Study and Implementation of Advanced Algorithms for Speech Enhancement

Capstone Report

Dilbar Zhumabayeva

Nazarbayev University
Department of Electrical and Computer Engineering
School of Engineering and Digital Sciences

This project report was created on TexStudio editing platform using LaTeX. All the figures were drawn using draw.io online software tool.

Electrical and Computer Engineering
Nazarbayev University
http://www.nu.edu.kz

Title:
Study and Implementation of Advanced Algorithms for Speech Enhancement

Theme:
Speech Enhancement

Project Period:
Fall 2023 – Spring 2024

Project Group:
Applications of Signal Processing Laboratory (ASP-Lab)

Participant(s):
Dilbar Zhumabayeva

Supervisor(s):
Muhammad Tahir Akhtar

Copies: 1

Page Numbers: 28

Date of Completion:
May 1, 2024

Abstract:

This report presents a study of speech enhancement algorithms aimed at reducing ambient noise and improving speech quality. The report focuses on traditional methods such as spectral subtraction as well as advanced methods based on adaptive filters. Implementation of these algorithms was done using MATLAB and evaluation was done using objective measures such as Perceptual Speech Quality Score (PESQ) and Short-Term Objective Intelligibility (STOI). The results indicate the effectiveness of these algorithms in improving speech signals.

Contents

Dedication

To my family, whose unwavering support has been my motivation throughout my studies. Their support and belief in me has been the driving force behind my achievements, and I am deeply grateful for their love and support.

Preface

Speech enhancement is a critical area of research with wide applications in telecommunications, multimedia, and human-computer interaction. This project examines various speech enhancement algorithms, implements them using MATLAB, and evaluates their performance using objective measures such as Perceptual Evaluation of Speech Quality (PESQ) and Short-Term Objective Intelligibility (STOI).

The main goal of this project is to develop and compare different speech enhancement techniques to improve speech quality in noisy environments. Algorithms studied include spectral subtraction, minimum mean square error estimation (MMSE), and adaptive filter-based methods such as least mean square (LMS) and normalized LMS (NLMS).

The implementation of these algorithms involves processing noisy speech signals to reduce background noise while maintaining speech quality and intelligibility. The performance of each algorithm is assessed using objective metrics such as PESQ, which measures the perceived quality of improved speech, and STOI, which assesses speech intelligibility.

This project is a comprehensive study of speech enhancement algorithms to understand their performance in different noise scenarios. The MATLAB implementation and evaluation results presented here make valuable contributions to the field of speech processing and serve as a basis for further research in this area.

Nazarbayev University, May 1, 2024

Dilbar Zhumabayeva
<dilbar.zhumabayeva@nu.edu.kz>

Acknowledgments

I would like to express my deepest gratitude to my family for their support and understanding throughout this project. Their love, support and belief in me have been a constant source of strength. I am truly fortunate to have such a loving family who have supported me every step of the way.

During the course of the project, I had the privilege of conducting ongoing project in the laboratory under the supervision of Prof. Muhammad Akhtar. I am deeply grateful to him for his exceptional support and guidance throughout the project. Prof. Akhtar's ideas, experience and mentorship were invaluable, especially during our regular meetings and discussions. His leadership significantly influenced the direction and success of this research.

I thank my friends Gulzat, Aktilek and Diana for their friendship, support and understanding. Their presence made this journey more enjoyable and meaningful. Their encouragement and belief in me has been a constant source of motivation and I am grateful for their continued support.

I would also like to thank the ASP-Lab research laboratory and its collaborators for their cooperation and support throughout the project. Their contributions and insights were invaluable in advancing this research.

Thank you everyone for your invaluable contribution and support.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background.

The real world is full of different noises. The noise appears in different shapes and forms in daily life for example the noise of people talking to each other, the noise of the car engine, or the noise of fans, etc. The noise can be stationary or non-staionary. In daily life, noise is not stationary such as the restaurant noise where multiple people talking in the background. Dealing with the non-stationary noise is more complicated as the spectral characteristics of the noise vary. So, the speech enhancement is essential as different types of speech enhancement algorithms is needed to solve problems based on the noise type. The goal of speech enhancement is to distinguish a desired speech signal from unwanted background sounds such as ambient noise, overlapping speech, and echoes caused by the surrounding environment [1]. It can be used in a wide range of practical contexts, including mobile speech communication, digital hearing aids and robust automatic speech recognition. Algorithms for improving speech include filtering approaches, spectrum subtraction techniques, model-based methods, and wavelet-based techniques [2]. The spectrum subtraction algorithm was one of the first noise reduction algorithms proposed [3]. In spectral substation, noise is assumed to be stationary, which does not best solve the problem of subtracting noise from pure speech as noise is inherently non-stationary [4]. Deep learning technology was first applied to voice recognition and produced very good results, and then researchers applied deep learning technology to speech enhancement.The effect of deep learning-based speech enhancement is substantially stronger than that of conventional speech enhancement methods, especially when dealing with stationary noise [5].

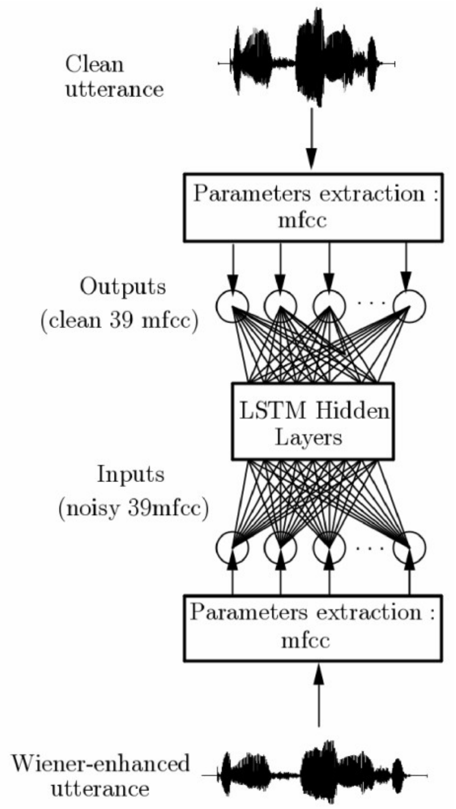

Figure 1.1: Denoising autoencoder with LSTM units in hidden layers[12]

1.2 State-of-the-art and related works.

Complex mathematical modeling is utilized behind the processing of big data in deep neural network (DNN), which consists of an input layer, output layer, and at least one hidden layer. The DNN can manage the intricate nonlinear interactions between noisy and pure speech and subsequently improve speech [6]. A large number of extracted features increases the accuracy of pure speech recognition [7, 8]. Mapping and masking targets are widely used in recent DNN-based speech enhancement [9, 10]. Different masking types and matching approaches were compared, and it was concluded that none of the approaches is the best with respect to all objective measures, since the choice of target type depends on the application of the speech enhancement algorithm [11].

Recent research in signal processing has shown that traditional methods of speech enhancement [12, 13, 14], acoustic echo cancellation [15] can be combined with methods based on machine learning. The approach to enhance noisy speech involves a hybrid two-stage Wiener filtering process combined with a set of long short-term memory (LSTM) networks [12]. These LSTM networks are used to map the Wiener filter output to pure features, thereby improving the quality of noisy speech signals. The Wiener filter is applied to the noisy utterances as the first stage in order to improve the enhancement of noisy utterances. The waveform at the filter's output is parametrized and an LSTM autoencoder is trained in the second stage [12]. The Wiener-enhanced parameters are mapped to clean parameters by each network. Figure 1.1 illustrates the method used to denoise 39 Mel-frequency cepstral coefficients (MFCCs). Training is done to make sure that the outputs closely resemble the corresponding clean vector. As a result, the hybrid technique has shown notable improvements in speech noise reduction despite the high computational cost of training the required networks [12]. To improve performance for particular noise types and intensities, additional noise-aware systems must be added for practical implementation. Most speech improvement algorithms did not take into account information about the phase of impaired speech, since phase information does not significantly reduce the quality of improved speech [16]. But using information about the phase of the speech signal to improve the speech signal improves the intelligibility of the speech signal [17]. Phase-based speech improvement methods were developed using machine learning methods [18, 19, 20, 21]. The use of phase of original signal showed a significant improvement in improved speech.

There is another hybrid approach that combines the advantages of classical noise reduction, known for its effectiveness against quasi-stationary noise, with the exceptional effectiveness of LSTM neural networks in suppressing rapidly changing noise and interference signals [14]. Hybrid methods not only produced clear speech, but also suppressed background noise [22]. The author concludes that the use of different target extraction methods plays an important role and the architecture of the neural network is equally important [23]. Kalman filtering, adaptive filters and other traditional speech enhancement techniques can be used to develop improved methods [24, 25, 26].

