
N: Non-speech
Classification of each window signal, Vector size 1 x $\frac{N}{W}$
[S,S,S,N,N,N,S,S,S,S,N,...]

Figure 2.1 Classifier's Structure

According to Fig. 2.1, the noisy signal will be processed first by a window block. Therefore, the signal will be split by windows through a windowing process, with which we apply the classification in each segment of the audio signal.

As we can see in Fig. 2.2, the noisy signal has N samples, so the windowing process will consist of splitting the audio signal into windows of size W, where W is the size of the window. In this way, we will have N/W windowed signals, which we have conveniently placed in a matrix of size W×N/W, each column of this matrix being a window signal.

After the audio signal windowing process, the clipped window signal matrix goes through a feature extraction process. The utility of this block is to compute, for each window signal present in the clipped matrix, the following metrics which we discussed earlier:

- Zero crossings
- RMS value
- std value
- Maximum value
- Power of the signal

To achieve this, we implemented a function that estimates each of the metrics listed above and is responsible for getting the features of the signal. Moreover, the argument of this function is the matrix of clipped window signals. For its part, the output of this function will be a matrix of features of dimension 5×N/W, since each window signal is summarized by the 5 features mentioned before.

Finally, as we can see in Fig. 2.1, the features matrix is evaluated on the ANN classifier. Naturally, the utility of this block is to classify each window signal as speech or non-speech. Therefore, the output of our ANN classifier will be a vector of dimension 1×N/W, that is, for each window signal we will have a classification label depending on the type of window signal Speech (S) or Non-speech (N). Here is an example of what the output of our classifier will look like:

[S,S,S,N,N,N,S,S,S,S,N,...]

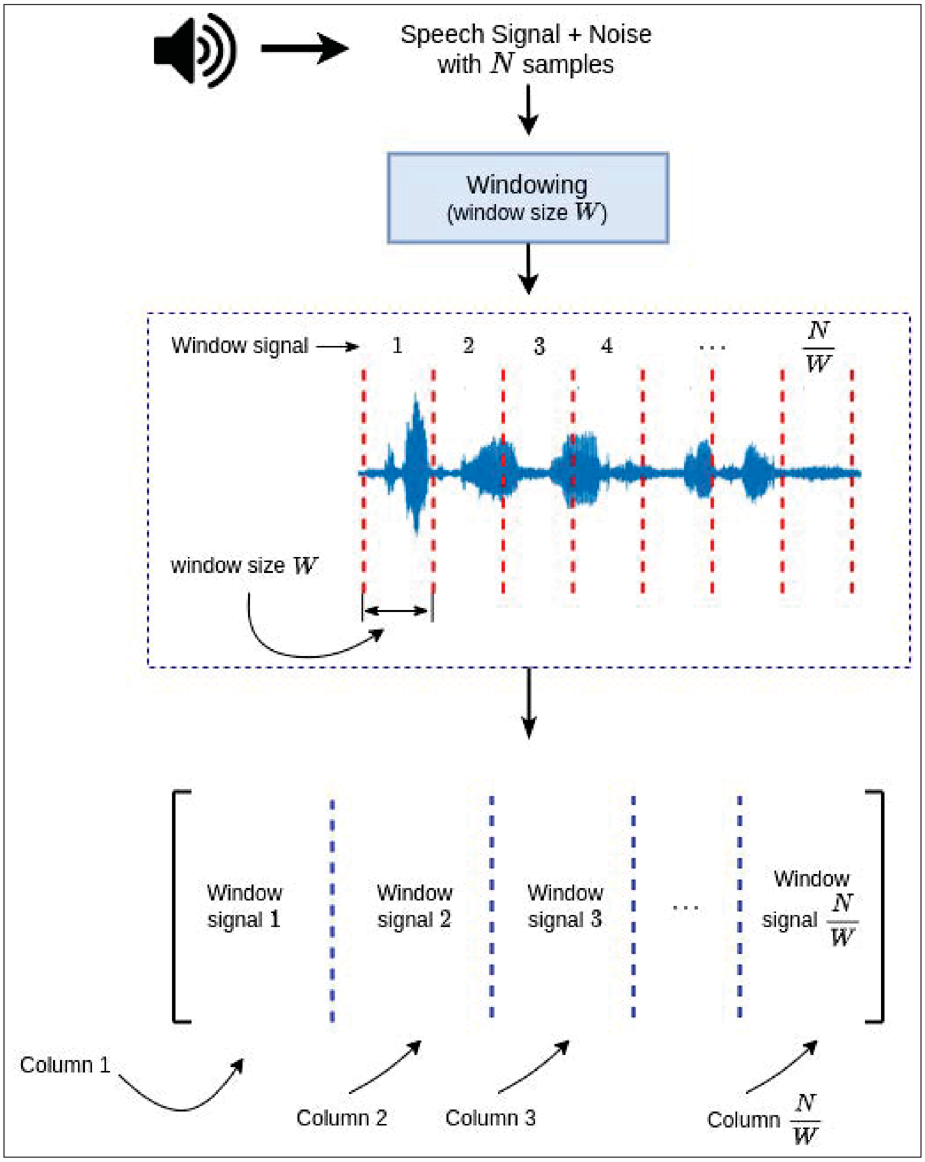

Figure 2.2 Windowing Process

Everything explained above describes the structure of our speech and non-speech classifier in audio signals. However, before using the described structure we must train our ANN classifier.

2.4 Training of classifier model

Once the training data has been prepared, it is time to move on to the training stage of our neural network. For this purpose, we have used MATLAB's Deep Learning Toolbox, which offers us a whole range of MATLAB's functions specially designed to build, train and evaluate any kind of neural network. Thus, our training process starts with splitting the dataset stage, which consists of taking appropriate amounts of signals for training, validation and testing. Usually, in ML/ANN literature it is common to use 70% of the samples as training samples and 30% for validation and testing (15% for validation and 15% for testing). It is important to highlight that the amount of training window signals is one of the critical parameters for the training of our neural network. However, the percentages previously established have been set according to the standards used in the field of signal processing using Machine Learning and Artificial Neural Networks.

Algorithm 2.1, provides a detailed procedure for the program which is responsible for training our neural network models for classification. The algorithm initiates by defining several key parameters and inputs such as $clean\_folder$ and $noisy\_folder$, which presumably store clean and noisy audio signals, respectively. This is followed by desired values for SNR which can be set by the user for setting the noisy signal's SNR and a clipping parameter which we used the value 500. The initialization steps, 1 to 4, set up the environment for populating variables like $all\_clean\_signals$ and $all\_noisy\_signals$ using the audio data from these respective folders. These will later serve as the foundation for training and validating our neural network models.

The "Preparation of Input Data" segment (steps 5 to 13) focuses on setting up the environment by preparing our input variables and feature extraction. In steps 5 to 6, for each audio signal in the $clean\_folder$, a corresponding noisy signal in the $noisy\_folder$ is updated with the specified SNR set by the user using a function which we will discuss later in equation 2.6 to 2.10. These signals are then concatenated into arrays $all\_clean\_signals$ and $all\_noisy\_signals$ (steps 7 to 8). Moreover, in steps 10 to 11, the concatenated signals are used for the windowing process and we store these windowed signals into $clean\_signal\_windows$ and $noisy\_signal\_windows$ for

Algorithm 2.1 Training of Neural Network Models

```
Input: $clean\_folder, noisy\_folder, SNR, clip$
     $all\_clean\_signals, all\_noisy\_signals$
Output: NN_MODEL
     Initialisation:
 1: $clean\_folder \leftarrow 'clean\ signal1.wav', clean\ signal2.wav', ...$
 2: $noisy\_folder \leftarrow 'noisy\ signal1.wav','\ noisy\ signal2.wav', ...$
 3: $SNR \leftarrow Desired\,Value$
 4: $clip \leftarrow 500$
     Preparation of Input Data:
 5: for $i = 1$ to [size of clean_folder] do
 6:    $noisy\_signal = Update\_noisy\_signal(noisy\_signal, SNR)$
 7:    $all\_clean\_signals = Concat(clean\_folder)$
 8:    $all\_noisy\_signals = Concat(noisy\_folder)$
 9: end for
10: $clean\_signal\_windows = Window$(all_clean_signals, clip)
11: $noisy\_signal\_windows = Window$(all_noisy_signals, clip)
12: $features$ = Get_features($noisy\_signal\_windows$)
13: $labels$ = Get_labels($clean\_signal\_windows, 0.01$)
     Training:
14: $layer\_size \leftarrow '10'$
15: Divide training data to 70% for training, 15% for validation and 15% for testing
16: $NN\_model$ = Train(net, features, labels)
17: Save NN_model with the selected SNR and clip values.
```

later use. Subsequently, in step 12, features are extracted from the noisy signal windows, and as we mentioned before features could be statistical properties of the signals like zero crossings, RMS value, standard deviation, and so on, as indicated by the $Get\_features$ function. In step 13, labels are derived from the windowed clean signals to label parts of the signal as speech or non-speech using a certain threshold which was 0.01 in our case as we discussed in previous sections. The derived labels serve as ground truth for our neural network model and are used to evaluate its performance on the validation set during the training process.

Finally, the "Training" section in steps 14 to 17 outlines the process of training the neural network model. The size of the hidden layer for the neural network is set in our training process (step 14). In the context of neural networks, the hidden layer is the layer between the input layer and the output layer and the value of 10 sets the size of the hidden layer to 10 neurons. The size of the hidden layer is an important hyperparameter in the design of neural networks and can impact the model's performance. A smaller hidden layer may lead to underfitting, where the model may not capture the complexity of the data. A larger hidden layer may lead to overfitting, where the model captures noise in the data and does not generalize well to new, unseen data. Continuing from the specifics of our neural network training, we also adopted the Scaled Conjugate Gradient (SCG) as our optimization algorithm. This choice was informed by SCG's computational efficiency and reduced memory requirements, which are beneficial given our large dataset and the complex nature of our neural network. Moreover, SCG automates the tuning of certain hyperparameters, like learning rates, making the training process more straightforward. Thus, similar to our careful selection of the hidden layer size, the use of SCG contributes to creating a model that is both effective and efficient. In step 15, the data is divided into training, validation, and testing sets. Using the features and labels prepared earlier, the neural network model (NN_model) is trained (step 16). The trained model is then saved, incorporating the selected SNR and clip values in the filename of the file in order for later use during the enhancement process (step 17).

