DEVELOPMENT OF NOVEL SPEECH ENHANCEMENT TECHNIQUES UNDER LOW SNR ENVIRONMENTS

Submitted in partial fulfilment of the
requirementsfor the award of the degree of

DOCTOR OF PHILOSOPHY

by

VENKATA SRIDHAR KONERU

(Roll No. 718168)

Under the Supervision of

Prof. T. KISHORE KUMAR

Professor

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY
WARANGAL – 506004, T.S, INDIA

November – 2022

Dedicated to my beloved

Wife, Children

and Teachers,

APPROVAL SHEET

This thesis entitled "**Development of Novel Speech Enhancement Techniques Under Low SNR Environments**" by **Mr. Venkata Sridhar Koneru** is approved for the degree of **Doctor of Philosophy**.

Examiners

_____

_____

Supervisor

_____

Prof. T. Kishore Kumar

Professor, Electronics and Communication Engineering
Department,NIT WARANGAL

Chairman

_____

Prof. P. Srihari Rao

Head, Electronics and Communication Engineering
Department,NIT WARANGAL

Date:

Place: Warangal

DECLARATION

I, hereby, declare that the matter embodied in this thesis entitled "**Development of Novel Speech Enhancement Techniques Under Low SNR Environments"** is based entirely on the results of the investigations and research work carried out by me under the supervision of **Prof. T. Kishore Kumar**, Department of Electronics and Communication Engineering, National Institute of Technology Warangal. I declare that this work is original and has not been submitted in part or full, for any degree or diploma to this or any other University.

I declare that this written submission represents my ideas in my own words and where other ideas or words have been included. I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/date/fact/source in my submission. I understand that any violation of the above will cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Venkata Sridhar Koneru

Roll No: 718168

Date:

Place: Warangal

Department of Electronics and Communication

EngineeringNational Institute of Technology

Warangal – 506 004, Telangana, India

CERTIFICATE

This is to certify that the dissertation work entitled **"Development of Novel Speech Enhancement Techniques Under Low SNR Environments"**, which is being submitted by Mr. Venkata Sridhar Koneru (Roll No. 718168), a bonafide work submitted to National Institute of Technology Warangal in partial fulfilment of the requirement for the award of the degree of Doctor of Philosophy to the Department of Electronics and Communication Engineering of National Institute of Technology Warangal, is a record of bonafide research work carried out by her under my supervision and has not been submitted elsewhere for any degree.

Prof. T. Kishore Kumar
(Supervisor)
Professor, Department of ECE
National Institute of Technology
Warangal, India – 506004

ACKNOWLEDGEMENTS

I would like to thank several people who have contributed to my Ph.D. directly or indirectly and in different ways through their help, support, and encouragement.

It gives me immense pleasure to express my deep sense of gratitude and thanks to my supervisor **Prof. T. Kishore Kumar** (National Institute of Technology Warangal, NITW), for his invaluable guidance, support, and suggestions. His knowledge, suggestions, and discussions helped me to become a capable researcher. He has shown me the interesting side of this wonderful multidisciplinary area and guided me to get profound knowledge as well as publications in this area.

I am thankful to the present Head of the Dept. of E.C.E., **Prof. P Srihari Rao**, and the former Head, **Prof. N. Bheema Rao** for giving me the opportunity and all the necessary support from the department to carry out my research work.

I take this privilege to thank all my Doctoral Scrutiny Committee members, Dr. **D. M Vinod Kumar**, Professor, Dept. of EEE, **Dr. L. Anjaneyulu**, Professor, Dept. of ECE, **Dr**. **N. Bheema Rao**, Professor, Dept. of ECE and **Dr. V.V. Mani** Professor, Dept. of ECE for their detailed review, constructive suggestions, and excellent advice during the progress of this research work.

I take this opportunity to convey my regards to my speech lab-mates, NIT Warangal, K Sunil Kumar, and Govind for being always present next to me in my time of need.

I also appreciate the help rendered by teaching, non-teaching members, and the fraternity of Dept. of E.C.E. of N.I.T. Warangal. They have always been encouraging and supportive.

I acknowledge my gratitude to all my teachers and colleagues in various aspects for supporting and cooperating to complete this work.

Finally, I appreciate and respect my family members (my Wife Smt. K. Sripadamaja,

my elder daughter and Son-in-law: Mrs. Gayathri Devi Mr. Hema Reddy, my younger daughter and Son-in-law: Mrs. Jahnavi Sai, Mr. Ravi Raja, and my beloved grandchildren: Akshaya Siri, Viraj Raman) for being very supportive while giving me mental support and inspiration that motivated me to complete the thesis work successfully. Especially, I thank Prof. Raja Viswanathan who has been a strong support throughout my Ph.D. period and also helped me as an English professional to proofread my publications as well as my thesis.

VENKATA SRIDHAR KONERU

ABSTRACT

Speech enhancement (SE) is important when the speech signal is degraded by real-world background noise. Degradation impacts the quality and intelligibility of speech and decreases the performance of speech processing systems. The main goal of the SE method is to reduce or suppress the noise from the degraded speech signal while preserving the perceptual quality of speech (PESQ) and intelligibility (STOI) with the least distortion. Noise estimation is a crucial stage in SE, and it commonly necessitates the use of prior models for speech, noise, or both. Prior models, on the other hand, can be ineffective in dealing with nonstationary noise, especially at low signal-to-noise (SNR) levels. Estimating noise-related parameters in SE techniques is challenging task in low SNR and nonstationary noise environments. Existing SE techniques are limited to various types of stationary noise and lack robustness to some forms of real-world noise, which are in general nonstationary. These techniques are able to improve speech quality while introducing considerable signal distortion, leading to poor intelligibility. Specifically, due to the enhancement process, these techniques introduce significant distortion under strong noisy conditions, i.e., low SNR (< 0 dB). Signal subspace decomposition methods for denoising offer superior performance interms of low signal distortion and residual noise.

Recently, sparse coding has gained popularity in signal denoising. A sparse representation decomposes a signal into a small set of over-complete (dictionary) components tailored to the processed data. Determining the best appropriate criterion to identify the principal components (called atoms) from the learnt dictionary to generate the principal (signal) subspace and removing the other is difficult for signal subspace decomposition based on a sparse representation. Dictionary learning based on compressive sensing is hard due to insufficient training a noise dictionary. The performance degrades in real world scenarios due to mismatch between unknown background noise and the training noises. Therefore, unsupervised models are needed to enhance the noisy speech. Estimating noise-related parameters in unsupervised SE techniques is challenging in low SNR and nonstationary noise environments. In the recent SE approaches, best results are achieved by partitioning noisy speech spectrogram into low rank noise and sparse speech parts.