1.3 Motivation.

Improving speech is challenging because of the complexity of speech signals and the wide range of noise sources and acoustic conditions. Deep learning and other machine learning techniques have shown promising results in solving this problem, making it an attractive and active area of research. Machine learning for speech enhancement is an interdisciplinary field that combines knowledge in signal processing, machine learning, and acoustics. Recent studies have shown that traditional speech enhancement methods can be combined with machine learning. There is variety of machine learning architectures that can be used to improve speech. Studying this topic gives a variety of skills and information that can be applied to different areas besides voice enhancement. Moreover, the research results can be used in various fields such as speech recognition and acoustics as they improve clear speech in noisy environments.

1.4 Report Organization.

The first section provides an overview of speech enhancement and its importance in improving speech quality in noisy environments. It discusses recent advances in speech enhancement, highlighting the need for more efficient algorithms to deal with different types of noise.

Second chapter details traditional methods such as spectral subtraction and minimum mean square error (MMSE) estimation. Their options and limitations are also discussed. Additionally, it covers adaptive filter-based approaches such as least mean squares (LMS), normalized LMS (NLMS), kernel LMS (KLMS), and normalized kernel LMS (NKLMS), explaining their principles and applications in improving speech.

The third chapter describes the methodology for implementing the discussed algorithms using MATLAB. It involves tuning simulation parameters such as noise types, signal-to-noise ratio (SNR) levels, and evaluation metrics. Simulation results are presented and analyzed, comparing the performance of different algorithms under different conditions.

The final chapter summarizes the study and draws conclusions based on the results. It discusses the effectiveness of each algorithm in reducing noise and improving speech quality. The section also provides information on future research directions and potential improvements to the algorithms studied.

Chapter 2

Speech Enhancement Algorithms

There are three main types of the traditional space enhancement methods.

- Spectral subtraction-based algorithms: These are the simplest and straightforward algorithms to implement for speech enhancement applications. The main principle algorithm is that noise is assumed to be additive and stationary, and speech signal is obtained from subtraction of the noise from the noisy signal by estimating the noise when there is no speech .

- Statistical-model-based algorithms: The speech enhancement problem is posed in a statistical estimation framework. Given a set of measurements, corresponding to the Fourier transform coefficients of the noisy signal, we wish to find a linear (or nonlinear) estimator of the parameter of interest, namely, the transform coefficients of the clean signal. The Wiener algorithm and MMSE algorithms, among others, fall in this category.

- Subspace: are based on the linear algebra theory. The main principle behind the subspace algorithms is that the clean signal may be limited to a subspace of noisy Euclidean space.Based on this principle, a noisy signal can be decomposed into two subspaces: the first subspace contains mainly the clean signal, and the second subspace contains mainly the noise signal. Therefore, the clean signal can be estimated by removing the noise vector component lying in the "noise subspace".

Traditional speech enhancement methods primarily aim to improve the quality and intelligibility of speech signals under various acoustic conditions. This section discusses and implement one of the commonly used approaches, spectral subtraction. Statistical model-based algorithms, including simple MMSE algorithms and their variants, will also be implemented. Furthermore, this section discusses adaptive filter-based speech enhancement algorithms. The section includes a detailed

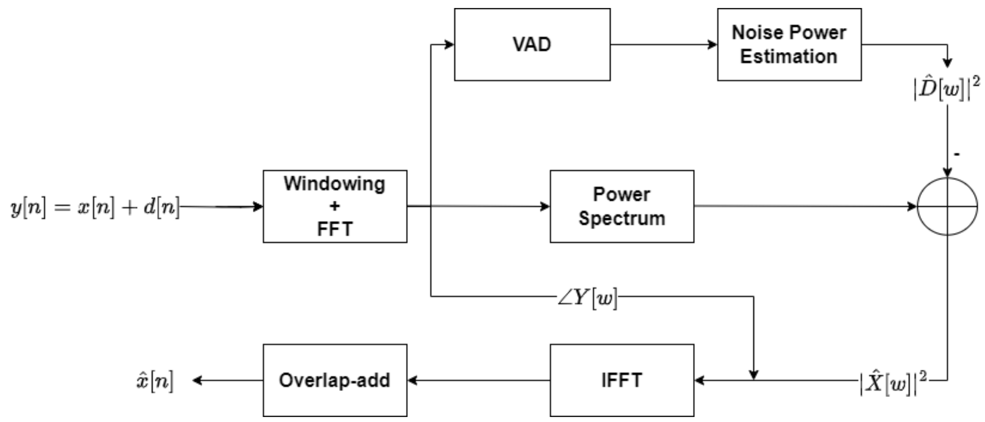

Figure 2.1: Block diagram of spectral subtraction

explanation of the LMS algorithm, which is a widely used adaptive signal processing technique for system identification and denoising. In addition, the kernel-based LMS algorithm will be further developed, with a focus on its application in nonlinear adaptive filtering problems.

2.1 Spectral Subtraction-based Algorithms

The spectral subtraction is a conventional and widely used algorithm used for single channel speech enhancement. The algorithm is simple and there are several variants of the algorithm. In single channel speech enhancement, it is assumed that only one microphone is available and the noise is extracted during the periods of pauses, which requires a stationary assumption of the background noise. During the subtraction process, care must be taken to avoid distortion of the speech signal. Too much subtraction can result in the removal of important speech information, and too little subtraction can result in a significant amount of interfering noise being introduced into the signal [27] .

As shown in Figure 2.1, spectral subtraction involves several stages, including windowing, FFT (Fast Fourier Transform), and clean speech signal recovery. Here's a step-by-step explanation:

1. Window: The noisy speech input signal is divided into frames of short duration, typically 20–30 milliseconds. Each frame is multiplied by a window function to reduce spectrum leakage. FFT: The windowed frames are then passed through an FFT to convert them from the time domain to the frequency domain. This process creates a power spectrum of the noisy speech signal for each frame.

2. Estimation of noise power. Voice Activity Detection (VAD) algorithms are used to estimate the noise power at each frequency bin of the power spectrum. VAD determines whether the frame is dominated by noise or speech. Noise power is often assessed during non-speech segments.

3. Noise subtraction: The estimated power spectrum of the noise is subtracted from the power spectrum of the noisy speech signal. This results in estimated clean speech.

4. Reconstruction: The resulting modified power spectrum is converted back to the time domain using inverse FFT (iFFT). The reconstructed frames are overlapped and summed to produce the final clean speech signal. Information about the phase of noisy speech is used to reconstruct enhanced speech, since phase is not assumed to be critical to speech quality and intelligibility.

5. Overlapping and adding. To reconstruct a clear speech signal, the frames are typically overlapped by 50% and added together. This helps smooth out transitions between frames and reduce distortion.

In spectral subtraction noise is assumed that it is uncorrelated with clean speech. Noisy speech can be represented as

$$y[n] = x[n] + d[n], \tag{2.1}$$

where $y[n]$ is an input signal which corrupted by noise, $x[n]$ is clean speech signal and $d[n]$ is additive noise signal.

The noisy speech signal is processed on a frame-by-frame. The representation of noisy speech signal in the short-time Fourier transform (STFT) domain is given by

$$Y[w,k]) = X[w,k] + D[w,k], \tag{2.2}$$

where k is the frame number.

In further derivation, we drop k for simplicity because it is assumed that the speech signal is segmented into frames . So, the short term power spectrum of $y[n]$ is given by:

$$|Y[w]| = |X[w]| + |D[w]|. \tag{2.3}$$

The clean speech can be estimated by subtracting the magnitude spectrum of noise from the input signal (Fig.2.1).

$$|\hat{X}[w]|^2 = |Y[w]|^2 + |\hat{D}[w]|^2, \tag{2.4}$$

where "$\hat{X}$" indicates the estimated parameter of interest. The noise spectrum estimated $|D[w]|^2$ by averaging the the recent pauses during speech:

$$|\hat{D}[w]|^2 = \frac{1}{M}\sum_{j=1}^{M-1} |Y_{sp_j}[w]|^2|, \tag{2.5}$$

where M is the number of consecutive frames of speech pauses and typically 5 frames is used for noise spectrum estimation. The speech pauses estimated using the VAD.

Although the spectral subtraction method dramatically reduces noise, it has some serious limitations. It is evident that the efficiency of spectral subtraction is heavily dependent on correct noise estimation, which is a difficult undertaking in most cases. When the noise estimate is not accurate, two primary issues arise: remnant noise with musical structure and speech distortion.