In step 6, we discussed a function used for updating and modifying the SNR of our noisy signals which is also going to be useful in our enhancement algorithm. This function proves especially advantageous considering that the NOIZEUS database did not contain noisy signals with an SNR of -5dB. The utilization of this function enabled us to successfully generate noisy variants with an SNR of -5dB for all signals within the diverse environments of the database during the training and enhancement phases. Below are the corresponding sets of equations for this process.

$$\text{noise} = \text{noisySignal} - \text{originalSignal} \tag{2.6}$$

$$E_{\text{originalSignal}} = \sum_{i=1}^{N} |\text{originalSignal}(i)|^2 \tag{2.7}$$

$$E_{\text{noise}} = \sum_{i=1}^{N} |\text{noise}(i)|^2 \tag{2.8}$$

$$\text{noise} = \text{noise} \times \sqrt{\frac{E_{\text{originalSignal}}}{10^{\frac{\text{SNR}}{10}} \times E_{\text{noise}}}} \tag{2.9}$$

$$\text{noisySignal} = \text{originalSignal} + \text{noise} \tag{2.10}$$

In the provided set of equations, noise is the initial noise computed by subtracting the cleanSignal from the noisySignal. The energy of the cleanSignal, denoted as $E_{\text{cleanSignal}}$, and the energy of the noise, denoted as E_{noise}, are calculated as the sum of the squares of the absolute values of their respective samples as we discussed in previous chapter. Subsequently, the noise is adjusted to satisfy the desired SNR by multiplying it with the square root of the fraction of $E_{\text{cleanSignal}}$ to the product of $10^{\frac{\text{SNR}}{10}}$ and E_{noise}, aligning the noise energy to the intended SNR level relative to the clean signal energy. Finally, the adjusted noisySignal is obtained by adding the adjusted noise to the cleanSignal. In conclusion, this covers the entire process we used for training our models, giving a detailed guide that outlines every step from start to finish.

Fig. 2.3 shows a structured ANN model as we discussed earlier in the training algorithm. This model is organized into three layers: an Input Layer, consisting of five neurons, each representing a specific feature extracted from the signal data, including Zero-crossing rate, RMS value, Standard deviation, Maximum value, and Average power of the signal; a Hidden Layer, consisted of ten neurons, responsible for identifying and learning the complex patterns and representations within the input data, where each neuron receives a weighted input from every neuron in the input layer; and an Output Layer, containing two neurons, indicative of the application of

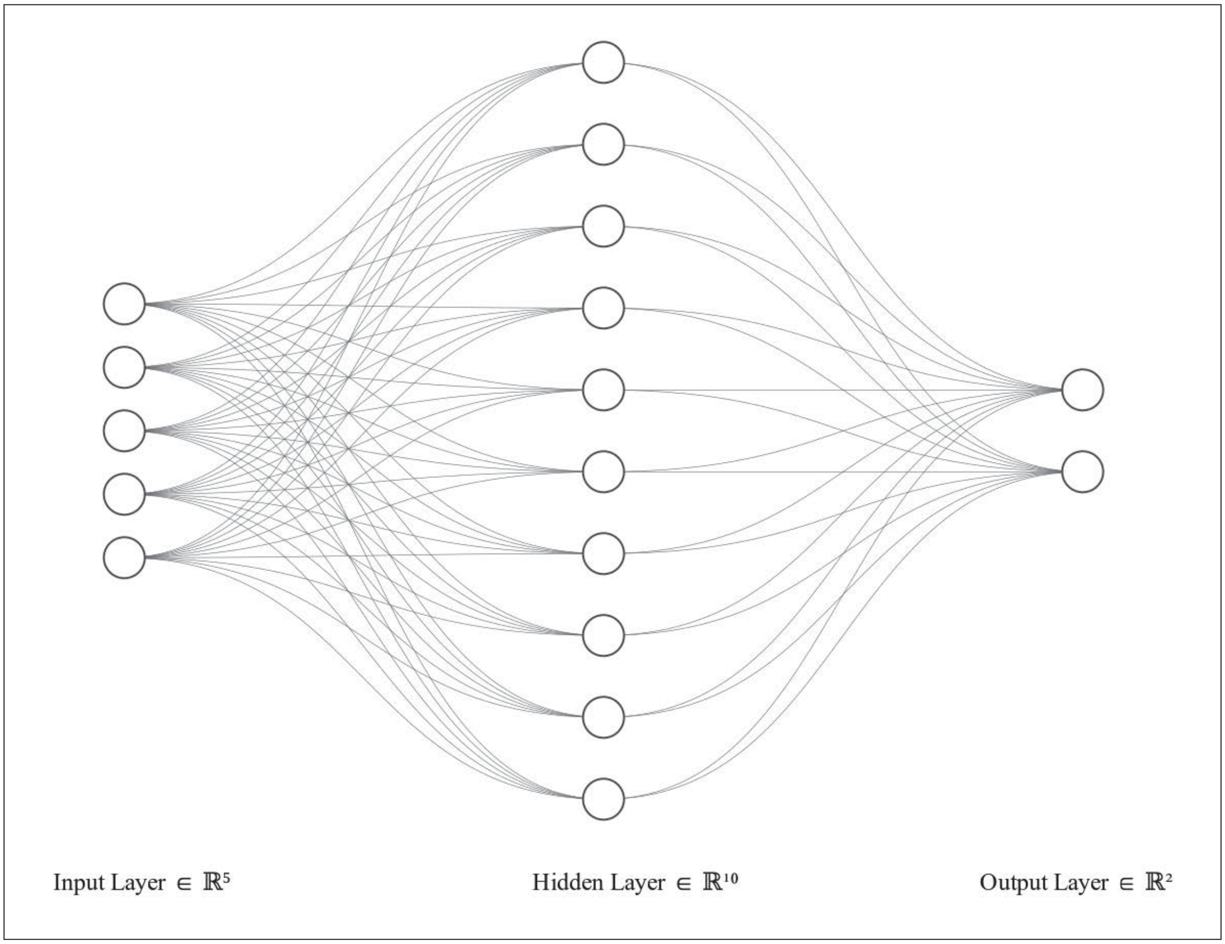
Input Layer $\in \mathbb{R}^5$
Hidden Layer $\in \mathbb{R}^{10}$
Output Layer $\in \mathbb{R}^2$

Figure 2.3 Illustration of the Neural Network Model for the Proposed Method

one-hot encoded labels, tasked with generating the final output of the network, representing the probabilities corresponding to the two classes of speech and non-speech.

The training of this network is executed using the Scaled Conjugate Gradient backpropagation algorithm which we discussed in the previous chapter. The design and configuration of this network enable it to efficiently understand and learn the complex relationships between the input features, thereby facilitating accurate predictions or classifications. Throughout the training process, the network continuously modifies the weights of the connections to minimize prediction error, aiming to optimize the model's performance on unseen or new data.

2.5 Enhancement

At this point, we know that our classifier will classify each window signal as speech or non-speech. With this as a premise, the speech enhancement will consist of applying a specific treatment depending on whether the window signal is speech or non-speech, as illustrated in Fig. 2.4.

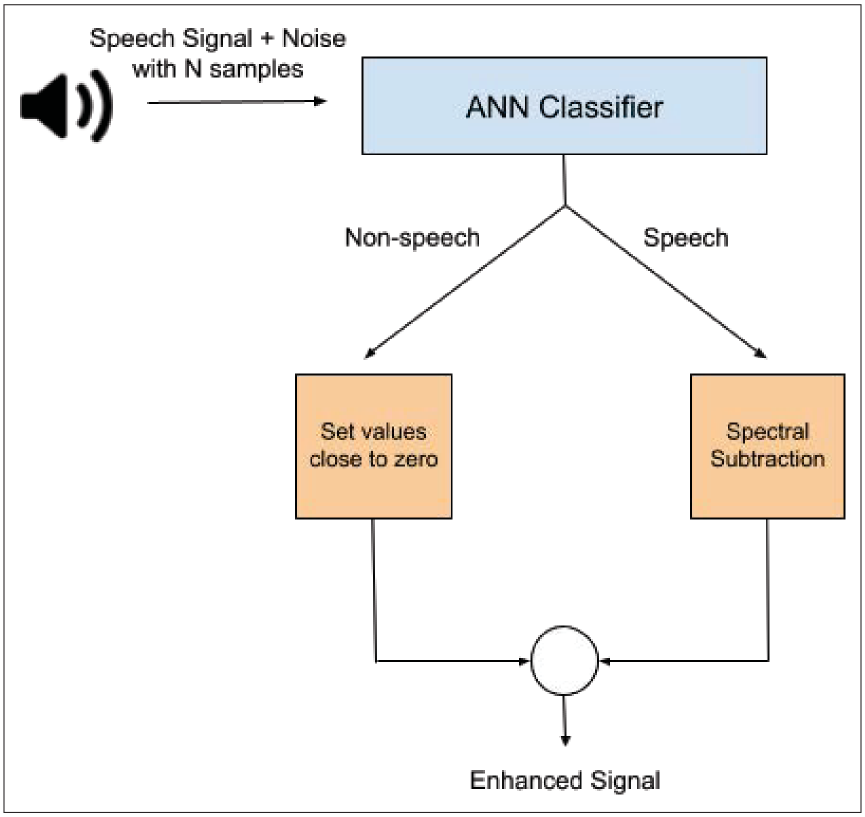

Figure 2.4 Enhancement Process

According to Fig. 2.4, once the window signal is classified, it goes to an enhancement block, which can be a block where all samples are set to a low value or a block where spectral subtraction is applied. In this sense, when we are in the presence of a window signal classified as non-speech, the samples of the entire window are set close to zero which is achieved by multiplying them by 0.4. This value was determined after experimentation and tweaking to optimize the results. Through this approach, we perform an enhancement on the non-speech window signals. Typically, these non-speech window signals have amplitudes near zero in the absence of noise. By resetting these segments closer to zero, we are inherently applying a denoising procedure to the non-speech window signals, enhancing the clarity and reducing ambient noise within these segments.

In the enhancement process, before scaling down the amplitudes of non-speech segments, a refined approach involving fade-in and fade-out techniques has been integrated into the non-speech segments to further refine the enhancement process. These implemented techniques are crucial in establishing smooth transitions between different segments, and they act as a buffer to mitigate any abrupt changes, ensuring the natural flow and coherence of the signal are maintained throughout the enhancement process. The gradual reduction and introduction of amplitudes in non-speech segments through fade-out and fade-in, are essential in preserving the integrity and the auditory quality of the signal, making the listening experience more fluid and less tiring, especially in environments with huge amounts of background noise or interference.

This transition technique is essential for mitigating abrupt changes and ensuring the continuity and coherence of the signal, effectively preserving its auditory quality, especially in noisy environments. These techniques also contribute to a smoother, more natural listening experience, allowing listeners to perceive speech with enhanced clarity and less effort. The focus on the transitions between speech and non-speech segments is crucial, introducing a layer of denoising that improves the perceptual quality and intelligibility of the audio signal, making the overall speech enhancement methodology more efficient and effective.

$$y(t) = x(t) \cdot \left(0.4 + 0.6 \cdot \left(1 - \frac{t}{T}\right)\right) \tag{2.11}$$

$$y(t) = x(t) \cdot \left(0.4 + 0.6 \cdot \left(\frac{t}{T}\right)\right) \tag{2.12}$$

The fade-out and fade-in effects are mathematically represented by the equations above. The first equation 2.11 represents the fade-out effect, where $y(t)$ is the amplitude of the faded signal at time t, $x(t)$ is the original signal amplitude at time t, t is the current time or sample index, and T is the total duration or total number of samples of the fade-out. This equation implies that the signal is gradually scaled down, and at $t = T$, the signal is scaled down to 40% of its original amplitude. Similarly, the second equation 2.12 represents the fade-in effect, where the parameters have the same representation. In this case, the equation shows that the amplitude of the signal increases linearly from 40% to 100% of its original amplitude as time progresses from 0 to T. The value of 0.4 in these equations was selected after fine-tuning and experimentation, as it gave us the most optimal results in our tests.