This research investigates an unsupervised SE strategy, Robust Principal Component analysis (RPCA) to estimate noise and speech when neither is available beforehand by decomposing the input noisy spectrum into low-rank noise and sparse speech components. Due to approximation of rank of the noise, these strategies are limited and don't directly use low-rank in optimization. Nuclear norm minimization (NNM) can recover the matrix's rank under specific theoretical guarantee conditions. In many instances, NNM can't correctly predict matrix rank. Weighted nuclear norm minimization (WNNM) addresses NNM's drawbacks and delivers a better matrix rank approximation than NNM. In this work, a weighted low rank and sparsity constraints is employed to differentiate speech and noise spectrograms. However, few limitations reduce the performance of these SE methods due to the use of: overlap and add in STFT process, noisy phase, due to inaccurate estimation of low- rank in nuclear norm minimization and Euclidian distance measure in cost function. These aspects can cause loss of information in the reconstructed signal when compared to clean speech. To solve this, a novel wavelet-based weighted low-rank sparse decomposition model is developed for enhancing speech by incorporating a gamma-tone filter bank and Kullback Leibler Divergence(KLD). The proposed framework differs from other strategies in which the SE is carried entirely in Time-domain without the need for noise estimation. The experimental findings using Noizeus speech corpus indicate that the proposed integrated model has shown significant improvement under low SNR conditions over individual and traditional methods with regard to objective metrics such as SDR, PESQ, STOI, SIG, BAK and OVL. Further, to reduce the Word Error Rate, these algorithms were trained and tested on a typical Automatic Speech Recognition module.

Finally, the enhanced speech signals are trained and assessed using standard khaldi automatic speech recognition (ASR) engine to lower Word Error Rate (WER). Extensive studies on the impact of real-world noise on speech signals reveal that the proposed model surpasses the existing state-of-the-art methods, resulted in improving the recognition accuracy in terms of WER.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ALM - Augmented Lagrange multiplier

BAK- Background-noise intrusiveness

BSS_Eval – Blind source separation _ evaluation

CS- Compressive Sensing

DCT – Discrete Cosine Transform

DFT – Discrete Fourier Transform

DWT- Discrete Wavelet Transform

ED- Euclidian Distance

e.g.- exempli gratia (latin: for example)

et al.- et alii (latin: and others)

FFT – Fast Fourier Transform

GMM – Gaussian Mixture Model

Go-Dec-Go Decomposition

ICA – Independent Component Analysis

i.e.-id est (latin: that is)

ISTFT-Inverse Short Time Fourier Transform

KLD- kullback-leibler divergence

K-NN – K-Nearest Neighborhood

K-SVD- K-means Singular Value Decomposition

LPC – Linear Prediction Coefficient

MSE- Mean Square Error

NMF – Non-Negative Matrix Factorization

OVL- Overall quality

PCA – Principal Component Analysis

PCP-Principal Component Pursuit

PDF- probability density function

PESQ- Perceptual Evaluation of Speech Quality

RPCA-Robust Principal Component Analysis

SDR- Signal to Distortion Ratio

SE - Speech Enhancement

SIG - Signal Distortion

SNR – Signal to Noise Ratio

SS-Godec – Semi- Soft Go Decomposition

STFT-Short Time Fourier Transform

STOI- Short-Time Objective Intelligibility

SVD – Singular Value Decomposition

WDO-Windowed disjoint Orthogonality

WKLWNNM Wavelet based kullback-leibler divergence WNNM

WNNM- Weighted Nuclear Norm

WSNM- Weighted Schatten P-Norm

CONTENTS

Chapter 1

INTRODUCTION

Speech is an effective way to express thoughts, as well as one's feelings and interests. It is one of the most typical forms of human communication. Due to technological advancements speech communication is currently performed not only in person-to-person interactions but may also be done across long distances, such as through telecommunications, or it can even be used as a natural manner of human-machine connection. As computationally complex computer hardware has been readily accessible to users, speech processing devices like smartphones, tablets, and notebooks have become very common. As a result, speech has a big impact on a lot of applications, like hands-free phones, digital hearing aids, speech-based computer interfaces, home entertainment systems, low-quality audio devices, ASR systems, etc.

However, the presence of undesirable real-world background noise significantly reduces speech quality and intelligibility in many speech processing systems. As a result, they have poor performance, which makes communication difficult and limits their use. Enhancing speech degraded by background noise is both a necessary and difficult task. The difficulty in improving speech quality in these applications arises from the nature of the noise encountered, which is frequently non-stationary and probably speech-like, thereby inducing a noticeable and time-varying spectral overlap between speech and noise.

High quality of speech perceived as being more pleasant to listen to for prolonged periods, while intelligibility of speech is measured by minimal word error rates in speech recognition scenarios. Speech enhancement (SE) is a very hard problem to solve if we don't know anything about the noise signal we're trying to remove. The majority of traditional speech enhancement techniques are constrained to stationary noise, limited to noise with specific characteristics like white/pink, and as a result, denoising has the consequence of degrading the speech signal's intelligibility. Due to this, the performance of SE methods entails a trade-off between noise

reduction and the intelligibility of speech signals. Understanding speech in the presence of these noises is dependent on Signal to Noise Ratio (SNR) and the type of noise. It gets harder to perceive speech that is significantly masked by noise as the signal-to-noise ratio lowers (i.e. as noise becomes more dominating). It's interesting to note that while most listening scenarios have moderate noise levels, only around one in ten involves dominant noise. Therefore, developing a successful SE method becomes extremely difficult, especially when dealing with a variety of non-stationary noises that are typically present in real-world scenarios under low SNR settings.

The focus of this research is to develop and analyse new low-rank sparse decomposition techniques for speech enhancement that address various issues with existing techniques. This chapter begins by giving a brief overview of speech enhancement, the principles of hearing and generating speech, applications, and classification of speech enhancement techniques. Next, a detailed description of the low-rank sparse decomposition approach is given, which serves as the theoretical foundation for the algorithms developed in the following chapters. A brief description of the methods currently employed in low-rank sparse decomposition algorithms for speech enhancement under low SNR is provided. The windowing and overlapping techniques are then introduced, which is used to pre-process speech signals for improvement.