The phase estimate of the speech is also required to reconstruct the resulting signal. The phase of the noisy signal is commonly used as the phase of the expected clean speech signal, based on the assumption that short-term phase is largely insignificant to human ears. So, the speech signal in a frame is estimated as:

$$\hat{X}(w) = |\hat{X}[w]|e^{j\angle Y[w]}. \tag{2.6}$$

The estimated form of the speech signal is reconstructed in the time domain by the inverse Fourier transform using the overlap and add method.

The simple spectral subtraction algorithm, although effective in reducing noise, has a noticeable disadvantage known as musical noise. This noise results from over-subtraction, especially when the noise level is lower than expected, resulting in negative power and signal distortion. To solve this problem, several modifications of the simple spectral subtraction algorithm have been proposed. One such modification involves the introduction of an over-subtraction factor, which helps eliminate the effects of over-subtraction and reduce the occurrence of musical noise. This modification showed promise in improving the performance of the spectral subtraction algorithm and reducing signal distortion.

2.2 Statistical-Model-Based Algorithms

The Minimum Mean Square Error short-time spectral amplitude based model based algorithms have simple implementation and low computational complexity. In this algorithm , noise is assumed as additive white Gaussian and stationary. According to the algorithm , minimized mean-square error between the estimated and true magnitudes is calculated using:

$$e = E(\hat{X}_k - X_k)^2. \tag{2.7}$$

The Log-MMSE is modification of the simple MMSE estimator with main difference taking squared error of the log-magnitude spectra . The estimator that minimizes the mean-square error of the log-magnitude spectra is:

$$E(logX_k - log\hat{X}_k)^2. \tag{2.8}$$

The optimal Log-MMSE estimator is derived by calculating the conditional mean of the logarithm of X_k:

$$log\hat{X}_k = E\{logx_k|Y[w_k]\}. \tag{2.9}$$

From equation above $\hat{X}_k$:

$$\hat{X}_k = exp\{E\{logx_k|Y[w_k]\}. \tag{2.10}$$

The solution to the Log-MMSE problem can be expressed as:

$$\hat{X}_k[m] = \frac{\hat{\xi}_k(m)}{\hat{\xi}_k(m)+1} exp\{\frac{1}{2}\int_{\hat{v}_k[m]}^{\infty} \frac{e^{-t}}{t} dt\} Y_k[m]. \tag{2.11}$$

The a priori SNR $\xi_k[m]$ is

$$\xi_k[m] = \frac{\lambda_{x_k}[m]}{\lambda_{d_k}[m]}, \tag{2.12}$$

where $\lambda_{x_k}(m)$ is the spectral power of clean speech and $\lambda_{d_k}(m)$ is the spectral power of noise

In general, Log-MMSE estimator is superior to simple MMSE estimator due to the fact that Log-MMSE estimator reduces residual noise and, most importantly, does not affect the speech signal itself, i.e., it does not introduce much speech distortion.

The previously discussed methods assumed that speech was constantly present. However, in reality, speech often includes numerous pauses even during periods of speech activity. In addition, speech may be absent at a certain frequency even during voiced segments. An algorithm that takes into account speech presence uncertainty (SPU) and is called log-MMSE with SPU. A log-MMSE with SPU derived as:

$$log\hat{X}_k = E\{logX_k|Y[w_k], H_1^k\}P(H_1^k|Y[w_k]), \tag{2.13}$$

where H_1^k is the hypothesis that speech is present. Overall, incorporating speech presence uncertainty into the estimator improves results.

2.3 Adaptive Filter-Based Speech Enhancement Algorithms

In 1960, Widrow and Hoff proposed the least mean square (LMS) algorithm . Due to its computational simplicity, the LMS algorithm is widely utilised in a variety of adaptive filtering applications. Also, because of unbiased convergence to the Wiener solution LMS is widely used and there are a lot of variants of LMS algorithm . There are various modifications of LMS algorithm such as normalized, filtered, etc.

The LMS adapts the filter tap weight which minimizes the mean-square error. The practical scheme for ralizing Weiner filters is LMS algorithm [28]. The operation of the LMS is based on a weight vector, denoted $w[n]$:

$$\mathbf{w}[n] = [w_0[n], w_1[n], \cdots, w_{M-1}[n]]^\top. \tag{2.14}$$

When LMS algorithm processes new samples , the new weight vector updated to:

$$\mathbf{w}[n+1] = \mathbf{w}[n] - \mu \nabla_w J[n], \tag{2.15}$$

where µ represents learning or adaptation step and J represents gradient of algorithm and defined as:

$$\nabla_{\mathbf{w}} J[n] = -\mathbf{u}[n]\, e^*[n], \tag{2.16}$$

where u(n) is the input signal, while e(n) represents the error signal. The error signal can be defined as the difference from the desired signal d(n)

$$e[n] = d[n] - \mathbf{w}^H[n]\, \mathbf{u}[n]. \tag{2.17}$$

The LMS algorithm uses the steepest descent method to update the weight vector to achieve the optimal Wiener solution. The optimal solution of Weiner filter shown below:

$$\mathbf{w}_o = \mathbf{R}^{-1}[n]\, \mathbf{p}[n]. \tag{2.18}$$

where R is the autocorrelation matrix and p is the cross-correlation vector.

One of the main disadvantages of the LMS algorithm is its sensitivity to input data scaling. This sensitivity makes it difficult to choose a learning rate µ that guarantees the stability of the algorithm. To solve this problem, a normalized LMS was introduced. Unlike the LMS algorithm, the NLMS algorithm is not only stable, but also deterministic at each sampling moment [29]. Step size normalization affects the weight update equation and changes to:

$$\mathbf{w}[n+1] = \mathbf{w}[n] - 2\, \frac{\mu}{\|\mathbf{u}[n]\|^2}\, \nabla_w J[n], \tag{2.19}$$

where μ is determined based on the amplitude of the input signal.

2.3.1 Kernel Based-Adaptive Filters

The class of algorithms that solve nonlinear problems used for pattern analysis in machine learning is called the kernel method. Kernel methods involve transforming linear classifiers to solve nonlinear problems. The concept of the kernel method is to map the input data into a high-dimensional feature space, which allows for more efficient processing of the data. The adaptive kernel filtering is the branch of online kernel-based learning that deals with nonlinear regression [30]. The inner products in feature space can be calculated using a positive definite kernel function that satisfies the Mercer condition, which is

$$\kappa(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle. \tag{2.20}$$

The kernel trick allows to implement inner product-based algorithms in feature space by replacing all inner products with kernels. The Gaussian kernel is the most used kernel function :

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right). \tag{2.21}$$

According to the Representator's theorem , problems' nonlinearities can be represented as an extension of the kernel in terms of training data.

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}). \tag{2.22}$$

The Kernel Least Mean Squares (KLMS) is novel approach that can improve estimation and prediction of certain time series [30]. The fundamental concept involves implementing the linear LMS algorithm in kernel space.

$$\Omega[n+1] = \Omega[n] + 2\mu \times e[n] \times \Phi(u[n]). \tag{2.23}$$

In the High-Dimensional Space (HDS), the weight vector represented as $\Omega(n)$. The expected output, $y[n]$, will be determined by:

$$y[n] = \langle \Omega[n], \Phi(u[n]) \rangle, \tag{2.24}$$

where $\Phi(u[n])$ is the transformation of input vector into the infinite feature vector, and its components are united linearly by an infinite-dimensional weight vector. The non-recursive form can be expressed as follows:

$$\Omega[n] = \Omega[0] + 2\mu \sum_{i=0}^{n-1} e[i]\Phi(u[i]). \tag{2.25}$$

When $\Omega[0] = 0$, Equation simplifies to:

$$\Omega[n] = 2\mu \sum_{i=0}^{n-1} e[i]\Phi(u[i]). \tag{2.26}$$

Based on above equations:

$$\begin{aligned} y[n] &= \langle \Omega[n] \times \Phi(u[n]) \rangle \\ &= \langle 2\mu \textstyle\sum_{i=0}^{n-1} e(i)\Phi(u[i]), \Phi(u[n]) \rangle \\ &= 2\mu \textstyle\sum_{i=0}^{n-1} e(i) \langle \Phi(u[i]), \Phi(u[n]) \rangle. \end{aligned} \tag{2.27}$$

The $y[n]$ can be calculated :

$$y[n] = \mu \textstyle\sum_{i=0}^{n-1} e[i] \mathrm{K}(u[i], u[n], \tag{2.28}$$

where K is defined as Kernel function. Above equation is defined as the Kernel LMS algorithm. As the system's error rate decreases over time, we can ignore the error term $e(n)$ after a certain number of samples ξ and make predictions for new data based on the past errors:

$$y[n] = \mu \textstyle\sum_{i=0}^{\xi} e[i] \mathrm{K}(u[i], u[n]). \tag{2.29}$$

One of the algorithm's advantages is its ability to predict in non-linear channels. But the results of the KLMS are sensitive to step-size and signal amplitude stability, which can be improved by NKLMS. The normalization process is identical to that in the Normalized Least Mean Squares (NLMS), but it occurs after the mapping to the Hilbert space, resulting in [31]:

$$\mathbf{w}[n+1] = \mathbf{w}[n] - 2 \frac{\mu}{\|\boldsymbol{\kappa}_\gamma[n]\|^2} e[n] \, \boldsymbol{\kappa}_\gamma[n]. \tag{2.30}$$

NKLMS has various advantages over NLMS such as :

1. simple implementation,
2. simplicity of working parameters selection,
3. convergence time is faster ,
4. little additional processing time.