On the other hand, when a window signal is classified as a speech signal, the enhancement consists of applying a spectral subtraction process. The spectral subtraction process is, in essence, a classical signal enhancement, where the spectral properties of the speech signals are isolated and enhanced. The noise signal, characterized by its spectral features, is subtracted from the mixed speech and noise signal. This subtraction happens in the spectral domain where the differences between noise and speech signals are typically more evident. The resulting enhanced speech signal is then converted back to the time domain. The goal of this procedure is to suppress the noise elements while preserving as much of the original speech signal as possible, thus improving the clarity and comprehensibility of the speech. Returning to Fig. 2.4, we can see that the enhanced speech and non-speech window signals are combined and concatenated, in an orderly fashion, to generate the resulting enhanced speech signal.

Algorithm 2.2 Proposed Speech Enhancement Algorithm

```
**Input:** $clean\_signal, noisy\_signal, SNR, clip, NN\_model$
**Output:** enhanced_signal
     *Initialisation:*
 1:  $clean\_signal \leftarrow 'clean\ signal.wav'$
 2:  $noisy\_signal \leftarrow 'noisy\ signal.wav'$
 3:  $SNR \leftarrow Desired\ Value$
 4:  $clip \leftarrow 500$
     *Load Pre-trained Neural Network:*
 5:  $net = NN\_model(SNR, clip);$
     *Preparation of Input Data:*
 6:  $noisy\_signal = Update\_noisy\_signal(noisy\_signal, SNR)$
 7:  $clean\_signal\_windows = Window$(clean_signal, clip)
 8:  $noisy\_signal\_windows = Window$(noisy_signal, clip)
 9:  $features$ = Get_features($noisy\_signal\_windows$)
     *Classification & Enhancement:*
10:  $net(features);$
11:  **for** $i$ = 1 to [size of noisy_signal_windows] **do**
12:      **if** window == non-speech **then**
13:          $window = Fade\_out(window\_start)$
14:          $window = Fade\_in(window\_end)$
15:          $window = window * 0.4$
16:      **else**
17:          spectral_subtraction(window)
18:      **end if**
19:  **end for**
     *Concatenation:*
20:  $enhanced\_signal = concat(noisy\_signal\_windows)$
21:  **return** enhanced_signal
```

The Algorithm 2.2 starts by initializing the $clean\_signal$ and $noisy\_signal$, which represent the audio signal without and with noise, respectively. It also sets the desired SNR and a clipping value, to indicate the size of each window. Once the initial variables are set, in step 5 the algorithm proceeds to load a pre-trained neural network model which was made using our training Algorithm 2.1, tailored to the specified SNR and clipping value. This model will later be useful in classifying the audio segments as either speech or non-speech.

Next, the algorithm focuses on data preparation. In steps 6 to 8, it first adjusts the *noisy_signal* to meet the desired SNR and then segments both the *clean_signal* and *noisy_signal* into windows based on the clipping value. Features are then extracted from these noisy signal windows in step 9, and these features are subsequently fed into the neural network for classification. With this setup, the algorithm moves on to the audio enhancement phase in steps 10 to 11. First, the features we got from the windows of the noisy signal are fed into our neural network and subsequently, a loop iterates through each window of the noisy signal. Depending on whether a window contains speech or not the algorithm either applies a fade-in and fade-out to non-speech segments and scales down their amplitude or performs spectral subtraction on segments that contain speech.

Finally, in step 20 after all the windows have been processed and appropriately enhanced, they are concatenated to form a single enhanced audio signal. This final *enhanced_signal* represents the algorithm's attempt to clean up the noisy input signal, and it is returned as the output in step 21. The aim is to make speech clearer and to reduce background noise, leveraging machine learning and signal processing techniques to do so.

2.6 Discussion

In conclusion, our method's primary focus was to enhance speech signals using a framework rooted in Machine Learning and Artificial Neural Networks. We managed to accomplish our two main objectives. We have successfully achieved our two primary research objectives. First, we designed and implemented a neural network that utilizes specific features of a speech signal, enhancing the accuracy and robustness of the speech enhancement process. Secondly, we developed a comprehensive speech enhancement framework. This framework uses both conventional and modern AI-based techniques to apply the right enhancing technique to each part of the speech signals, resulting in a significant improvement in the perceptive quality and intelligibility of the speech signals.

This framework opens new perspectives for real-life applications where high-quality speech is crucial, such as in telecommunication systems, speech recognition software, and hearing aids. However, further improvements can be made by employing more advanced neural network architectures or by considering more sophisticated features for the classification process. Future work could also explore other signal enhancement techniques or adaptive methods for spectral subtraction in the context of non-stationary noise conditions. Optimization techniques for machine learning and neural network tasks such as genetic algorithms are another way to improve and expand upon our proposed method. Based on the results shown in the next chapter, we can say that our ANN classifier together with the spectral subtraction method is able to significantly enhance the resulting speech signal.

CHAPTER 3

IMPLEMENTATIONS AND EXPERIMENTAL RESULTS

In this chapter, first, we present the setup we use for our implementation and the tools we use to implement our framework. In the next section, we illustrate the different experiments carried out to evaluate the performance and accuracy of the proposed method. In addition, we compare our results to a conventional method using objective metrics. Finally, we analyze the performance of the proposed method with corresponding results and discussions.

3.1 Implementation Setup

The implementation phase of this thesis started with establishing the necessary hardware and software environment. This comprised a system equipped with an Intel Core i5 processor, 32GB RAM, and an NVIDIA GeForce RTX 4070 Ti graphics card. The computational efficiency of this setup effectively supported the demands of our machine learning computations.

The software platform used was MATLAB R2022a, augmented with its Deep Learning Toolbox. This environment enabled the design and execution of our complex neural network models, supporting diverse architectures and training options.

3.2 Experiments Setup

The goal of the experiments was to validate and assess the performance of our innovative neural network-based model designed for speech signal enhancement using spectral subtraction techniques. Our experimental setup was broadly divided into two integral stages: the first focused on the training of neural network models, and the second on the enhancement of noisy speech signals using these trained models and getting the final results.

For the training phase, we relied on the NOIZEUS dataset that was introduced in the previous chapter, an industry-standard resource rich in diverse environmental noise conditions. This dataset comes with eight distinct types of ambient noise, such as restaurant and airport, each accompanied by four varying noise intensity levels of 0dB, 5dB, 10dB, and 15dB. Particularly, each environment and noise level within the NOIZEUS dataset contains 30 unique audio files that consist of different sentences pronounced against a backdrop of the respective noisy environments. To maintain a balance between our training and testing data, we intentionally separated the last four audio files in each category for later use in model evaluation. The remaining audio files, amounting to 832 different files, along with an equal number of corresponding clean speech files, were used for model training. While in the NOIZEUS dataset speech signals of only 0db, 5db, 10db and 15db noise levels are present, a unique feature of our training program was used which allowed us to set our desired global SNRs to our models. This allowed us to train separate neural network models at five strategically chosen global SNR settings of -5dB, 0dB, 5dB, 10dB, and 15dB. This configuration ensured that we had a broad spectrum of trained models, capable of handling various noise conditions and a more general understanding of the performance of our models.

Moving on to the second stage, the enhancement of noisy speech signals, we employed the previously separated four audio files from each of the eight noise environments and five noise levels within the NOIZEUS dataset. These files served as the test bed for our trained neural network models. The procedure involved running our enhancement algorithm on these selected files and carefully recording the SNR values obtained post-enhancement. To establish a comparative baseline, the same set of files was also subjected to the traditional spectral subtraction method, and the resulting SNR values were documented. This side-by-side comparison was helpful in evaluating the relative performance and advantages of our neural network-based approach over existing, traditional methods. The key criteria for comparison were improvements in SNR and overall listening quality, allowing us to conclude how much more effective our proposed model is in the realm of speech signal enhancement.

3.3 Results

The result of our proposed method in general was a significantly enhanced speech signal in which noise from various environments was effectively minimized, and the clarity of speech was substantially improved. The final output was a cleaner and higher-quality speech signal, free from the impediments of its earlier form. By reducing background noise in the non-speech segments and enhancing the clarity in the speech segments, our method achieves two key objectives. First, it improves intelligibility, making it easier for human listeners to understand the spoken content. Second, the cleaner signal also aids automated speech recognition systems in accurately interpreting the words. In contrast to our proposed method, speech signals enhanced by traditional spectral subtraction still contained noticeable background noise during the speech and non-speech parts. This residual noise made it more challenging to distinguish spoken words, ultimately reducing the listener's ability to understand the speech clearly and causing increased fatigue during extended listening sessions.

The effectiveness of our model is best understood through a series of tables that present detailed SNR data for both the Proposed Method (PM) and the traditional Spectral Subtraction (SS). As shown in Tables 3.1–3.4, the SNR results for the last four speech signals from the NOIZEUS dataset are presented in each table separately. These tables are organized in a way to facilitate comparisons of our proposed method with spectral subtraction, across various types of background noise and noise levels. They have eight columns that represent the eight distinct noise environments, along with five rows that depict different initial noise levels, including -5dB, 0dB, 5dB, 10dB, and 15dB. In every cell, two SNR values are reported, one represents the Proposed Method result and the other represents the resulting values for Spectral Subtraction. This layout permits a nuanced understanding of how each method performs across different noise environments and at various noise levels.

Table 3.1 Enhanced Speech Signal's (No.27) SNR Results for Proposed Method (PM) and Spectral Subtraction (SS). The Corresponding Speech Utterances are "Bring your best compass to the third class"

SNR	Babble		Car		Exhibition		Restaurant		Street		Airport		Train		Station	
	SS	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**
-5dB	3.63	6.30	3.53	5.72	3.74	6.02	3.74	6.62	3.41	5.90	4.14	5.89	3.77	5.20	3.29	5.96
0dB	5.09	8.73	5.06	7.68	5.14	7.76	5.34	8.96	4.91	8.04	5.88	9.36	5.47	8.11	4.77	7.72
5dB	6.74	10.36	7.79	10.80	8.37	11.85	6.90	11.60	7.27	10.74	7.16	10.44	7.51	10.95	7.47	10.66
10dB	9.79	14.12	10.60	13.54	10.76	13.81	10.28	13.83	11.21	14.56	10.43	14.15	10.81	14.15	10.26	13.34
15dB	13.90	18.22	14.37	16.65	14.27	17.39	13.86	17.10	13.97	18.17	14.24	17.59	14.56	17.15	13.91	16.54

Table 3.2 Enhanced Speech Signal's (No.28) SNR Results for Proposed Method (PM) and Spectral Subtraction (SS). The Corresponding Speech Utterances are "The club rented the rink for the fifth night"

SNR	Babble		Car		Exhibition		Restaurant		Street		Airport		Train		Station	
	SS	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**
-5dB	3.33	6.27	3.33	7.07	3.29	6.35	3.33	3.80	3.25	6.14	3.42	6.48	4.00	5.81	3.57	6.50
0dB	4.87	8.35	4.90	8.82	4.69	8.17	4.79	7.40	4.89	8.07	4.95	8.60	5.76	8.72	5.22	8.42
5dB	7.35	11.65	7.19	11.36	7.14	10.96	7.33	10.34	7.05	10.11	7.82	12.50	8.24	11.07	7.49	11.30
10dB	10.24	14.32	10.40	13.96	10.26	14.02	10.64	14.91	10.97	14.45	10.34	15.64	10.98	14.48	10.55	14.13
15dB	14.02	17.14	14.23	16.97	14.07	17.37	13.99	17.52	14.21	17.60	14.38	17.56	15.13	16.98	14.46	17.04