The motivation for low-rank sparse decomposition techniques is presented, followed by problem statement, research objectives, and finally organization the thesis.

1.1 Essentials of Speech Signal

It is important to understand the speech signal before addressing speech enhancement. It is essential to understand the qualities and characteristics of the speech signal in order to develop efficient speech enhancement strategies that preserve high quality and intelligibility in noisy conditions with least amount of distortion.

It is helpful to understand the properties and characteristics of the speech signal. This section provides a brief description of the human speech production mechanism and hearing process.

1.1.1 Speech Production Mechanism

The acoustic output of voluntarily regulated movements of the respiratory and masticatory systems is speech.

Figure 1.1: Schematic diagram of the human speech production mechanism

The auditory result of deliberate and structured movements of the respiratory and masticatory systems is speech. Figure 1.1 depicts the parts of the speech production system schematically. The illustration shows a mid-sagittal segment of an adult's vocal tract. Expanding the rib cage, lowering lung pressure, and pulling air into the lungs through the nostrils, nasal cavity, velum port, and trachea are the main goals of inhalation (windpipe). Normally, the air is ejected along the same path. Mastication occurs in the oral cavity when eating. The structures at the tracheal entry are pushed up under the epiglottis when food is swallowed. The latter protects the voice cord aperture and keeps food from entering the windpipe. In order to provide a passageway

for the stomach, the esophagus, which ordinarily lays collapsed against the rear wall of the throat, is simultaneously pulled open. The actual vocal tract is a nonuniform cross-sectional area acoustical tube. At one end, it is terminated by the lips, and at the other, by the constriction of the vocal cords at the top of the trachea. An additional channel for sound transmission is the nasal tract. It starts at the nasal velum and ends beyond. The nasal septum divides the cavity along some of its front-to-back extents. The size of the opening at the velum affects the acoustic connection between the nasal and vocal tracts. The vellum is depicted in Fig. 1.1 at its widest opening. The sound may emanate from the mouth and nostrils in such a situation. Generally speaking, nasal coupling has a significant impact on the nature of sound emitted from the mouth. The vellum is pushed tightly up in order to produce non-nasal noises, essentially blocking the nasal cavity's entry. The thoracic and abdominal muscles provide the power needed to produce speech. By widening the chest cavity and bringing the diaphragm down, the air is pulled into the lungs. By tightening the rib cage and raising the lung pressure, it is ejected.

1.1.2 Hearing Mechanism

Hearing initially begins with the outer ear (shown in figure 1.2) called the pinna. When an external sound is made, sound waves, or vibrations, travel through the external auditory canal and strike the eardrum (tympanic membrane). It vibrates inside the ear. Three tiny bones in the middle ear called the ossicles receive the vibrations after that. Sound is amplified by the ossicles. The hearing organ, which is filled with fluid, receives the sound waves and sends them to the inner ear (cochlea). The cochlea is a coiled tube with two membranes—the basilar and Reisner’s membranes—enclosed in a fluid. When vibration is applied to the ear, the basilar membrane's width and stiffness taper along its length, and the location on the membrane resonate depending on the frequency [1]. This process gives the ear frequency selectivity, which is improved by active processing by the auditory cortex.

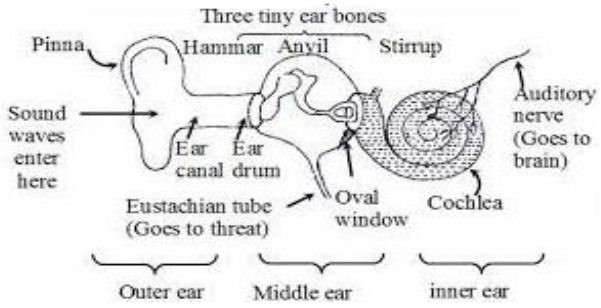


Figure 1.2 Schematic diagram of the human hearing mechanism

Upon entering the inner ear, sound waves are transformed into electrical impulses. These impulses travel to the brain via the auditory nerve [2]. When these electrical impulses reach the brain, they are translated into sound.

1.1.3 Auditory Masking

Masking, which is attributed to the mechanical vibrations of the basilar membrane, is the lowering of a listener's capacity to hear the target sound in the presence of other sounds. A listening situation known as "auditory masking" occurs when the presence of one sound prevents the detection of another occurring nearby.

Reducing the noise won't be essential if the speech component of noisy speech can hide the noise component. In this situation, more speech distortion will result from any processing. On the other hand, if the noise component cannot be masked, one must lower the noise distortion level below the masking threshold to make the noise undetectable. The masking effect of the human acoustic system is depicted in Figure 1.3, where a high-energy speech signal can increase the normal hearing threshold level in its proximity. The masking sound (red color) in its neighborhood raises the hearing threshold, and the sounds (green and blue colors) at close frequencies have been hidden. In the presence of background sound, the amount of masking is calculated by subtracting the absolute threshold from the masked threshold.

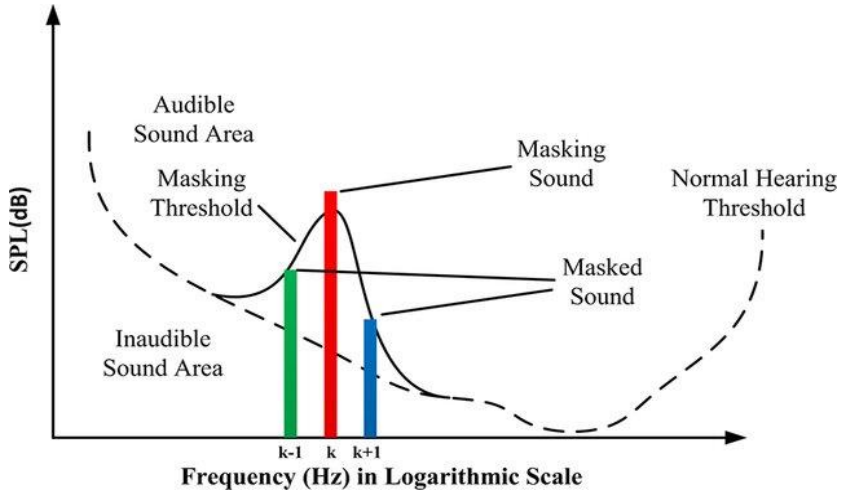


Figure 1.3 Masking characteristics of the human hearing system (Courtesy IET Journal ISSN 1751-9675)