Chapter 3

Results and discussion

3.1 Simulation Parameters

The simulation utilized the NOIZEUS database. The noisy database includes 30 IEEE statements (said by three male and three female speakers) that have been distorted by eight different real-world noises at various SNRs. The noise signals were obtained from the AURORA database, which includes various recordings from different environments such as babble (crowd noise), car, exhibition hall, restaurant, street, airport, train station and train. Noise signals were added to speech signals at SNR of 0, 5, 10, and 15 dB. The NOIZEUS speech corpus was used to evaluate the quality of various speech enhancement algorithms. Table 3.1 provides a summary of key information about the NOIZEUS database. The simulation parameters listed in Table 3.2 were utilized to evaluate the performance of Spectral Subtraction, MMSE, Log-MMSE, and Log-MMSE with SPU in the context of speech

Table 3.1: Description of the NOIZEUS database

Aspect	**Description**
Database Name	NOIZEUS
Usage	Modeling and evaluation of speech enhancement algorithms
Contents	Noisy speech signals distorted by eight different real-world noises at various SNRs - 30 IEEE statements said by three male and three female speakers - Noises from AURORA database: babble (crowd noise), car, exhibition hall, restaurant, street, airport, train station, and train
Input SNRs	SNRs of 0, 5, 10, and 15 dB

Table 3.2: Simulation Parameters for Conventional Speech Enhancement Algorithms

Parameter	Value
Sampling Frequency	8 kHz
Number of Different Noises	6
SNR Levels	0, 5, 10, 15 dB
Hamming Window Size	20 ms
Overlap-Add Percentage	40%
Input Speech Signal	6 (3 male and 3 female speakers)

Table 3.3: Simulation Parameters for Adaptive Filter-based Speech Enhancement Algorithms

Parameter	Value
Step Size μ (LMS/NLMS)	{0.1; 0.2; 0.3; 0.4}
Filter Order (LMS/NLMS)	{16; 32; 64}
Step Size μ (Kernel Methods)	{0.05; 0.01; 0.1}
Dictionary Size/Filter Order (Kernel Methods)	{16; 32; 64; 128}
Gaussian Kernel Parameter (Kernel Methods)	{0.1; 1; 2}
SNR	0 dB
Sampling Frequency	8 kHz
Noise Type	White Gaussian

enhancement."

Table 3.3 provides information about simulation parameters employed for adaptive-filter based speech enhancement algorithms, detailing key factors such as step sizes, filter orders, and other relevant parameters essential for the evaluation and comparison of these methods.

3.2 Effect of Input SNR

This section evaluates the impact of input signal-to-noise ratio (SNR) levels on algorithm performance, focusing on low SNR values such as 0 dB and 5 dB. The main goal is to analyze how the algorithm works in difficult conditions when noise significantly affects the speech signal. First, to demonstrate the effect of input SNR on clean speech, the waveform and spectrogram of both clean and distorted speech at various SNRs have been plotted.

Comparing Figure 3.1 with Figures 3.2–3.4, it becomes obvious that with SNR=0 dB, speech signal distortion is significantly higher compared to SNR=10 dB. This highlights the challenge of processing speech signals under low signal-to-noise ratio conditions.