Table 3.3 Enhanced Speech Signal's (No.29) SNR Results for Proposed Method (PM) and Spectral Subtraction (SS). The Corresponding Speech Utterances are "The flint sputtered and lit a pine torch"

SNR	Babble		Car		Exhibition		Restaurant		Street		Airport		Train		Station	
	SS	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**
-5dB	3.48	6.20	3.87	6.20	4.54	6.23	3.78	6.42	3.57	6.43	3.82	6.91	4.04	7.40	4.86	7.92
0dB	5.19	8.07	5.56	8.18	6.43	8.55	5.30	8.07	5.19	8.36	5.49	8.42	5.91	9.17	6.73	10.06
5dB	7.91	11.26	7.78	11.01	7.87	10.74	7.71	10.84	7.76	10.97	9.10	11.77	8.02	10.89	7.56	10.44
10dB	11.01	14.33	11.03	13.44	12.03	14.67	10.67	15.24	11.27	14.11	12.07	14.48	11.04	14.16	11.20	13.69
15dB	14.56	17.22	14.67	17.05	15.40	18.74	14.59	17.82	14.38	17.48	14.55	16.99	15.19	18.27	14.25	16.63

Table 3.4 Enhanced Speech Signal's (No.30) SNR Results for Proposed Method (PM) and Spectral Subtraction (SS). The Corresponding Speech Utterances are "Let's all join as we sing the last chorus"

SNR	Babble		Car		Exhibition		Restaurant		Street		Airport		Train		Station	
	SS	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**
-5dB	3.62	6.18	3.70	6.35	3.46	6.34	3.59	6.38	3.65	5.64	4.65	6.97	3.88	5.78	3.45	6.67
0dB	5.27	8.10	5.38	8.12	4.99	8.23	5.23	9.22	5.25	8.55	6.55	9.32	5.56	7.81	5.01	8.38
5dB	7.40	10.81	7.67	10.75	7.62	11.02	7.86	11.62	7.06	10.37	7.34	10.55	7.98	10.82	6.80	10.31
10dB	10.43	14.26	10.64	13.40	10.69	14.31	10.15	14.02	10.44	14.22	12.76	15.21	11.03	13.83	10.32	13.34
15dB	13.95	16.83	14.24	16.72	14.45	17.40	13.79	16.42	14.07	16.70	14.20	17.51	15.25	17.52	14.16	16.46

Following these individualized assessments, Table 3.5 below summarizes the findings by showing the average SNR values that were extracted from these four initial tables. This summary table uses the same structure as the preceding tables, presenting a broader perspective on the overall effectiveness of both our Proposed Method and the Spectral Subtraction across all test conditions.

Table 3.5 Average SNR Results for 4 Enhanced Noisy Speech Signals using Proposed Method (PM) and Spectral Subtraction (SS)

SNR	Babble		Car		Exhibition		Restaurant		Street		Airport		Train		Station	
	SS	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**	**SS**	**PM**
-5dB	3.51	6.23	3.60	6.33	3.75	6.23	3.61	5.80	3.47	6.02	4.00	6.56	3.92	6.04	3.79	6.76
0dB	5.10	8.31	5.22	8.20	5.31	8.17	5.16	8.41	5.06	8.25	5.71	8.92	5.67	8.45	5.43	8.64
5dB	7.35	11.02	7.60	10.98	7.75	11.14	7.45	11.10	7.28	10.54	7.85	11.31	7.93	10.93	7.33	10.67
10dB	10.36	14.25	10.66	13.58	10.93	14.20	10.43	14.50	10.97	14.33	11.40	14.87	10.96	14.15	10.58	13.62
15dB	14.10	17.35	14.37	16.84	14.54	17.72	14.05	17.21	14.15	17.48	14.34	17.41	15.03	17.48	14.19	16.66

Analyzing the presented Table 3.5, it is evident that the Proposed Method (PM) outperforms the Spectral Subtraction (SS) in every environment and noise level, showcasing its effectiveness in enhancing noisy speech signals. Across all listed noise levels, ranging from -5dB to 15dB, the proposed method consistently reports higher SNR values compared to the spectral subtraction method. For instance, in a ‘Babble’ noise environment at -5dB, the proposed method achieves an SNR of 6.23, which is significantly higher than the 3.51 SNR achieved by the spectral subtraction method. This trend of superiority continues with the increasing noise levels, reaching a noteworthy improvement at 15dB, where the proposed method attains an SNR of 17.35 compared to SS’s 14.10 in the same noise environment.

The proposed method’s consistent outperformance is further illustrated in various other noise environments such as ‘Car’, ‘Exhibition’, and ‘Street’. In the ‘Car’ environment at 0dB, the PM method reveals an SNR of 8.20, while the spectral subtraction method shows an SNR of 5.22, proving the enhanced robustness of the proposed method in different noise situations.

Furthermore, assessing the values in each row, a progressive increase in SNR is observable for both methods as the noise level escalates from -5dB to 15dB. This pattern suggests that both the Proposed Method and Spectral Subtraction are more effective in higher SNR conditions. However, the rate of increase in SNR is more pronounced for the proposed method, reflecting its superior noise reduction capabilities in diverse SNR scenarios.

Additionally, while analyzing various noise environments, the ‘Airport’ and ‘Train’ environments tend to exhibit relatively higher SNR values for the proposed method, indicating its enhanced adaptability and performance in these specific situations compared to others. For example, at 10dB in the ‘Airport’ environment, the proposed method achieves an SNR of 14.87, one of the highest among the listed environments.

In conclusion, the consistent superiority of the proposed method across different noise environments and levels underscores its potential as a robust and effective technique for enhancing noisy speech signals. The incremental advancements in SNR values across increasing noise levels, in addition to the distinct adaptability in varied noise environments, emphasize the proposed method’s versatility and effectiveness in comparison to the traditional spectral subtraction method. The observed trends and patterns in the table show the promising capabilities of the proposed method in delivering improved performance in diverse real-world noise conditions.

3.4 Performance Analysis

As shown in Table 3.6 below, the average SNR results for all seven noise environments, along with the Mean Increased SNR Values are presented. This table compares the average SNR results for all four enhanced speech signals across 8 noise environments and 5 levels of noise. The first two columns offer average SNR values across all noise environments, thus summarizing the overall effectiveness of each technique. The last two columns, ΔSNR Spectral Subtraction and ΔSNR Proposed Method, offer a calculation of how much each method improved the SNR, measured as the difference between the original and the enhanced signal's SNR for each given noise level and method.

Table 3.6 Average SNR Results for 4 Enhanced Speech Signals across 7 Noise Environments, with Mean Increased SNR Values

SNR	**Spectral Subtraction**	**Proposed Method**	**ΔSNR Spectral Subtraction**	**ΔSNR Proposed Method**
-5dB	3.70	6.24	8.70	11.24
0dB	5.33	8.41	5.33	8.41
5dB	7.56	10.96	2.56	5.96
10dB	10.78	14.18	0.78	4.18
15dB	14.34	17.26	- 0.66	2.26

Upon analyzing the gathered data, it's apparent that our proposed method consistently outperforms the traditional spectral subtraction method across various metrics and conditions. In particular, the ΔSNR values for our proposed method reveal a more substantial improvement over the original noisy speech signals, especially in harsher noise environments. At a low SNR level of -5dB, our proposed method achieves an improvement of 68% over spectral subtraction. This impressive performance extends across all evaluated SNR levels; at 0dB, 5dB, 10dB, and 15dB, the percentage improvements were 57%, 45%, 31%, and 20%, respectively.

This robustness in performance not only suggests the adaptability of our method to varying noise conditions but also its proficiency in significantly improving the initial SNR levels of noisy speech signals. Specifically, at lower SNR levels where speech intelligibility is often critically compromised, our model delivers more significant improvements, thereby emphasizing its utility in real-world scenarios with substantial background noise. As the noise level decreases (i.e., as the initial SNR increases), the relative benefit compared to spectral subtraction decreases in terms of SNR. However, in our listening tests, the improvement in the speech intelligibility and the overall audio quality can be easily perceived by the listener. Even at higher SNR levels like 15dB, the proposed method still outperforms spectral subtraction with a lower percentage improvement of 20.36%. The data corroborates that our machine learning-based approach offers a robust and efficient way to enhance speech signals, presenting a substantial advancement in the realm of speech processing technologies.