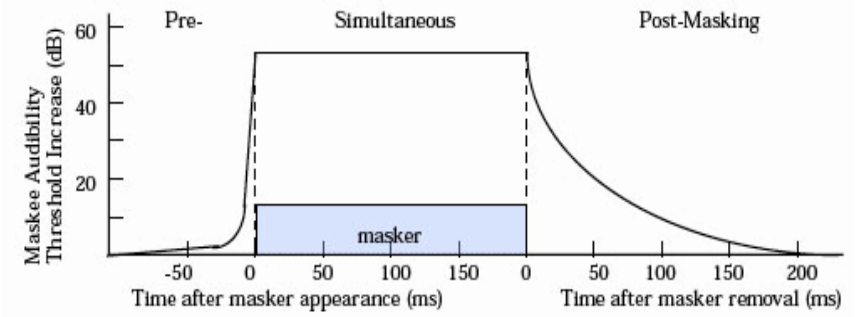


Figure 1.5 Illustration of pre, Simultaneous, and Post Masking concepts

Masking can be categorized as simultaneous or frequency masking in the frequency domain, and in time domain, as temporal masking. Frequency masking happens when the masker and the target happen at the same time while temporal masking happens when they don't happen simultaneously but are close to each other. Even though simultaneous masking is considered to have a more significant effect than temporal masking, perception is still impacted by it. Temporal masking may be further divided into pre-masking (Shown in figure 1.5), where the target appears before the masker in time for roughly 20 ms, and post-masking when the masker appears before

the signal for up to 200 ms and has a more pronounced masking effect. Most frequently, only post-masking is taken into account and the pre-masking is removed in temporal masking models since the influence of post-masking lasts longer and that of pre-masking is very minimal. Widespread use of the masking technique in applications for speech such as hearing aids, speech coding, and make speech pickup and transmission of high quality. Najafzadeh et al. [3] introduced pre and post-temporal masking models into MPEG psychoacoustic model to obtain a considerable coding gain. The majority of speech enhancement techniques do not entirely eliminate residual noise. Algorithms, therefore, try to increase the residual noise perceptually undetectable and increase speech intelligibility by utilizing the qualities of the auditory system of humans.

1.1.4 Applications of Speech Enhancement Techniques

The most important aspect of speech enhancement is noise reduction, which is utilized in a variety of real-world applications, including telecommunication systems, Automatic speech recognition, electronic hearing aids, etc. The predominant use of speech enhancement is in telecommunications systems, but there are several other applications as well, including hearing aids, the restoration of damaged audio recordings, etc. This section discusses some of the most significant applications to highlight the significance of speech enhancement in our daily lives.

1.1.4.1 Telecommunication

The field of telecommunications uses a lot of speech enhancement techniques however, they struggle to perform in noisy conditions. Mobile communication typically takes place in a variety of noisy situations including traffic and cars, crowds, trains, airports, streets, etc., making it uncomfortable and unpleasant for the users. The noise and distortion in coded speech are caused by noisy environments. For many voice communication systems to work properly, speech restoration blocks are necessary. Reduced background noise and easier conversational flow may result from the integration of a speech enhancement algorithm into the communication system [4-6].

1.1.4.2 Electronic Hearing aids

The primary function of the hearing aid is to improve the user's ability to understand speech. Despite numerous advancements in hearing aid technology, users with cochlear prostheses, sensorineural impairment, and other sensory aids, where it is impractical to employ a second microphone to provide reference input for noise suppression through adaptive filtering, experience great difficulty in speech perception in noisy environments. Along with the targeted speech, the microphone tries to pick up noise signals. For hearing-impaired listeners using hearing aids, the amplified speech must be free from any undesirable noises or disturbances which, if heard, would negate the benefits provided by the hearing aid. The noise content is also amplified along with the speech, reducing the intelligibility of the speech signal. Hence there is a need for a speech enhancement stage in the hearing aids, to filter out the noise before amplification [7,8] for improving speech quality and intelligibility.

1.1.4.3 Automatic Speech Recognition Systems

Automatic speech recognition (ASR) systems are designed to convert spoken user commands into readable text or other useful system input. The speaker needs to be some distance away from the sound-capturing equipment for ASR to be used in real-world situations like meeting transcription, customer services, education, and human-robot interaction. The target signal is further degraded by noise, which has a detrimental effect on ASR accuracy. In noisy environments, the system either misinterprets spoken words or fails to appropriately match the dictionary's list of terms. Since the system must be able to distinguish between words with similar sounds, particularly in critical applications like banking or disability assistance, high intelligibility is a necessity for these systems. Therefore, speech enhancement is used as a preprocessor to speech recognizer. Various solutions have been proposed to provide robustness in ASR: improvement of the audio signal, front-end-based approaches that enhance the signal in the feature domain, and back-end procedures (Haeb-Umbach and Krueger, 2012).

1.1.4.4 Audio surveillance and Restoration of recordings

The circumstances of audio surveillance applications are frequently harsh and seldom ever allow for microphone adjustment during recording. Speakers who are being watched don't try to speak directly and clearly into the microphone. Audio surveillance recordings are strong contenders for speech enhancement because of the multiple degradations and sources of interference that obscure the speech signals on such tapes. The use of substandard recording tools and settings also influences the quality of recorded speech because they introduce noise or hiss into the recordings. The goal of audio restoration is to eliminate any audio artifacts that were not meant to be included in the recording, such as storage media disruptions and background noise that was recorded. Speech enhancement improves speech quality and recovers data that was previously believed to be missing. The significant application in the retrieval of old magnetic tapes may contain important speeches or facts captured in poor settings.

1.2 Introduction to Speech Enhancement

The main objective of enhancing speech is to improve both the perceived quality and the intelligibility of speech, by eliminating undesirable noise, without degrading the speech substantially [9]. The speech enhancement stage is often used as a preprocessing block in several applications, including speech communication systems, automatic speech recognition, etc. Since they eliminate or reduce noise in the speech signal, speech enhancement algorithms are also referred to as noise suppression algorithms.

It is important to distinguish between speech quality and intelligibility, two terms that are frequently used interchangeably but are quite different from one another. The term "quality of speech" relates to the manner in which a speaker presents a statement and encompasses characteristics like naturalness and speaker identification [10]. Quality can be defined as a measurement of how closely the speech being examined resembles the original speech and how pleasant the speech sounds. Intelligibility is concerned with what was said, i.e., the information contained or the meaning underlying the words. It focuses on the information-carrying content of

speech and measures how easily it is understood.