Figure 3.5 illustrates a comparison between clean speech signal, noisy speech,

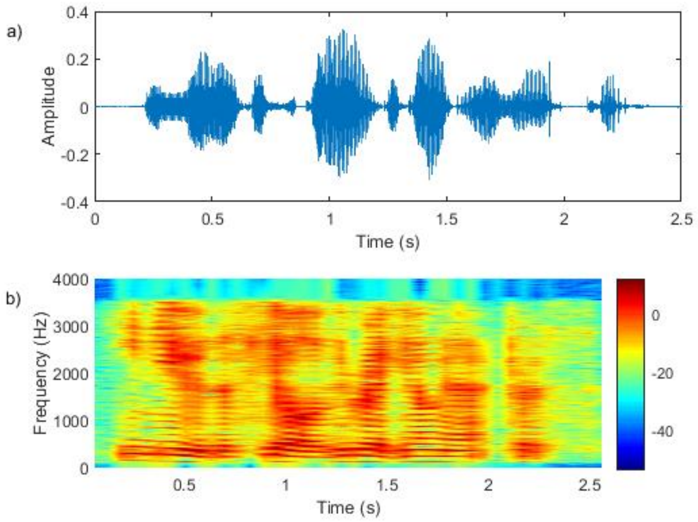

Figure 3.1: Plot of Clean Speech a) waveform and b) spectrogram

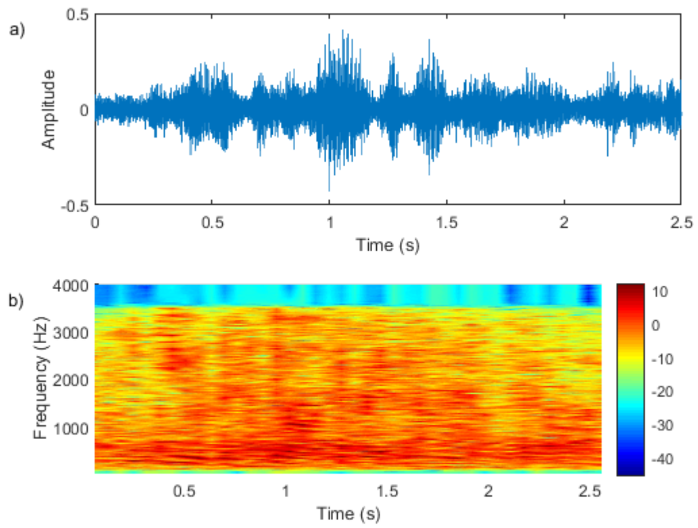

Figure 3.2: Plot of Noisy Speech at SNR=0 dB a) waveform and b) spectrogram

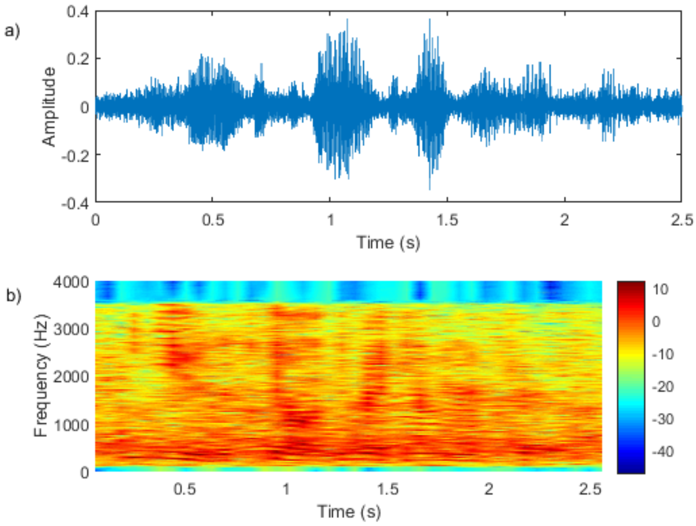

Figure 3.3: Plot of Noisy Speech at SNR=5 dB a) waveform and b) spectrogram

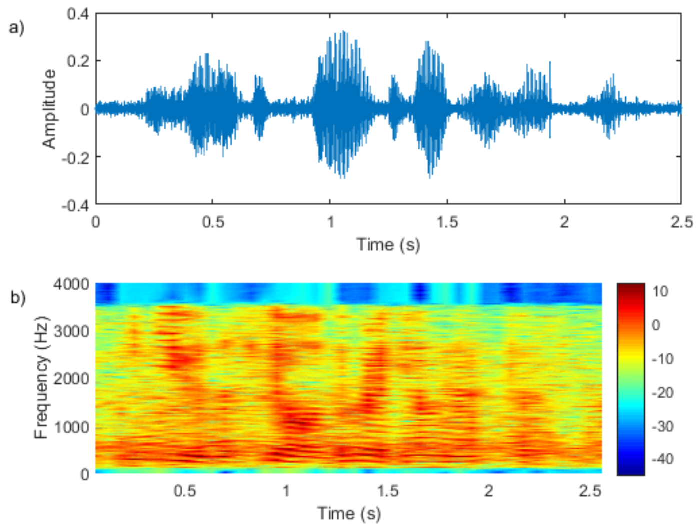

Figure 3.4: Plot of Noisy Speech at SNR=10 dB a) waveform and b) spectrogram

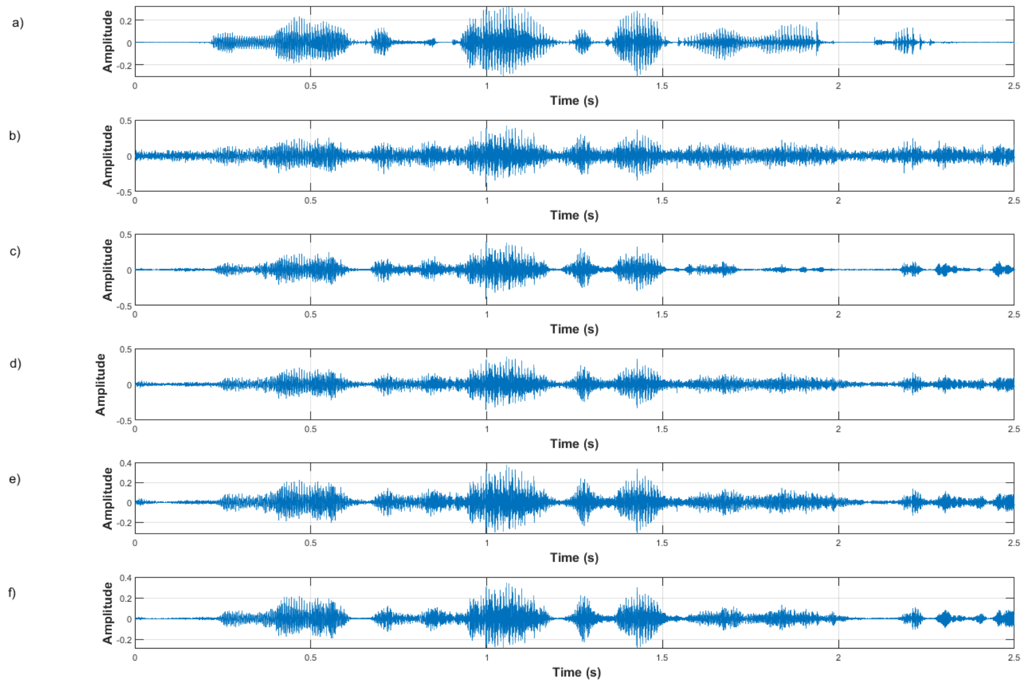

Figure 3.5: Plot of a) clean speech b) noisy speech with input SNR=0db c) enhanced speech using spectral subtraction d) enhanced speech using MMSE e) enhanced speech using log-MMSE f) enhanced speech using log-MMSE with SPU

and enhanced speech. The input signal-to-noise ratio is 0 dB. The noisy speech signal exhibits significant distortion, and the background noise is babble, which is a combination of the voices of a crowd of people, making it difficult to understand.

The improved signal shows a marked improvement over noisy speech, resembling the characteristics of clean speech. The improvement process effectively reduces the noise level. However, some residual/musical noise is still present, indicating that the enhancement process is not entirely ideal.

Overall, the comparison in Figure 3.5 demonstrates the effectiveness of enhancement algorithms in improving the quality of the speech signal, bringing it closer to the pure reference signal.

When comparing the plots in Figure 3.5 (c-f), spectral subtraction initially appears to outperform the MMSE algorithm due to less distortion between 2.0 and 2.5 seconds. However, spectral subtraction comes with a serious disadvantage: oversubtraction between 1.5 and 2.0 seconds. This oversubtraction results in the loss of important information in the signal, which significantly affects the quality

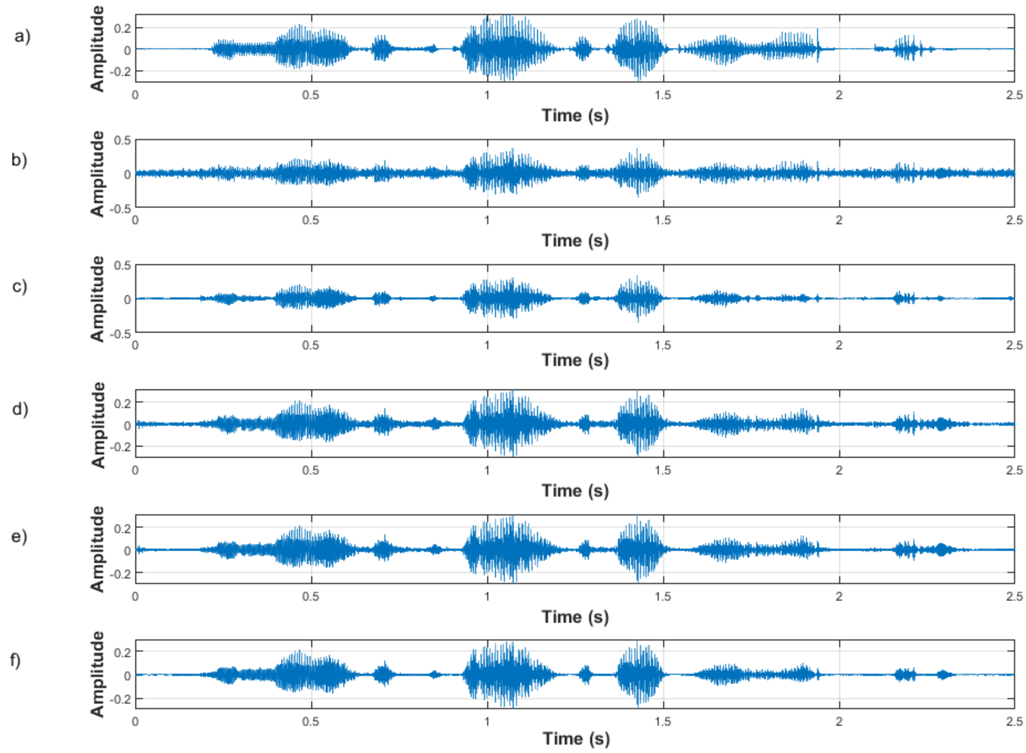

Figure 3.6: Plot of a) clean speech, b) noisy speech with input SNR=5dB, c) enhanced speech using spectral subtraction, d) enhanced speech using MMSE, e) enhanced speech using log-MMSE, f) enhanced speech using log-MMSE with SPU.

of the enhanced speech. Despite the apparent better performance in some aspects, the tendency of spectral subtraction to over-subtract in certain segments highlights a limitation compared to the MMSE algorithm. When comparing MMSE-based speech enhancement algorithms in Figure 3.5(d-f), the performance of the log-MMSE based algorithms outperforms that of simple MMSE estimation. The log-MMSE algorithm exhibits improved noise reduction and speech feature preservation, resulting in a clearer, more intelligible speech signal, especially noticeable between 2.0 and 2.5 seconds, when noise artifacts are significantly reduced. Moreover, among the three algorithms compared, log-MMSE using SPU shows the best performance. This algorithm exhibits the least amount of distortion, indicating its ability to effectively suppress noise while maintaining speech clarity and quality.