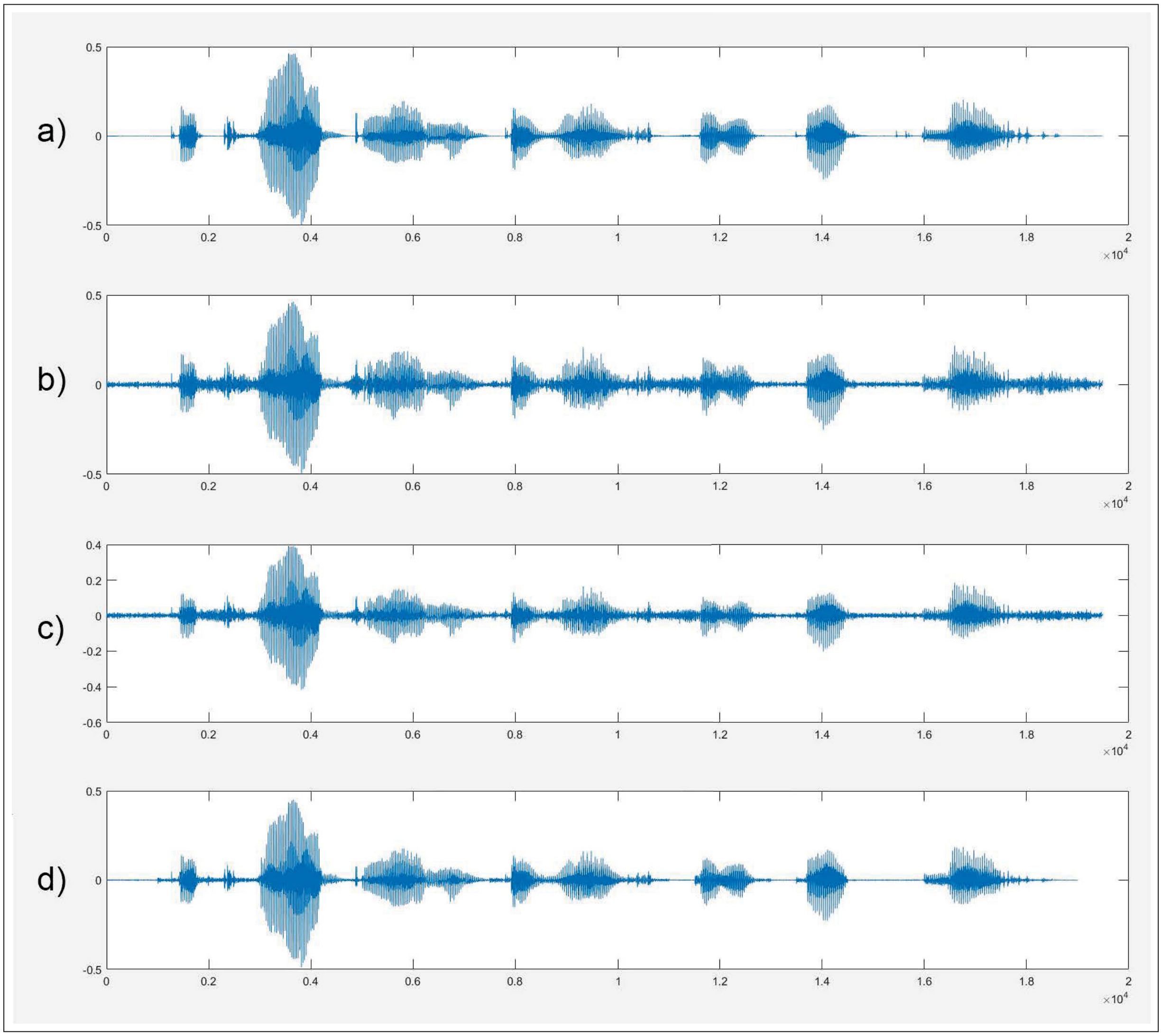

Figure 3.1 Time Domain Representations of: a) Clean Speech Signal (No.28). The Corresponding Speech Utterances are "The club rented the rink for the fifth night".; b) Noisy Speech Signal of Airport Noise with SNR = 10.00 dB; c) Enhanced Speech Signal using Spectral Subtraction with SNR = 10.34 dB and d) Enhanced Speech Signal using Proposed Method with SNR = 15.64 dB;

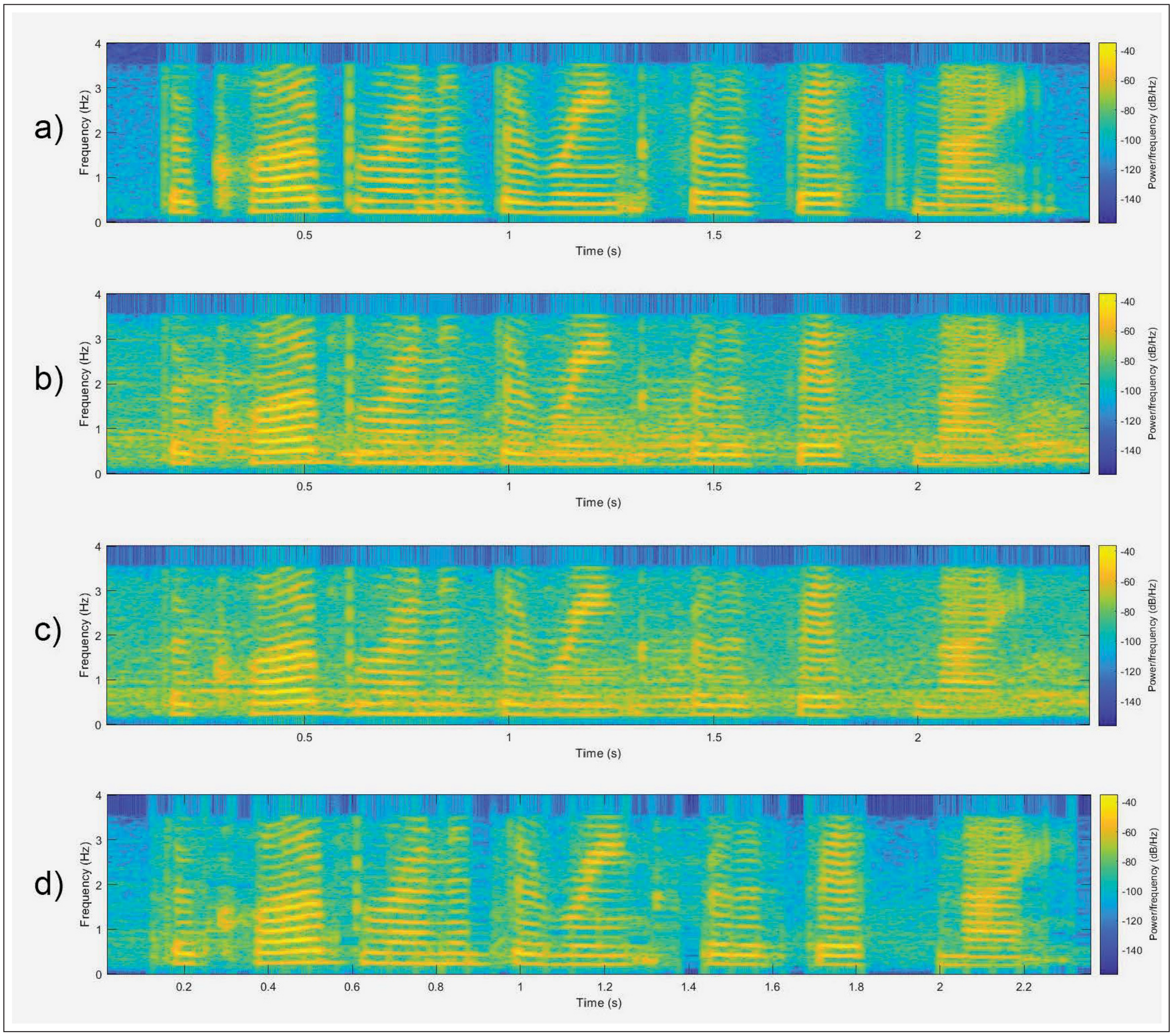

Figure 3.2 Frequency Domain (Spectrogram) Representations of: a) Clean Speech Signal (No.28). The Corresponding Speech Utterances are "The club rented the rink for the fifth night".; b) Noisy Speech Signal of Airport Noise with SNR = 10.00 dB; c) Enhanced Speech Signal using Spectral Subtraction with SNR = 10.34 dB and d) Enhanced Speech Signal using Proposed Method with SNR = 15.64 dB.

Fig. 3.1 and Fig. 3.2 offer compelling evidence of the superiority of the proposed speech enhancement method over traditional spectral subtraction. Fig. 3.1 presents the time-domain representations of clean, noisy, and enhanced speech signals. Similarly, Fig. 3.2 depicts their frequency domain (spectrogram) counterparts. For these figures, we specifically employed the No. 28 speech signal file from the dataset, which was subjected to an airport noise level of 10 dB as a representative example for testing the efficacy of both speech enhancement methods. Using a standard speech signal with the speech utterance of "The club rented the rink for the fifth night" under consistent noise conditions allowed us to make an objective assessment of our method.

In Fig. 3.1, it is clear that the proposed method delivers a substantial improvement in SNR, elevating it to 15.64 dB compared to the 10.34 dB achieved by spectral subtraction. The increased SNR by more than 5 dB in the proposed method signifies not only an enhanced speech quality but also a remarkable suppression of airport noise. In the time domain, our proposed method exhibits distinct advantages over the spectral subtraction technique. A critical observation is the treatment of non-speech segments within the signal. In the spectral subtraction-enhanced output, these segments still contain residual noise, degrading the overall intelligibility and quality of the speech. In contrast, our method effectively minimizes noise in these regions, resulting in amplitudes that are close to zero. This brings the enhanced signal much closer to the clean, original audio, particularly in the non-speech segments.

A spectrogram is a visual representation of the spectrum of frequencies in a sound or other signal as they vary with time or some other variable. It provides insights into the frequency content of a signal, depicting how these frequencies change over time. In a spectrogram, frequencies are represented on the y-axis, time on the x-axis, and the amplitude or intensity of frequencies is usually represented using color or intensity. The spectrogram is computed using the Short-Time Fourier Transform (STFT). The STFT of a signal $x(t)$ can be represented as:

$$X(\tau, \omega) = \int_{-\infty}^{\infty} x(t) w(t - \tau) e^{-i\omega t} dt \tag{3.1}$$

In equation 3.1, $x(t)$ is the time-domain signal being analyzed, $w(t)$ is a window function centered around time τ, and ω denotes the angular frequency. The window function $w(t)$ is typically short, like a Gaussian, Hamming, or Hanning window, effectively concentrating most of their energy in a limited duration. The process involves multiplying the signal $x(t)$ by this window to focus on a small segment of $x(t)$ and then determining the Fourier transform of this segment. By shifting the window across the signal (changing τ), the evolution of the frequency content of $x(t)$ over time is captured. In practical computations, the STFT is usually calculated using the Fast Fourier Transform (FFT) on successive frames or windows of the input signal. The resulting spectrogram then offers a detailed view of the evolution of the frequency content of the signal over time.

The spectrogram in Fig. 3.2 is proof that the proposed method better preserves the frequency components of the original clean speech signal. The enhanced speech through the proposed method shows a closer resemblance to the clean signal, both in terms of time and frequency representations. This result underscores the method's effectiveness in lowering noise while preserving the essential characteristics of the speech signal. Thus, based on the observed SNR values and visual inspections from Fig. 3.1 and Fig. 3.2, it is apparent that the proposed speech enhancement algorithm outperforms the spectral subtraction method. In general, the analysis of both time-domain waveforms and spectrograms reveals that a substantial amount of noise has been effectively removed while maintaining the integrity of most of the clean speech structure at the same time.

Formants are the resonant frequencies of the vocal tract when we speak or sing, producing a harmonic, sustained sound like a vowel. They are responsible for the characteristic timbre of voice sounds, especially vowels. The vocal tract acts as a resonant cavity with certain frequencies being amplified, creating peaks in the frequency spectrum. These peaks are the formants.

The first three formants are pivotal elements in the analysis of speech sounds, primarily owing to their substantial influence on the characteristics and perception of vowel sounds. The First Formant, or F1, is the lowest among the formants and is a product of the resonance created by the entire vocal tract. Its predominant influence lies in modifying the perceived height of the vowel sound, determining how open or close the vowel is. Essentially, F1 plays a crucial role in shaping the basic characteristics of a vowel sound.

Moving to the Second Formant, F2, it is mainly influenced by the configuration and length of the oral cavity. It has a significant impact on the perceived frontness or backness of the vowel sound, determining the relative positioning of the tongue during articulation. It adds another layer to the characteristic sound of the vowel, making it an integral component in the analysis of speech sounds.

Lastly, the Third Formant, F3, is associated predominantly with the rounding of the lips and encompasses several other articulation details. It further refines the characteristics of vowel sounds and is instrumental in distinguishing between similar vowel sounds, contributing to the richness and variety of speech sounds.

Analyzing the first three formants is paramount as they collectively define the unique timbre and quality of vowel sounds, allowing for a nuanced understanding and representation of speech signals. They help in differentiating vowel sounds and contribute to the comprehensibility and distinctive nature of individual voices, making them essential in speech and linguistic studies. Formants are typically calculated using Linear Predictive Coding (LPC), a method that models the vocal tract as a linear system and tries to predict the current sample of an audio signal based on a linear combination of its past samples. LPC coefficients can be used to find the formant frequencies. The LPC equation can be represented as:

$$s(n) = \sum_{k=1}^{p} a_k \cdot s(n-k) + G \cdot e(n) \tag{3.2}$$

$$F = \frac{\text{angle}(r) \cdot \text{Fs}}{2\pi} \tag{3.3}$$

The first equation 3.2 is the Linear Predictive Coding (LPC) equation, which is fundamental in analyzing speech signals. It predicts the current sample, $s(n)$, as a linear combination of its p past samples, with a_k being the LPC coefficients, and G being the gain factor. The term $e(n)$ represents the error in prediction.