In unfavorable acoustic conditions, speech communication systems' performance rapidly deteriorates. The quality and intelligibility of speech are affected by background noise. In the presence of background noise, the performance of speech communication devices, such as mobile phones, ASR, etc., which depend on speech processing systems to communicate and store speech signals, suffer noticeably, leading to inaccurate information exchange on listener fatigue between the speaker and the listener. Thus, a noisy environment limits communication between the speaker and the listeners. For example, background noise from traffic & car, crowd, machines, wind, etc. at the transmitter end makes it hard for the listener at the receiving end to understand the speaker during voice communication using cellular telephony networks. Therefore, there are many different situations wherein improving speech is desirable. The performance of other speech applications, including speech communication systems, automatic speech recognition (ASR), hearing aids, speech coding, etc., is significantly improved by enhancing the quality and/or intelligibility of noisy speech. Based on the application, the speech enhancement system has different objectives such as minimizing listener fatigue, improving overall speech quality, improving intelligibility, etc., or a mix of all of these.

A speech enhancement technique can be used to pre-process the noisy speech signal before it is fed to the speech recognition system since the recognition accuracy of an ASR will suffer in the presence of noise. In military communication systems, it is more important to improve intelligibility than quality. It is always preferred to improve noisy speech by eliminating noise before amplification for hearing-impaired listeners using hearing aids. Additionally, the design and development of the speech enhancement system are influenced by the characteristics of the noise and how it interacts with clean speech signal, such as additive, convoluted, correlated, uncorrelated, etc.

The monaural scenario lacks any prior knowledge of noise and spatial information. Due to the approximation of noise characteristics, distortion is introduced. Signal distortion is an undesirable change of the sound waveform that is being received. Another significant issue in

speech enhancement is musical noise, which consists of tones at random frequencies. The problem has been addressed by several SE methods utilizing diverse methodologies. The trade-off between noise reduction and speech enhancement limits the speech enhancement system's effectiveness. Therefore, it becomes extremely challenging to develop an efficient SE algorithm, especially in low SNR real-world environments, that can improve speech signal quality without reducing intelligibility and while minimizing distortion.

The speech processing system as shown in Figure 1.1, consists of speech pre-processing (framing and windowing), SE, and speech recognition stages.

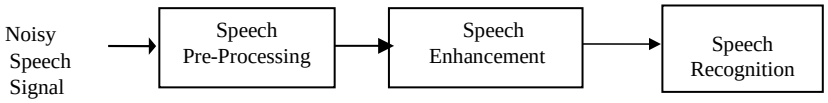


Figure 1.1: Speech Enhancement

The noisy speech frames are being processed by an SE algorithm stage to extract clean speech. The performance of the above system is influenced by various factors such as type of noise, length of the frame, hop size, type of window, and denoising capability of the SE algorithm. To examine the accuracy of recognition, the enhanced speech signals are validated by ASR stage.

1.2.1 Classes of Speech Enhancement Techniques

Speech enhancement methods can broadly be classified as single-channel and multi-channel techniques [6], depending on the number of Channels(microphones) used to acquire the acoustic signal and noise.

1.2.1.1 Single-Channel speech enhancement Systems

A single-channel speech enhancement approach is used when a single microphone is used

for the acquisition of the signal and to improve the signal. For speech enhancement in single-channel systems, it is presumed that the noisy speech signal, which consists of both clean speech and additive noise, is available from a single microphone. There is no second signal that could give details about the background noise or speech. A second microphone is typically not available in the majority of real-time applications, like speech and speaker recognition, mobile communication, and hearing aids. Since most real-world circumstances only allow for a single microphone, like speech communication, speech coding, and speech recognition in noisy environments [5], this is a topic of intense research due to its simplicity and universal applicability. In comparison to multi-input systems, these systems are comparably less expensive and simpler to implement.

Single-channel systems provide one of the most challenging situations for speech enhancement since there is no reference signal to the noise and the speech cannot be pre-processed before it is influenced by noise. Typically, they employ various speech and noise statistics. The suppression strategy a single microphone speech enhancement system employs has a significant impact on the system's quality. More signal distortion results from using a suppression rule with larger attenuation to remove a huge amount of noise. A suppression algorithm with lower attenuation, on the other hand, will result in less distorted speech signal but with only modest noise reduction. It is important to carefully balance the amount of noise suppression and distortion to get the best quality and intelligibility.

1.2.1.2 Multi-Channel Speech Enhancement Systems

The aim of multi-channel speech enhancement is to separate clean speech from a noisy mixture employing signals obtained from multiple microphones. The performance of the speech enhancement algorithms may be affected by the number of microphones available [5]. The task of improving speech is typically made simpler by using multiple microphones. When at least one microphone is put close to the source of the noise, adaptive canceling techniques can be used. In an adaptive canceling mechanism, the multi-microphone system makes use of the noise reference it has obtained. In multi-channel enhancement techniques, microphone arrays are employed to phase align and remove undesired noise components [6]. The complexity of these systems is

usually higher.

However, better speech enhancement results can be obtained by using a microphone array system, which consists of more than one microphone but is more complex and expensive. Most systems are single-microphone-based solutions because they are more cost-effective, and the output of the single microphone is where the speech enhancement is performed.

1.2.2 Classical speech Enhancement Methods

Speech enhancement is one of the classical topics of speech signal processing and numerous single microphone-based methods have been proposed.

The noisy observation can be interpreted as the summation of clean speech and non-speech interference signals in the time domain. Then, using only the noisy speech signal, the single-microphone speech enhancement method seeks to obtain an estimate of clean speech, that is, in some respects, "near to" clean speech.

1.2.2.1 Spectral Subtraction Methods

The simplest enhancement methods to use are spectral subtraction approaches, one of the earliest classes suggested for the enhancement of single-channel speech. It operates in the frequency domain and is based on the fundamental idea that the spectrum of the input signal may be described as a sum of the speech spectrum and the noise spectrum, assuming noise to be additive. Then, using a Voice Activity Detection (VAD) algorithm, for example, one may estimate the noise spectrum when speech activity is really not present and remove it from the noisy signal to get the enhanced speech signal. Weiss et al[11] proposed the first formulation of spectral subtraction algorithms in the correlation domain and Boll[ ]'s subsequent formulation in the Fourier transform domain led to the development of other versions. Even though spectral subtraction-based speech enhancement algorithms may significantly reduce speech noise, one of their major

drawbacks is that they cause signal distortion—commonly referred to as musical noise—by making inaccurate assumptions about the noise magnitude.