As the input SNR level increases, all algorithms show improved performance, as evidenced by clearer and less distorted speech signals in the graphs. This im-

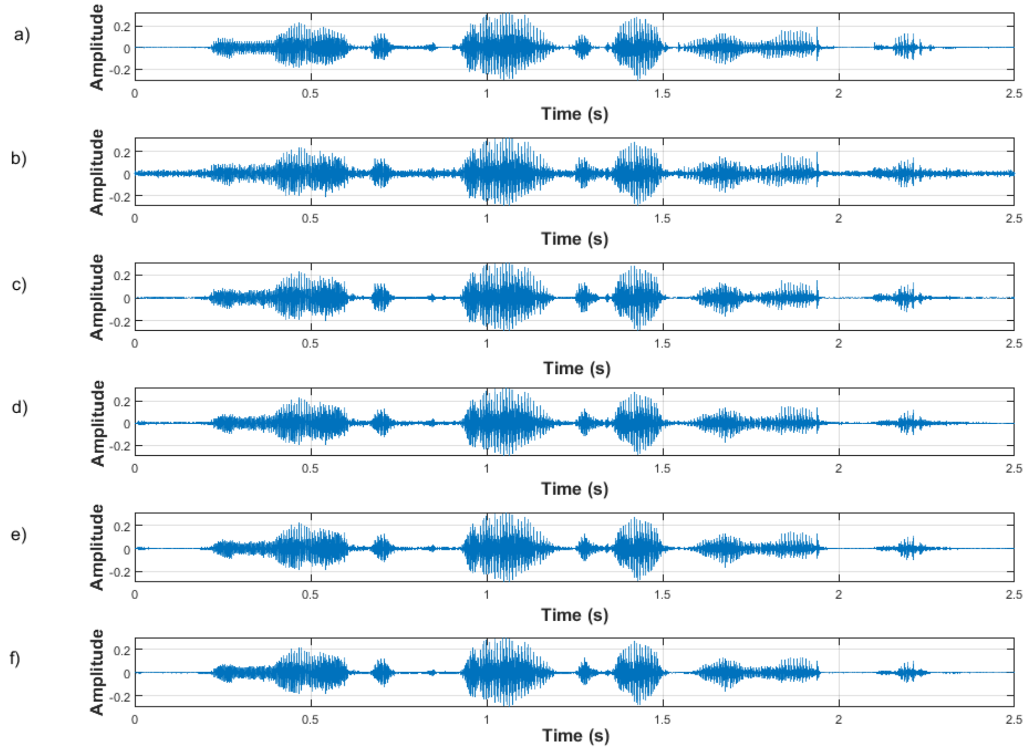

Figure 3.7: Plot of a) clean speech b) noisy speech with input SNR=10db c) enhanced speech using spectral subtraction d) enhanced speech using MMSE e) enhanced speech using log-MMSE f) enhanced speech using log-MMSE with SPU

provement is expected as higher SNR levels provide a clearer signal for the algorithms to operate on, resulting in more effective noise reduction.

From the graphs (Fig. 3.5-3.8) it can be seen that as the SNR level increases, the amount of distortion in the improved speech signals decreases. This reduction in distortion is especially noticeable in the spectral subtraction algorithm, which shows less overestimation compared to lower SNR levels. This indicates that at higher SNR levels, spectral subtraction allows better discrimination between noise and speech components, resulting in more accurate estimation and hence less distortion in the enhanced speech signal.

Overall, the results show that as the input SNR level increases, all algorithms are able to generate improved speech signals with less distortion and improved quality, with spectral subtraction showing better performance in terms of overestimation.

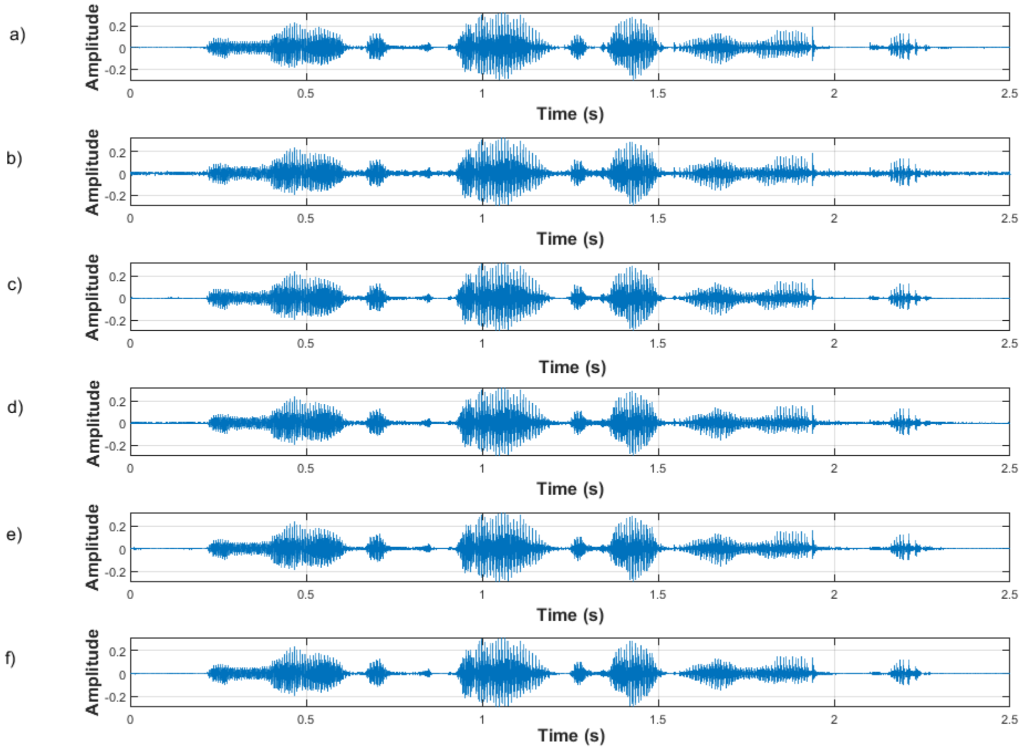

Figure 3.8: Plot of a) clean speech b) noisy speech with input SNR=15db c) enhanced speech using spectral subtraction d) enhanced speech using MMSE e) enhanced speech using log-MMSE f) enhanced speech using log-MMSE with SPU

Despite the overall improvement in performance at high SNR levels, a common problem observed in the spectral subtraction and MMSE algorithms is loss of speech between 2.0 and 2.5 seconds (Figure 3.8). This indicates that these algorithms require further improvements to better preserve speech and improve output quality. Speech loss in this segment suggests that the algorithms may be too aggressive in removing noise, resulting in unintentional suppression of speech components. To address this issue, modifications to the estimation algorithms and noise reduction mechanisms may be required. Additionally, the inclusion of more sophisticated speech enhancement techniques, such as deep learning-based approaches, can potentially improve the ability of algorithms to distinguish between noise and speech components, thereby reducing speech loss and increasing the overall quality of the enhanced speech signal.

Figure 3.9 shows that most LMS-based algorithms demonstrate similar perfor-

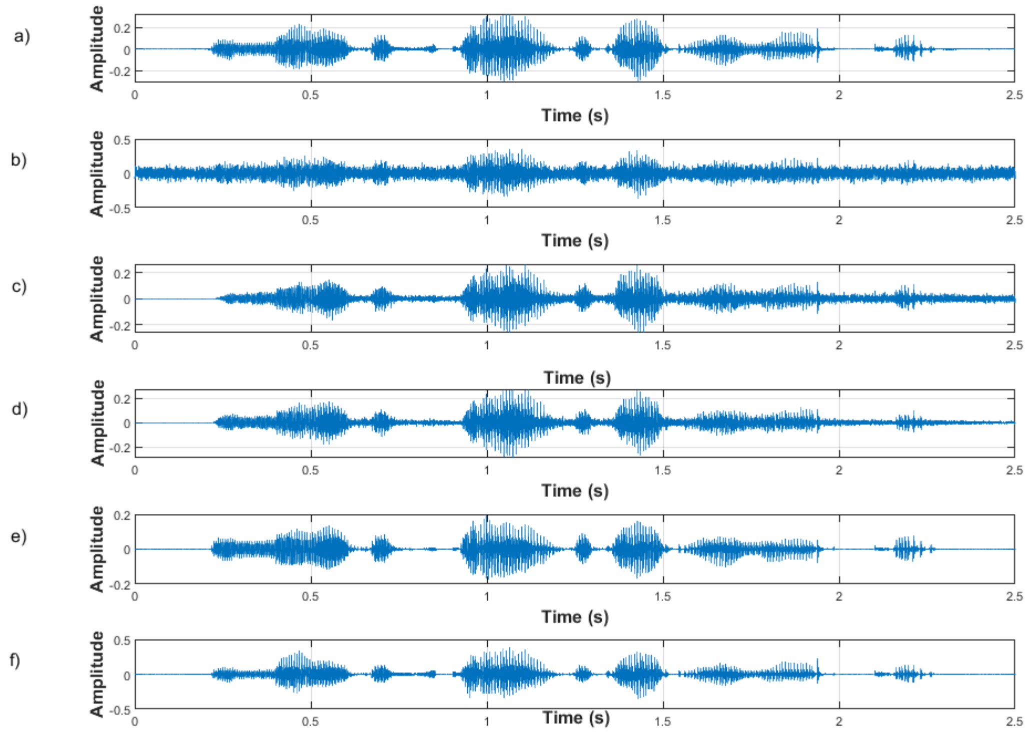

Figure 3.9: Plot of a) clean speech b) noisy speech with input SNR=0db c) enhanced speech using LMS d) enhanced speech using NLMS e) enhanced speech using KLMS f) enhanced speech using NKLMS

mance with minimal loss of speech signal. This indicates that these algorithms are effective in improving speech signals while maintaining speech integrity. The similarity observed among the LMS-based algorithms in Figure 3.9 is directly related to the fact that the noise is white Gaussian. White Gaussian noise is characterized by equal intensity at different frequencies, resulting in a constant power spectral density. This uniform distribution of noise across frequencies allows LMS-based algorithms to efficiently adapt and remove noise, resulting in similar performance across different variants of the LMS algorithm.