The second equation 3.3 calculates the formant frequency, F, using the angle of the roots of the LPC polynomial. The angle of each root is transformed into frequency, using the sampling frequency, F_s. This equation provides the frequency at which the resonant peaks, or the formants, occur in the speech signal, and these formants play a crucial role in characterizing vowel sounds and other speech sounds.

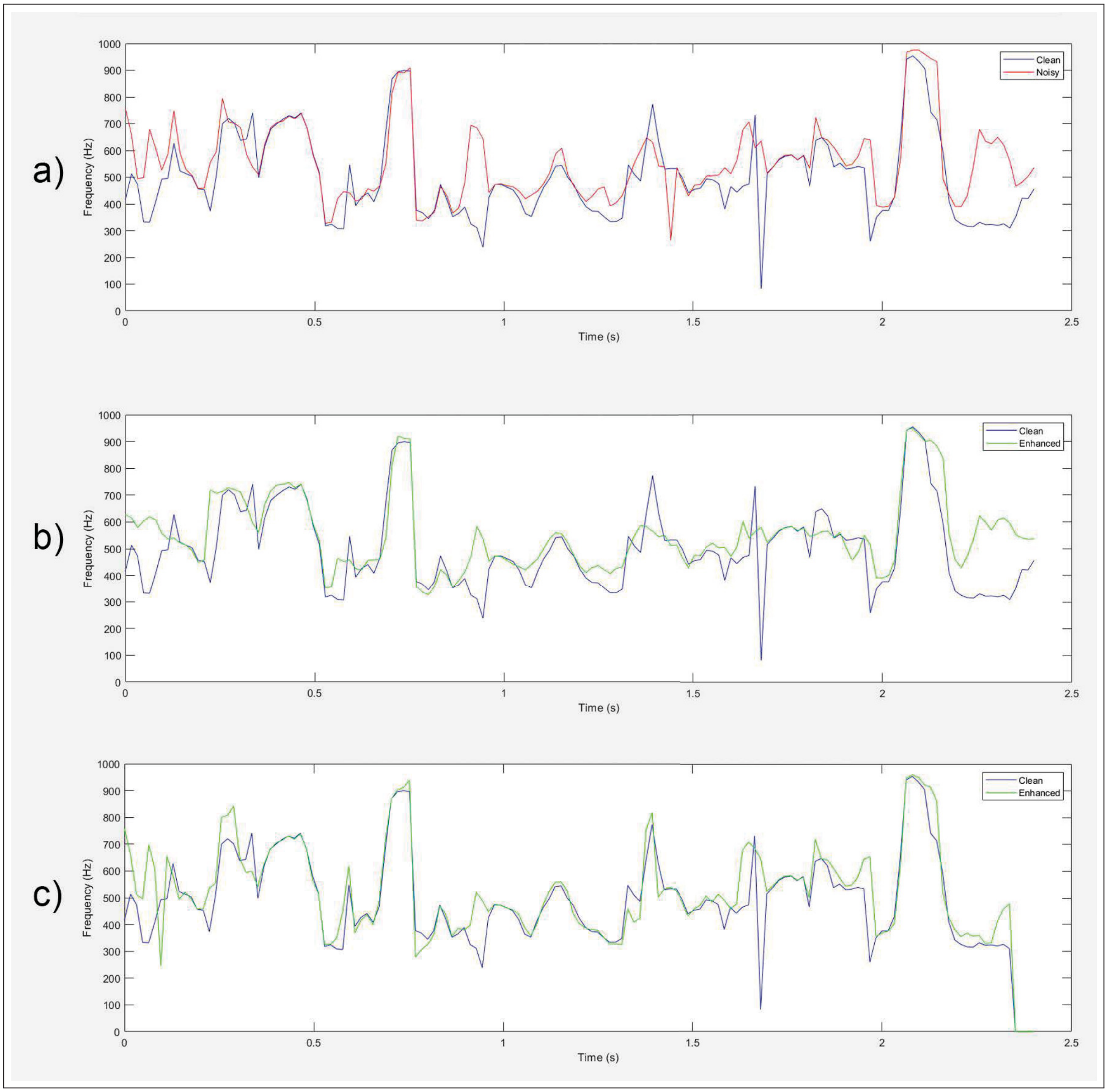

Figure 3.3 Formant 1 Plots of: a) Clean (Blue) Speech Signal (No.28) and Noisy (Red) Speech Signals of Airport Noise with SNR = 10.00 dB. The Corresponding Speech Utterances are "The club rented the rink for the fifth night"; b) Clean (Blue) and Enhanced Speech Signals (Green) using Spectral Subtraction with MSE = 13674.91; c) Clean (Blue) and Enhanced Speech Signals (Green) using Proposed Method with MSE = 9810.92.

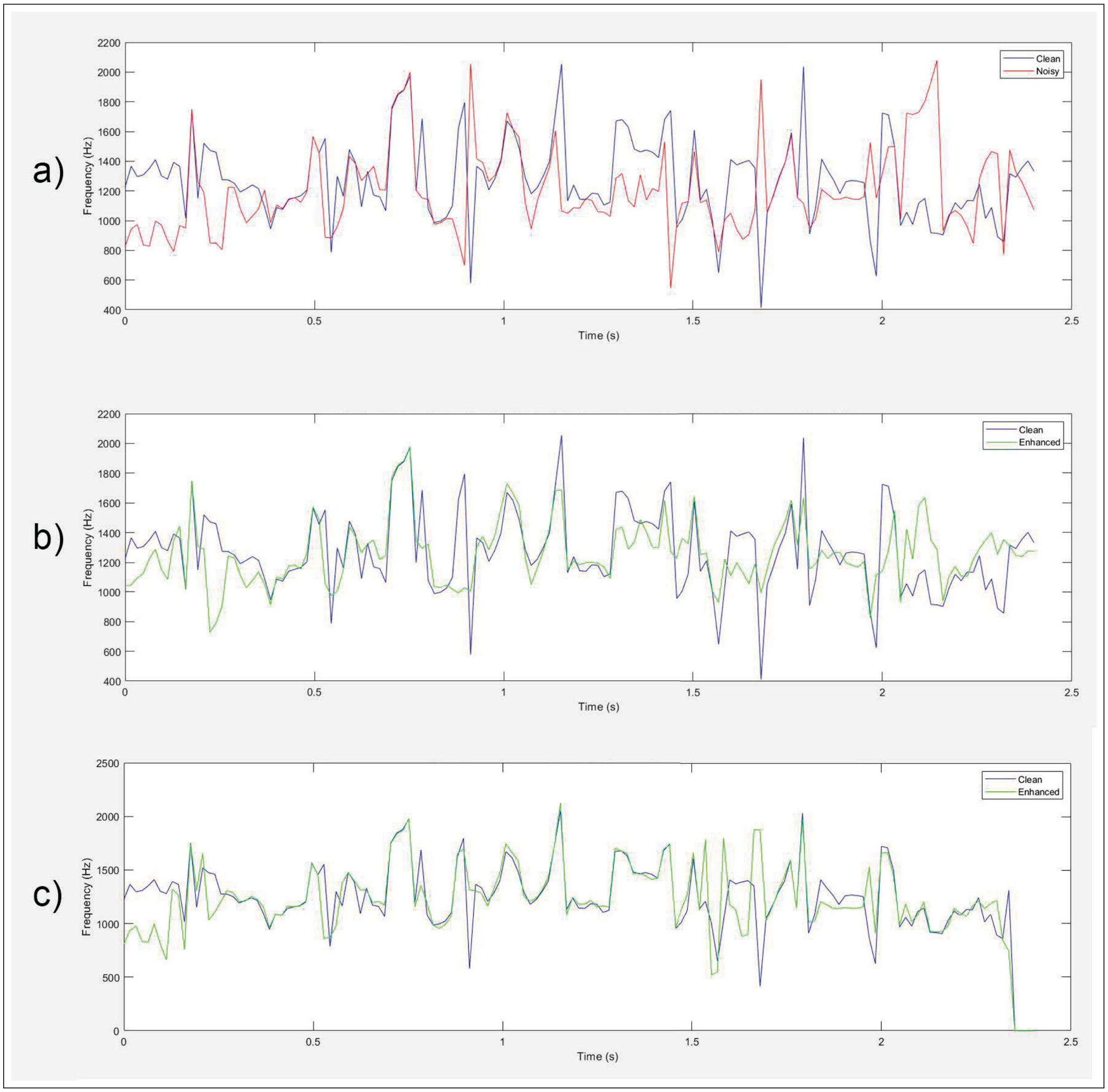

Figure 3.4 Formant 2 Plots of: a) Clean (Blue) speech signal (No.28) and Noisy (Red) Speech Signals of Airport Noise with SNR = 10.00 dB. The Corresponding Speech Utterances are "The club rented the rink for the fifth night"; b) Clean (Blue) and Enhanced Speech Signals (Green) using Spectral Subtraction with MSE = 53382.57; c) Clean (Blue) and Enhanced Speech Signals (Green) using Proposed Method with MSE = 60262.20.