The characteristics of single-channel subtractive-type algorithms include a trade-off between the level of noise reduction, the degree of speech distortion, and the amount of musical residual noise. By adjusting the subtraction settings, this trade-off may be changed. The only alternative offered with classic algorithms is often the use of preset, optimal parameters, which are difficult to choose for all speech and noise conditions.

1.2.2.2 Statistical Methods

Although spectrum subtraction-based methods for speech enhancement are effective, they use heuristics to improve the entire process rather than being specifically developed to be mathematically perfect. However, a class of perfect speech enhancement algorithms may be built if the speech enhancement task is formulated as a statistical estimation problem with well-defined optimality criteria and thoroughly established statistical assumptions. One of these categories is the Minimum Mean Squared Error (MMSE) estimators, which may be further subdivided into non-linear Short-Time Spectral Amplitude (STSA)-MMSE estimators and linear MMSE estimators, also known as Wiener filters.

1.2.2.3 Weiner Filtering Algorithms

The goal of Weiner filtering, a statistical filtering technique, is to produce output speech signals that are as similar as possible to the desired speech signals. To achieve this, the estimation error is calculated and mask it as low as possible. The wiener filter, named after the mathematician Nobert Wiener, is the optimal filter that minimizes the prediction error. When the complex noise and the speech discrete Fourier transform (DFT) coefficients are considered to be independent Gaussian random variables, the Wiener filter technique produces a linear estimator of the complex spectrum of the signal and is optimum in the minimum mean square error (MMSE) sense.

1.2.2.4 Basic STSA-MMSE Estimators

The Wiener filter is the best complex spectral estimator however, it is not optimum for estimating spectral amplitude. The best spectral amplitude estimators, also known as STSA-MMSE estimators, were developed as a result of this and the prevalent view at the time was that the phase was considerably less significant than amplitude for speech enhancement.

Statistical-model-based algorithms provide nonlinear estimators of the signal's amplitude, rather than its complex spectrum, using a range of statistical models and optimization criteria (as is done in the Wiener filter). These nonlinear estimators explicitly take into account the noise probability density function (PDF) and speech DFT coefficients and, in certain cases, use non-Gaussian prior distributions. It is necessary to derive from the noisy input a nonlinear approximation of the clean speech signal. These nonlinear estimators are created using a range of techniques, such as maximum likelihood (ML) estimators and Bayesian estimators, which differ in the assumptions they make about the parameter of interest and the optimization criteria they use.

1.2.2.5 Subspace Algorithms

Subspace-based algorithms (SSA) are the third category of enhancement algorithms, which differ from the previously stated algorithms in that they are principally drawn from the concepts of linear algebra rather than, to the same extent, from those of estimation theory and signal processing. Signal subspace approaches are an empirical linear technique used in dimensionality reduction and noise reduction. Principal component analysis (PCA), Singular value decomposition analysis, and Independent component analysis are three statistical analytic procedures that are used in SSA.

The foundation of the subspace algorithms, which are based on linear algebra, is the assumption that the clean speech signal may be confined inside a subspace of Euclidean space. The vector space is divided into a noise subspace that solely contains noise elements and a signal subspace that includes both clean signal and noise components. Once the noise subspace had been

nullified and the signal subspace had been cleared of noise elements, the clean signal could then be estimated. Using well-known orthogonal matrix factorization methods from linear algebra, such as Eigenvalue decomposition (ESD) or singular value decomposition, the signal and noise subspace is decomposed (SVD).

The subspace approach for noise reduction was developed by Pisarenko [9]. Understanding the sinusoidal frequencies and the noise covariance matrix allowed the approach to be used to detect p sinusoids in additive white noise. Later, Schmidt [10] devised the multiple signal classification (MUSIC) approach to analyze numerous wavefronts arriving at an antenna array that picked up transmitted sinusoidal frequencies. The signal component was extracted from the noisy data set using the Tufts et al's [12] approach, which used SVD of the Hankel data matrix and projected the noisy signal onto the signal subspace. It offered the clean signal's least square estimator, which eliminates the singular values associated with the noisy signal. De Moor [13] further enhanced this technique by employing a minimal variance estimator that reduced the mean square error of the reconstructed signal.

Although the fundamental SSA approaches assume that noise is additive, they have been extended to include various kinds of noise. Dendrinos [s-13] initiated the work in the field of speech enhancement using SSA and proposed using SVD on a data matrix comprising time-domain amplitude values. Later, Ephraim and Van Trees [15] proposed utilizing eigenvalue decomposition of the signal covariance matrix. It has been demonstrated that the aforementioned algorithms enhance speech quality and minimize listener fatigue.

1.2.3 Noise estimation in Speech Enhancement

There are many different types of noise, which significantly impact the intelligibility and quality of speech. The background noise, the presence of other speakers, a noisy channel, etc. are all factors that lower speech quality. Based on a variety of statistical, spectral, or spatial characteristics, this noise can be categorized.

Noise can also be categorized into additive background noise, speech-like noise, impulse noise, convolutive noise, and multiplicative noise depending on the type and characteristics of the noise sources. The performance of the speech enhancement system may be improved with a suitable speech model.

Noise statistics have to be calculated from the noisy speech in single-channel speech-enhancing systems. Particularly in an environment with non-stationary noise, noise variance estimation is crucial in determining the effectiveness of the speech enhancement system. For noise estimation, various approaches exist [16-19]. One of the most popular strategies is the use of Voice Activity Detection (VAD), which updates noise from noisy speech during speech absence frames. This method is popular due to its complexity and usability. Since it is difficult to detect speech active zones in these circumstances, the accuracy of the technique often declines at low SNRs. Additionally, under extremely non-stationary noise settings, the enhancement system's performance is compromised since the variations in the noise during the speech active areas are not detected. With no prior knowledge about the characteristics of noise, it is therefore more challenging to deal with non-stationary noise, especially in low SNR situations. The traditional approach for estimating noise from the initial intervals by assuming there is no speech signal is unsuccessful when attempting to estimate non-stationary noise. Another extensively used technique is the minimal statistics (MS) noise estimator[16] which also yields excellent tracking performance for non-stationary noises.