However, it is important to note that real noise is often non-stationary, meaning its characteristics change over time. Unlike white Gaussian noise, which is relatively easy to model and remove, non-stationary noise presents a more serious problem. Therefore, it is not possible to compare the simulation results of traditional methods with the adaptive filter-based speech enhancement algorithm. But

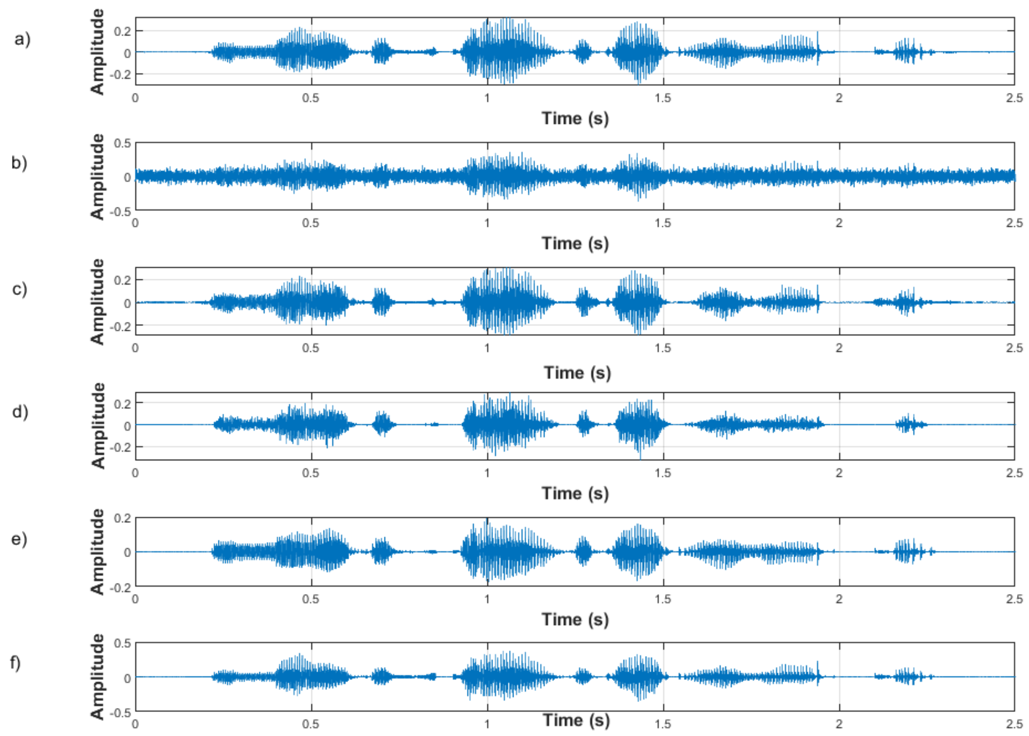

Figure 3.10: Plot of a) clean speech b) noisy speech with input SNR=5db c) enhanced speech using LMS d) enhanced speech using NLMS e) enhanced speech using KLMS f) enhanced speech using NKLMS

this suggests that an adaptive filter can be used to improve speech.

The kernel-based LMS algorithm shows superior performance compared to the simple LMS algorithm, especially between 2.0 and 2.5 seconds of the enhanced speech signal(Figure 3.9). In this segment, distortion is significantly reduced, indicating that the kernel-based approach effectively suppresses noise and preserves speech quality more effectively than the simple LMS algorithm.

By comparing Figures 3.9 and 3.10, it can be concluded that as the SNR increases, the performance of the algorithms improves. This improvement is evident in the clearer and less distorted speech signals at higher SNR levels, indicating that the algorithms are more effective in reducing noise and preserving speech quality when the input signal is less corrupted by noise.

3.3 Objective quality evaluation

Table 3.4: PESQ value for Conventional Speech enhancement methods

#	Method	0 dB	5 dB	10 dB	15 dB
1	Log-MMSE with SPU	1.944	2.259	2.573	2.990
2	Log-MMSE	1.915	2.197	2.495	3.002
3	MMSE	1.826	2.075	2.404	2.860
4	Spectral Subtraction	1.544	2.004	2.371	2.756

PESQ served as the objective quality evaluation metric for the enhanced speech signals. It's a recognized industry standard for audio quality that takes into considerations characteristics such as: audio sharpness, call volume, background noise, clipping, audio interference ect. PESQ returns a score between -0.5 and 4.5 with the higher scores indicating a better quality.

Table 3.4 provides a detailed overview of the PESQ values obtained for various speech enhancement algorithms, including spectral subtraction, MMSE, Log-MMSE, and Log-MMSE with SPU. These values are calculated as averages for different types of noise, giving a comprehensive assessment of the performance of each algorithm under different acoustic conditions.

From the PESQ values at 0 dB and 5 dB SNR levels, it is clear that Log-MMSE with SPU gives better results compared to other algorithms. This indicates that incorporating speech presence uncertainty (SPU) into the Log-MMSE algorithm significantly improves its ability to preserve speech quality and suppress noise, especially at lower SNR levels.

In contrast, spectral subtraction exhibits comparatively poor performance across the entire signal-to-noise spectrum, indicating limitations in its noise reduction capabilities. This is due to the oversubtraction that occurs during spectral subtraction. Log-MMSE-based algorithms show almost identical performance for 10 dB and 15 dB SNR input levels. This observation suggests that further improvements in SNR will not have a significant impact on the performance of these algorithms if the SNR level exceeds 10 dB.

The STOI is a evaluation metric used for assessing speech intelligibility. It quantifies the degree of intelligibility of the processed speech signal by comparing it with a clean reference signal. STOI works by analyzing short segments of processed and clean signals, measuring their similarity in amplitude and phase characteristics. By comparing these short segments, the STOI can produce a numerical score that reflects the overall intelligibility of the processed speech.

The STOI score ranges from 0 to 1, with a higher score indicating better quality of improved speech. A score closer to 1 means that the processed speech is more intelligible and closer in quality to the clean reference signal.

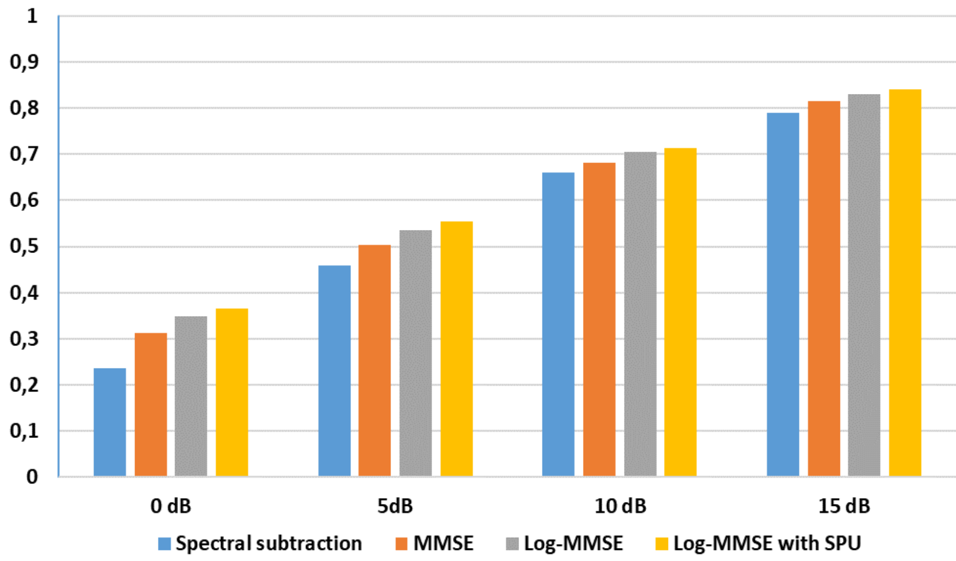

Figure 3.11: Comparison of STOI value

From Figure 3.11, it can be seen that speech quality improves significantly when using MMSE-based algorithms, as evidenced by their similar and relatively high performance levels. However, these algorithms have limitations, especially at low SNR levels where their performance degrades. This limitation is directly related to the assumption made by these algorithms that speech and noise are not correlated and that noise is stationary, which may not be true in real-world scenarios characterized by non-stationary and correlated noise.