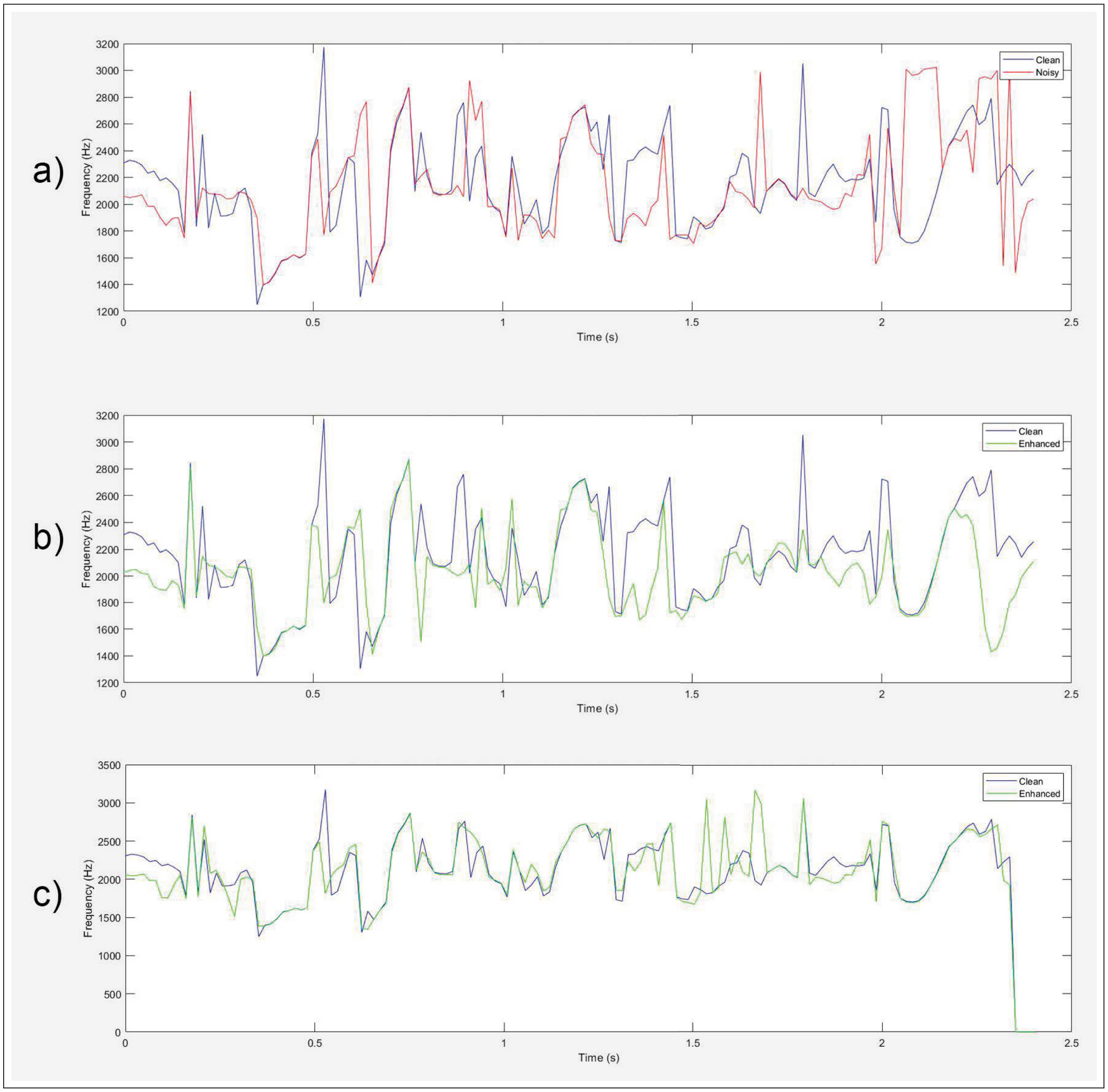

Figure 3.5 Formant 3 Plots of: a) Clean (Blue) Speech Signal (No.28) and Noisy (Red) Speech Signals of Airport Noise with SNR = 10.00 dB. The Corresponding Speech Utterances are "The club rented the rink for the fifth night"; b) Clean (Blue) and Enhanced Speech Signals (Green) using Spectral Subtraction with MSE = 116193.31; c) Clean (Blue) and Enhanced Speech Signals (Green) using Proposed Method with MSE = 68282.38.

The three figures presented show the formant tracks for the first three formants (F1, F2, F3) of speech signal No.28 that we analyzed before. Each figure outlines the comparison between the clean speech signal, the enhanced signal using spectral subtraction, and the enhanced signal using the proposed method. In order to have an objective comparison of our results we used Mean Squared Error (MSE) which we employed to evaluate the discrepancy between the enhanced and clean formant tracks.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(F_{\mathrm{clean},i} - F_{\mathrm{enhanced},i})^2 \tag{3.4}$$

In the MSE, shown in equation 3.4, $F_{\mathrm{clean},i}$ and $F_{\mathrm{enhanced},i}$ represent the frequency values of the clean and enhanced formant tracks, respectively, for the i^{th} data point, and N denotes the total number of data points in the formant track. A lower value of MSE indicates that the enhanced formant track is closely approximating the clean formant track, thus implying higher accuracy in enhancement. In the context of speech signal processing, achieving a lower MSE between clean and enhanced formant tracks is important, as it translates to retaining the essential characteristics of the original speech signal in the enhanced one.

In Figs. 3.3–3.5, the enhanced signal obtained through the proposed method displays a considerable overlap with the clean signal, indicating the effectiveness of the proposed method in preserving formant tracks, in comparison to the signal enhanced by spectral subtraction. This is noticeable in the analysis of the F1 formant in Fig. 3.3, where the proposed method achieves a lower MSE of 9810.92, compared to the 13674.91 MSE attained by the spectral subtraction method. This lower MSE implies a more accurate approximation of the clean signal by the proposed method, especially in terms of the first formant, which is critical as it primarily influences the perceived height of vowel sounds.

In the F2 formant as illustrated in Fig. 3.4, the spectral subtraction method secures a lower MSE of 53382.57 compared to the 60262.20 of the proposed method. This slightly higher MSE for the proposed method in the F2 formant might suggest a marginally better approximation by spectral subtraction in representing the second formant, which affects the perceived frontness or backness of the vowel sounds. However, the difference in MSE is rather trivial, emphasizing the negligible difference in accuracy between the two methods in approximating the F2 formant. Consequently, it is important to consider the overall clarity and quality of the enhanced speech.

Further analysis on the F3 formant in Fig. 3.5 reveals the robustness of the proposed method in approximating the third formant, securing a lower MSE of 68282.38, significantly less than the 116193.31 of the spectral subtraction method. This is a significant result as the third formant is associated with the rounding of the lips and other articulation details, playing a crucial role in the clarity of speech sounds.

In conclusion, analyzing the plots and MSE values for the three formants showcases the potential of the proposed method in delivering a more accurate representation of the original clean signal in most of the evaluated aspects, compared to the spectral subtraction method. While the overlap in the formant tracks is notable, the comprehensive enhancement in intelligibility and quality of the speech signal cannot be overlooked. The proposed method, with its emphasis on the precise representation of formants, significantly influences the preservation of the original characteristics of speech, emphasizing the necessity to prioritize formant preservation in the development of speech enhancement methods to ensure the uniform recovery of the original speech attributes.

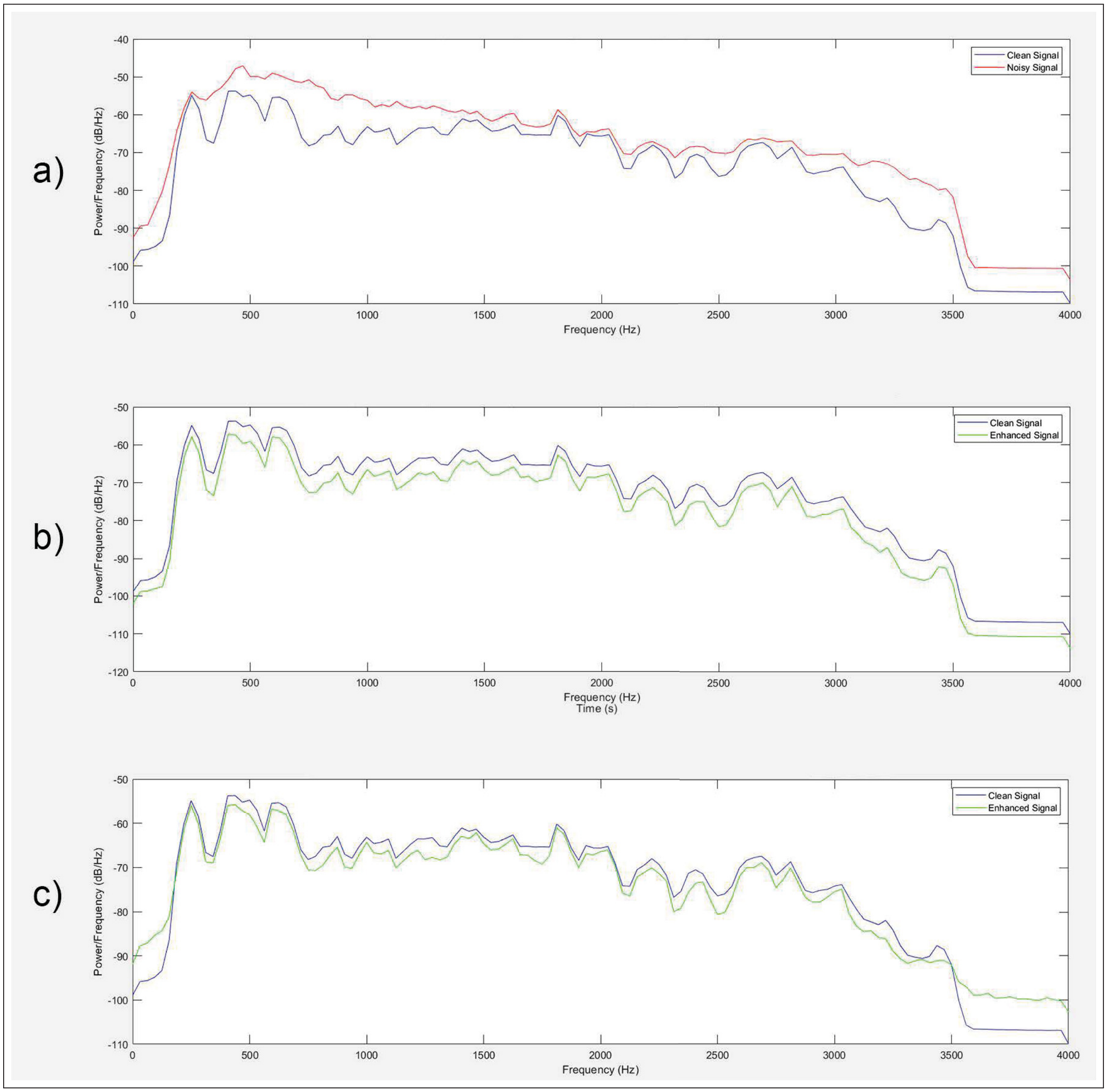

Figure 3.6 The Power Spectral Density Plots of: a) Clean Speech Signal (Blue) No.27 and Noisy Speech Signal of Restaurant Noise with SNR = −5.00 dB (Red) and MSE = 0.0051. The Corresponding Speech Utterances are "Bring your best compass to the third class".; b) Clean Speech Signal (Blue) and Enhanced Speech Signal (Green) using Spectral Subtraction with SNR = 3.74 dB and MSE = 0.0007; c) Clean Speech Signal (Blue) and Enhanced Speech Signal (Green) using Proposed Method with SNR = 6.62 dB and MSE = 0.0004.

The Power Spectral Density (PSD) plot represents how the power of a signal is distributed across different frequency components. It provides insights into the dominant frequencies in a signal and reveals the presence of various frequency components and their corresponding power levels.

$$S(f) = \int_{-\infty}^{\infty} R(\tau) e^{-j2\pi f \tau} d\tau \tag{3.5}$$

The equation 3.5 represents the power spectral density, $S(f)$, of a continuous-time signal. Here, $S(f)$ is the power spectral density as a function of frequency, f. It is obtained by taking the Fourier transform of the auto-correlation function, $R(\tau)$, of the signal, where $R(\tau) = E[x(t)x(t+\tau)]$ is the auto-correlation function, $E[.]$ denotes the expected value, and τ is the time lag. The integration is performed over all time lags from $-\infty$ to ∞. This equation provides a way to analyze the distribution of power over different frequency components of the signal. When analyzing the PSD plots, it is important to consider the relationship and proximity between the clean, noisy, and enhanced signals. In Fig. 3.6, in plot (a), the noisy signal has an MSE of 0.0051 with the clean signal, which is the highest among the three scenarios. This is consistent with the visual interpretation, as the noisy signal is higher than the clean signal, indicating a significant level of distortion and deviation due to the added noise. This suggests that the noise added is broadband and is contributing significant energy across the frequency spectrum. The deviation of the noisy signal from the clean signal at local minima implies that noise significantly corrupts the valleys, possibly affecting the intelligibility and quality of the speech signal.