1.2.4 Low SNR and Non –stationary Conditions

Many research studies have been conducted to enhance speech. Traditional algorithms like spectral subtraction, Wiener filtering, subspace methods, and statistical methods often assume that the noise is stationary which works effectively when SNR is high. They are not effective in handling non-stationary noise types. In many real-world scenarios, speech signals are often distorted or even totally submerged by strong noise (Low SNR). Consequently, it is necessary to develop enhancement methods under low SNR settings.

The removal of non-stationary disturbances was investigated using clean speech and noise data to train prior models for HMM-based [20] and codebook-driven speech enhancement [21]. During training, these methods accurately represented several non-stationary noise types. In relevant studies, they were extended to include undetectable non-stationary noise. [22] suggests an online noise estimation method using linear prediction coefficients (LPC) features by updating the auto-regression HMM (ARHMM) parameters of the noise in a recursive Expectation Maximization framework. Creating a functional observation probability density for each HMM state allowed for the development of an adaptive HMM [23]. Super-Gaussian HMM priors were used in [24] to offer spectral domain speech enhancement. These methods provide baselines for algorithms utilizing a pre-trained speech model, similar to the semi-supervised baselines used in this work. Modeling spectral noise has been suggested using nonnegative dictionary learning [25]. The dictionary has speech and noise bases. A convex combination of speech and noise bases is first used to describe the magnitude or power spectrogram of the input noisy speech before it is recreated entirely using speech. The estimated clear speech is from the reconstructed spectrogram. The primary estimate is improved using Wiener filtering or other smoothing methods. These methods' applicability is constrained since they rely on prior knowledge of speech, noise, or both. Using noise-only extracts as prior knowledge, dictionaries of the related noise were trained in [26], where a wind noise dictionary was estimated prior to enhancement. [27] describes the learning and fixing of a speech dictionary as well as the use of clean speech data to train dynamical systems with sparse non-negative speech dictionaries that can be adjusted to varying noise environments. Without prior noise and speech training, we must first distinguish speech from noise in an unsupervised manner and then filter or smooth the speech estimate. Inspired by Spectral Subtraction, the first step is to identify noise/speech areas in the input noisy speech spectrogram. Voice activity detection (VAD) can assist in this regard. VAD performance relies on noise level and type. Strong non-stationary noises might cause failure [28]. Time-frequency binary mask estimation is another solution [29], although it requires supervised training. Recently, speech processing from computer vision was applied by employing sparse and low-rank decomposition [30]. The quality of noisy speech was enhanced by pre-training a low-rank speech dictionary assuming sparse noise [27].

1.2.5 Effects of window and overlap processing on power Estimates from Spectra

The assumption that the signals are stationary and ergodic serves as the basis of Fast Fourier Transform (FFT) spectrum processing. As the speech signal is non-stationary, it is difficult to analyze in practice. This issue is solved by framing (segmenting) the noisy speech signal into an equal number of samples. The signal processing techniques are now employed and each frame length is taken individually to be stationary. Depending on the duration of the speech signal, the number of frames changes from one speech signal to another. Each of the speech signal frame edges experiences discontinuities during framing. If a Fast Fourier Transform (FFT) of such a signal is computed after segmentation using a no-window or rectangular window, the resulting spectrum leakage, also known as leakage error, occurs when the spectral energy spreads over a broad frequency range in the FFT instead of its actual frequency range, as shown in figure1.6. The signal's exact frequency content might be hard to distinguish, and its amplitude is smaller than its true value due to spreading (bias). The leakage and bias generated are minimized by windowing [31]. It requires multiplying the time record by a window with a finite length and an amplitude that gradually and smoothly decreases to zero at the edges, causing the signal to be periodic. When applying the windows, a compromise is required. A large main lobe often causes greater bias, but a small side lobe level typically reduces leakage. The two properties could not be optimized simultaneously. Reduced side lobe level results in decreased leakage error, increased main lobe width, and decreased spectral resolution.

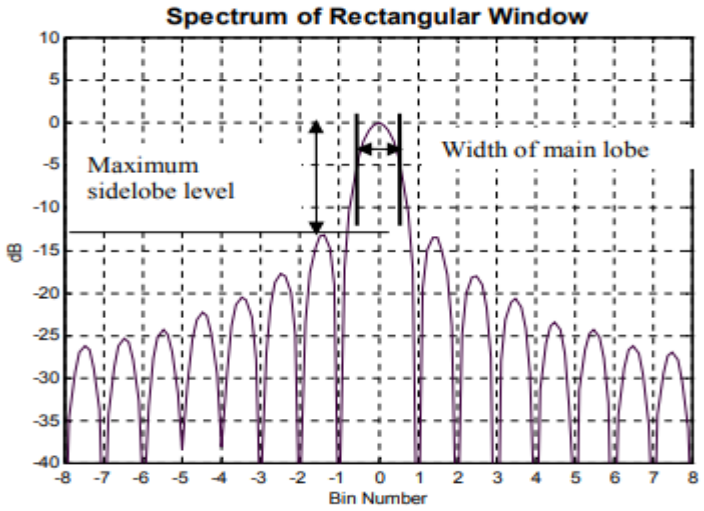


Figure 1.6 Illustration of spectral leakage in a rectangular window

Over time, a variety of windows have been suggested, some of which enhance frequency resolution and others that enhance amplitude accuracy. Smaller side lobes are required by the fact that most spectra have a small first derivative, hence Hanning, Tukey, and Hamming windows are chosen over other windows like rectangular and Bartlett (Triangular) windows.