In comparison, MMSE-based algorithms outperform spectral subtraction, indicating the effectiveness of MMSE-based approaches in speech improvement. This superiority is further supported by the PESQ and STOI values, which consistently show that log-MMSE-based algorithms produce better results than both spectral subtraction and simple MMSE algorithms. These results highlight the importance of using advanced algorithms such as log-MMSE to effectively address real-world noise characteristics and improve speech quality in noisy environments.

Chapter 4

Conclusion

The Capstone Project conducted a comprehensive study of speech enhancement techniques using MATLAB. This involved practical implementation and careful evaluation of several algorithms. Specifically, spectral subtraction has been implemented and evaluated using measures such as PESQ and STOI. In addition, the project included the implementation of MMSE and its various adaptations, as well as the implementation of speech enhancement methods based on adaptive filters. These efforts aimed to improve speech quality and intelligibility in various noisy environments, demonstrating the practical application and effectiveness of these algorithms in real-world scenarios.

Overall, the goals of the Capstone project were successfully achieved through the implementation and evaluation of various speech enhancement algorithms. However, to further improve performance and explore new opportunities for speech enhancement, it is planned to introduce approaches based on machine learning and adaptive filters.

Machine learning and adaptive filters has shown great promise in various signal processing tasks, including speech enhancement. By using machine learning algorithms such as deep learning models, even better results can be achieved in terms of noise reduction and speech quality improvement. These models can learn complex patterns from data and adapt to different noise environments, potentially outperforming traditional signal processing methods. Also, adaptive filters have high potential as it adapts to the environmental noise and can solve nonlinear problems.

Thus, the decision to explore machine learning-based speech enhancement for future work is motivated by the potential for improved performance and the opportunity to contribute to the development of speech processing technologies.

Reference

[1] Tian Lan et al. “Multi-scale informative perceptual network for monaural speech enhancement”. In: *Applied Acoustics* 195 (2022), p. 108787. ISSN: 0003-682X. DOI: 10.1016/j.apacoust.2022.108787. URL: https://www.sciencedirect.com/science/article/pii/S0003682X2200161X.

[2] Amol Chaudhari and S. B. Dhonde. “A review on speech enhancement techniques”. In: *2015 International Conference on Pervasive Computing (ICPC)*. 2015, pp. 1–3. DOI: 10.1109/PERVASIVE.2015.7087096.

[3] Anuradha R. Fukane and Shashikant L. Sahare. “Enhancement of Noisy Speech Signals for Hearing Aids”. In: *2011 International Conference on Communication Systems and Network Technologies*. 2011, pp. 490–494. DOI: 10.1109/CSNT.2011.105.

[4] Mehdi Yektaeian and Rassul Amirfattahi. “Comparison of spectral subtraction methods used in noise suppression algorithms”. In: *20 07 6th International Conference on Information, Communications & Signal Processing*. 2007, pp. 1–4. DOI: 10.1109/ICICS.2007.4449542.

[5] Fei Ge. “Brief Review of Recent Researches in Speech Enhancement from Filters to Neural Networks”. In: *2020 International Conference on Computing and Data Science (CDS)*. 2020, pp. 260–264. DOI: 10.1109/CDS49703.2020.00059.

[6] Yuxuan Wang, Arun Narayanan, and DeLiang Wang. “On Training Targets for Supervised Speech Separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.12 (2014), pp. 1849–1858. DOI: 10.1109/TASLP.2014.2352935.

[7] VS Vivek, S Vidhya, and P Madhanmohan. “Acoustic scene classification in hearing aid using deep learning”. In: *2020 International Conference on Communication and Signal Processing (ICCSP)*. IEEE. 2020, pp. 0695–0699.

[8] Naga Sandhya Devi Ganta and Vinay Kumar Mittal. “Enhancing Speech In Noisy Environment: A Review”. In: *2021 IEEE Region 10 Symposium (TENSYMP)*. 2021, pp. 1–8. DOI: 10.1109/TENSYMP52854.2021.9550835.

[9] Soha A. Nossier et al. "Mapping and Masking Targets Comparison using Different Deep Learning based Speech Enhancement Architectures". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. 2020, pp. 1–8. DOI: 10.1109/IJCNN48605.2020.9206623.

[10] Yan Zhao et al. "DNN-based enhancement of noisy and reverberant speech". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 6525–6529.

[11] Yan Zhao et al. "DNN-based enhancement of noisy and reverberant speech". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016, pp. 6525–6529. DOI: 10.1109/ICASSP.2016.7472934.

[12] Marvin Coto-Jimenez et al. "Hybrid Speech Enhancement with Wiener filters and Deep LSTM Denoising Autoencoders". In: *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*. 2018, pp. 1–8. DOI: 10.1109/IWOBI.2018.8464132.

[13] Huajian Fang et al. "Integrating Statistical Uncertainty into Neural Network-Based Speech Enhancement". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 386–390. DOI: 10.1109/ICASSP43922.2022.9747642.

[14] Yan-Hui Tu et al. "A Hybrid Approach to Combining Conventional and Deep Learning Techniques for Single-Channel Speech Enhancement and Recognition". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 2531–2535. DOI: 10.1109/ICASSP.2018.8461944.

[15] Hao Zhang et al. "Deep Adaptive Aec: Hybrid of Deep Learning and Adaptive Acoustic Echo Cancellation". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 756–760. DOI: 10.1109/ICASSP43922.2022.9746039.

[16] Douglas O'Shaughnessy. "Speech Enhancement—A Review of Modern Methods". In: *IEEE Transactions on Human-Machine Systems* 54.1 (2024), pp. 110–120. ISSN: 2168-2305. DOI: 10.1109/THMS.2023.3339663.

[17] Donald S. Williamson. "MONAURAL SPEECH SEPARATION USING A PHASE-AWARE DEEP DENOISING AUTO ENCODER". In: *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. 2018, pp. 1–6. DOI: 10.1109/MLSP.2018.8516918.

[18] Siarhei Y. Barysenka and Vasili I. Vorobiov. "SNR-Based Inter-Component Phase Estimation Using Bi-Phase Prior Statistics for Single-Channel Speech Enhancement". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 2365–2381. DOI: 10.1109/TASLP.2023.3284514.

[19] Xintao Liang et al. "A Dual Stream Generative Adversarial Network with Phase Awareness for Speech Enhancement". In: *Information* 14.4 (2023), p. 221.

[20] Huy Nguyen et al. "Phase-Aware Speech Enhancement With Complex Wiener Filter". In: *IEEE Access* 11 (2023), pp. 141573–141584. DOI: 10.1109/ACCESS.2023.3341919.

[21] Amarendra Jadda and Inty Santi Prabha. "Speech enhancement via adaptive Wiener filtering and optimized deep learning framework". In: *International Journal of Wavelets, Multiresolution and Information Processing* 21.01 (2023), p. 2250032.

[22] Yan-Hui Tu et al. "A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition". In: *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 2531–2535.

[23] Yong Xu et al. "An Experimental Study on Speech Enhancement Based on Deep Neural Networks". In: *IEEE Signal Processing Letters* 21.1 (2014), pp. 65–68. DOI: 10.1109/LSP.2013.2291240.

[24] Hongjiang Yu et al. "A deep neural network based Kalman filter for time domain speech enhancement". In: *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2019, pp. 1–5.

[25] Jonah Casebeer, Nicholas J Bryan, and Paris Smaragdis. "Meta-af: Meta-learning for adaptive filters". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2022), pp. 355–370.

[26] Sujan Kumar Roy, Aaron Nicolson, and Kuldip K Paliwal. "Deep learning with augmented Kalman filter for single-channel speech enhancement". In: *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2020, pp. 1–5.

[27] Philipos C Loizou. *Speech enhancement: theory and practice*. CRC press, 2007.

[28] Vinícius Almeida Dos Santos et al. "Improving speaker recognition in environmental noise with adaptive filter". In: *IEEE Access* 10 (2022), pp. 124523–124533.

[29] Dirk TM Slock. "On the convergence behavior of the LMS and the normalized LMS algorithms". In: *IEEE Transactions on Signal processing* 41.9 (1993), pp. 2811–2825.

[30] Steven Van Vaerenbergh and Ignacio Santamaría. "A comparative study of kernel adaptive filtering algorithms". In: *2013 IEEE Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)*. IEEE. 2013, pp. 181–186.

[31] Hamed Modaghegh et al. "A new modeling algorithm-normalized kernel least mean square". In: *2009 International Conference on Innovations in Information Technology (IIT)*. IEEE. 2009, pp. 120–124.