In plot (b), showing the result of Spectral Subtraction, the MSE is 0.0007, much lower compared to the noisy signal in plot (a). This confirms the visual assessment where the enhanced signal through spectral subtraction envelopes the clean signal but remains closer to it compared to the noisy signal, indicating a substantial reduction in distortion and an improvement in signal quality and intelligibility. In this plot, the enhanced signal is consistently lower and this suggests that spectral subtraction is potentially over-subtracting noise across frequencies, leading to a suppression of the original speech signal. The overlap at lower frequencies may suggest that this method preserves lower frequency components better, which are crucial for speech intelligibility.

Plot (c), showcasing the proposed method, presents the lowest MSE of 0.0004, suggesting that this method is the most accurate in approximating the clean signal, further affirming the visual interpretation. The enhanced signal is closest to the clean signal, demonstrating the effectiveness of the proposed method in reducing noise and maintaining the integrity of the speech signal. Initially, at lower frequencies, the enhanced signal is above the clean signal, suggesting a possible boost or over-enhancement in these regions, which may contribute to the increased SNR. The almost overlapping middle sections imply that the proposed method is more successful at approximating the clean signal in these frequency regions compared to spectral subtraction. The closeness of the enhanced signal to the clean signal suggests a more accurate noise reduction.

To sum up, the quantitative measure of MSE aligns well with the visual interpretations from the PSD plots. The proposed method, with the lowest MSE, shows superior performance in noise reduction and speech enhancement compared to spectral subtraction, presenting a more accurate reproduction of the clean signal. The higher MSE of the noisy signal in plot (a) underscores the level of degradation introduced by noise, emphasizing the effectiveness of the enhancement methods in plots (b) and (c) in mitigating such distortions. The improvements in MSE, along with the visual closeness and improvements in SNR, demonstrate the efficiency of the proposed method in preserving the essential characteristics of the clean signal while achieving a considerable amount of noise reduction.

In conclusion, the performance of our speech enhancement strategy was evaluated using objective metrics such as SNR, MSE, Time domain, Frequency domain, Formants and PSD plots. These measures provided a quantitative analysis of the improvement in speech signals post-enhancement. The performance analysis confirmed the effectiveness of our classification scheme and the subsequent enhancement of speech and non-speech signals. Consequently, these results show that our approach effectively demonstrated its capability to handle and improve noisy speech signals.

CONCLUSION AND RECOMMENDATIONS

In this thesis, we have explored the crucial area of speech enhancement with a focus on using machine learning techniques to improve the quality of speech signals. The study introduced a framework that initially classifies segments of speech signals as either speech or non-speech, followed by the enhancement of these classified segments to produce a speech signal of higher quality. Utilizing neural networks trained on the NOIZEUS dataset for the classification task, we successfully implemented a robust framework for this purpose. The windowing process and classification included in the algorithm ensure more granular and accurate analysis, outperforming traditional methods in both accuracy and reliability.

Beyond classification, the framework also enhances the speech and non-speech segments using a set of tailored enhancement methods. This approach resulted in improved speech signals in terms of SNR and overall listening quality. Notably, this work contributed to this field by developing a machine learning framework specifically for speech enhancement, introducing a new advanced method for accurately segmenting speech signals, and applying customized enhancement techniques to each identified segment. These efforts have led to a noticeable improvement in both the SNR and the perceptual quality of the speech signals. The effectiveness of this method has been validated through the objective evaluation of SNR in various noisy environments. Overall, this research has demonstrated promising results in the enhancement of speech signals using state-of-the-art methods, presenting a viable alternative to more traditional methods.

For future work, several avenues could be considered for exploration. Using a larger dataset with more sound files and diverse environments could make the model even more robust. The algorithm could be fine-tuned to optimize both computational efficiency and accuracy further. Practical, real-world testing of the model could offer additional validation. Integration into existing speech recognition software, telecommunications or audio-visual systems could be explored for more immediate practical applications, and a user experience study could shed light on the subjective quality improvements. By focusing on these areas, subsequent research could further refine the model, paving the way for more sophisticated applications and improvements in speech enhancement.

BIBLIOGRAPHY

Ajmera, J., McCowan, I. & Bourlard, H. (2003). Speech/music segmentation using entropy and dynamism features in a HMM classification framework. *Speech communication*, 40(3), 351–363.

Anusha, V., Indira, K. & Maheshwari, K. (2022). Mathematical Model of Spectral Subtraction Technique.

Atal, B. S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, 64(4), 460–475.

Aylett, M. & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, 47(1), 31–56.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2), 113–120.

Dash, T. K. & Solanki, S. S. (2019). Speech intelligibility based enhancement system using modified Deep Neural Network and adaptive multi-band spectral subtraction. *Wireless Personal Communications*, 111(2), 1073–1087. doi: 10.1007/s11277-019-06902-0.

Deller Jr, J. R. (1993). Discrete-time processing of speech signals. In *Discrete-time processing of speech signals* (pp. 137–143).

Deller Jr, J. (1999). JHL Hansen, and JG Proakis. Discrete-Time Processing of Speech Signals. *IEEE Press*, 445, 119–125.

Ephraim, Y. & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, 32(6), 1109–1121.

Eyben, F., Weninger, F., Squartini, S. & Schuller, B. (2013). Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 483–487.

Ghosh, P. K., Tsiartas, A. & Narayanan, S. (2010). Robust voice activity detection using long-term signal variability. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3), 600–613.

He, R., Tian, Y., Yu, Y., Chang, Z. & Xiong, M. (2022). A Speech Enhancement Method Combining Two-Branch Communication and Spectral Subtraction. *International Conference on Neural Information Processing*, pp. 110–122.

Hu, Y. (2007). Subjective evaluation and comparison of speech enhancement algorithms. *Speech Communication*, 49, 588–601.

Huang, X., Acero, A., Hon, H.-W. & Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.

Kamath, S., Loizou, P. et al. (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. *ICASSP*, 4, 44164–44164.

Khaldi, K., Boudraa, A.-O. & Turki, M. (2016). Voiced/unvoiced speech classification-based adaptive filtering of decomposed empirical modes for speech enhancement. *IET Signal Processing*, 10(1), 69–80.

Kwon, O.-W. & Lee, T.-W. (2003). Optimizing speech/non-speech classifier design using adaboost. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, 1, I–I.

Liang, R., Kong, F., Xie, Y., Tang, G. & Cheng, J. (2020). Real-time speech enhancement algorithm based on attention LSTM. *IEEE Access*, 8, 48464–48476.

Lin, J., Niu, S., Wei, Z., Lan, X., Wijngaarden, A. J., Smith, M. C. & Wang, K.-C. (2019). Speech enhancement using forked generative adversarial networks with spectral subtraction. *Proceedings of Interspeech 2019*.

Loizou, P. (2007). NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms [Online dataset]. Retrieved from: https://ecs.utdallas.edu/loizou/speech/noizeus/.

Loizou, P. C. (2013). *Speech enhancement: theory and practice*. CRC press.

Maganti, H. K., Motlicek, P. & Gatica-Perez, D. (2007). Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, 4, IV–1037.

Martin, A., Charlet, D. & Mauuary, L. (2001). Robust speech/non-speech detection using LDA applied to MFCC. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, 1, 237–240.

Mohammadi, M., Zamani, B., Nasersharif, B., Rahmani, M. & Akbari, A. (2008). A wavelet based speech enhancement method using noise classification and shaping. *Ninth Annual Conference of the International Speech Communication Association*.

Mohammadiha, N., Smaragdis, P. & Leijon, A. (2013). Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10), 2140–2151.

Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, 6(4), 525–533.

Obuchi, Y. (2016). Framewise speech-nonspeech classification by neural networks for voice activity detection with statistical noise suppression. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5715–5719.

O'Shaughnessy, D. (1987). *Speech communication: human and machine*. Universities press.

Park, S. R. & Lee, J. (2016). A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv:1609.07132*.

Rabiner, L. R. (1978). *Digital processing of speech signals*. Pearson Education India.

Ramırez, J., Segura, J. C., Benıtez, C., De La Torre, A. & Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech communication*, 42(3-4), 271–287.

Ryant, N., Liberman, M. & Yuan, J. (2013). Speech activity detection on youtube using deep neural networks. *INTERSPEECH*, pp. 728–731.

Salehi, M. & Mirzakuchaki, S. (2022). A Novel Approach to Speech Enhancement Based on Deep Neural Networks. *Advances in Electrical and Computer Engineering*, 22(2), 71–78.

Scalart, P. et al. (1996). Speech enhancement based on a priori signal to noise estimation. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 2, 629–632.

Shafiee, S., Almasganj, F. & Jafari, A. (2008). Speech/non-speech segments detection based on chaotic and prosodic features. *Ninth Annual Conference of the International Speech Communication Association*.

Shaw, A., Vardhan, R. K. & Saxena, S. (2016). Emotion recognition and classification in speech using artificial neural networks. *International Journal of Computer Applications*, 145(8), 5–9.

Shin, W.-H., Lee, B.-S., Lee, Y.-K. & Lee, J.-S. (2000). Speech/non-speech classification using multiple features for robust endpoint detection. *2000 IEEE International Conference* *on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, 3, 1399–1402.

Smith, D., Lukasiak, J. & Burnett, I. S. (2006). An analysis of the limitations of blind signal separation application with speech. *Signal Processing*, 86(2), 353–359.

Taha, T. M., Adeel, A. & Hussain, A. (2018). A survey on techniques for enhancing speech. *International Journal of Computer Applications*, 179(17), 1–14.

Thambi, S. V., Sreekumar, K., Kumar, C. S. & Raj, P. R. (2014). Random forest algorithm for improving the performance of speech/non-speech detection. *2014 First International Conference on Computational Systems and Communications (ICCSC)*, pp. 28–32.

Thomas, S., Ganapathy, S., Saon, G. & Soltau, H. (2014). Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2519–2523.

Xu, Y., Du, J., Dai, L.-R. & Lee, C.-H. (2013). An experimental study on speech enhancement based on deep neural networks. *IEEE Signal processing letters*, 21(1), 65–68.

Xu, Y., Du, J., Dai, L.-R. & Lee, C.-H. (2014). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1), 7–19.

Yang, Y. & Bao, C. (2018). DNN-based AR-Wiener filtering for speech enhancement. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2901–2905.

Yu, H., Ouyang, Z., Zhu, W.-P., Champagne, B. & Ji, Y. (2019). A deep neural network based Kalman filter for time domain speech enhancement. *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5.

Zhang, X.-L. & Wang, D. (2015). Boosting contextual information for deep neural network based voice activity detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2), 252–264.

Zhang, X.-L. & Wu, J. (2013). Denoising deep neural networks based voice activity detection. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 853–857.

Zhao, N., Xu, X. & Yang, Y. (2011). Sparse representations for speech enhancement. *Chinese Journal of Electronics*, 20(2), 268–272.

Zhao, Y., Xu, B., Giri, R. & Zhang, T. (2018). Perceptually guided speech enhancement using deep neural networks. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5074–5078.