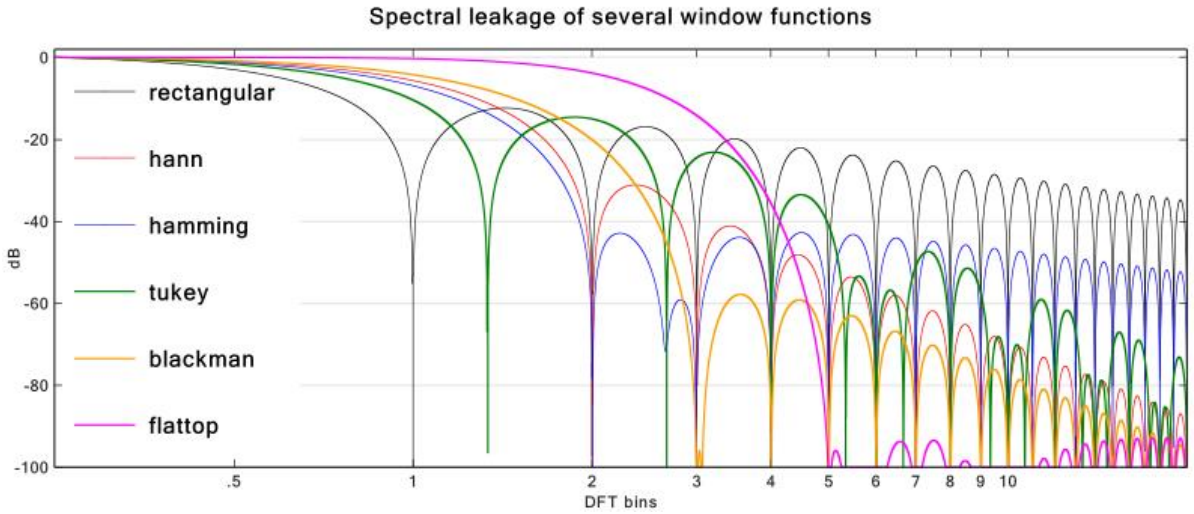


Figure.1.7 Comparison of spectral leakage of several windows (Courtesy: Bob K, CC0, via Wikimedia Commons)

The most popular window is the Hanning window because it provides good frequency resolution and offers less spectral leakage at frame boundaries than Hamming window does. The window size is chosen based on the frame length (‘N’). The Hanning window is represented as [1.1],

$$w_h[n]=0.5(1- \cos(2\pi n/(N-1))), \text{ where } 0\leq n\leq N-1 \quad (1.1)$$

Where ‘N’ is the size of the window, ‘n’ is the length of the speech signal. An overlap between the frames must be allowed so that there is no loss in the speech signal information.

The noisy speech signal ‘$h\,[n]$’ is multiplied with this window function $w[n]$ by allowing a frame overlap to obtain the resultant signal as :

$$x[n]=h\,[n]*w_h[n] \quad (1.2)$$

The technique, known as the overlap-add method, employs short-time signals that overlap and reconstructs the signals by adding partially overlapping frames. Since it is reasonable to suppose that speech signals produced by muscular movements are stationary for a duration of roughly 20 to 30 ms, overlapping windows of that length are necessary.

1.3 Motivation

Speech signal is vulnerable to noise in real-world environments. Typically, background noise in everyday life interferes with conversation, rather than in a fully calm environment. Background noise significantly impairs the performance of many speech signal processing applications, including automatic speech recognition (ASR) systems, telecommunication systems, man-machine interfaces, electronic hearing aids, poor audio devices, etc. As a result, they have poor speech quality and intelligibility, which makes communication difficult and so restricts their use.

Speech Enhancement (SE) methods attempt to improve the perceptual quality of speech and intelligibility with low distortion and residual noise. Numerous strategies have been proposed for SE over the years. The traditional SE techniques make several noise-related assumptions, such as stationarity, low magnitude, directly pre-fixing the spectra, and rank with input SNRs greater than 0 dB. But typically, real-world noise is strong, divergent, unseen, and non-stationary. When these presumptions are not considered, enhancement rapidly decreases. Most of the existing SE techniques improve speech quality at the expense of intelligibility and distortion. Particularly in low SNR(< 0 dB) and various background noise scenarios, there is a substantial amount of distortion introduced. As a result, the SE approaches fall short of expectations. Because noisy interference is so unpredictable, the SE performance cannot guarantee the functioning of real-world applications. Recently, low-rank sparse decomposition methods like Robust Principle Component Analysis (RPCA) have been demonstrated to be superior to other approaches in challenging situations, particularly when input SNR is low. Since speech intelligibility must be sacrificed to reduce noise in low SNR settings, residual noise and distortion are introduced into the enhanced speech. Therefore, there is a need to develop Speech Enhancement methods that can provide better speech quality and intelligibility at low SNR with the least distortion.

Further, none of the existing methods was tested for improving ASR performance. Therefore, there is a need to validate the effectiveness of the proposed SE method in recognizing the speech uttered under low SNR with the least error.

1.4 Statement of the Research Problem

The Speech Enhancement (SE) system should have the ability to improve both the intelligibility and perceptual quality of noisy speech input. Most of the existing SE techniques improve speech quality at the expense of intelligibility. Few of the existing approaches have achieved better speech quality with intelligibility with considerable signal distortion and residual noise. These methods have shown poor performance in challenging circumstances, particularly under low Input SNR conditions. Under low SNR situations, it is hard to suppress noise without

sacrificing speech intelligibility by introducing residual noise and distortion in enhanced speech. In comparison to existing strategies, the low-rank sparse decomposition (LRSD) methods have been shown to provide a better compromise between speech quality, and intelligibility along with low level of residual noise and output speech distortion. Therefore, there is a need to investigate the use of sparsity property in clean and noisy speech signals in dictionary-based methods for SE. The influence of parameters in the LRSD models using Robust Principle Component Analysis (RPCA)and Semi Soft Go-Decomposition (SS-GODEC) methods under extended range and various types of Non- Stationary noise environments are to be explored. To achieve less residual noise and speech distortion under low SNR conditions, there is a need to develop novel noise robust SE methods using LRSD Algorithms. Further, the recognition accuracy gets affected due to strong noisy environments and thus there is a need to validate the performance of the developed SE algorithms so that it is robust even under noise conditions**.**

1.5 Research Objectives

The objectives of the present research are:

1. To Investigate the use of sparsity property in clean and noisy speech signals for Single channel Speech Enhancement (SE) using traditional methods such as fixed dictionaries (DFT, DCT, CEPSTRAL, etc) and adaptive dictionaries (KSVD, NMF, etc).

1 a) To Investigate the influence of parameterization of existing Sparsity-based SE techniques using Robust Principle Component Analysis (RPCA)and Semi Soft Go-Decomposition (SS-GODEC) methods under extended range and various types of Non- Stationary noise environments. b) To develop novel noise robust speech enhancement methods using Low -rank Sparse decomposition Algorithms such as Weighted Nuclear Norm minimization (WNNM), and Weighted Schatten p-Norm Minimization (WSNM). c) To combine Wavelet, Gamma-tone filter bank, and KL Divergence with the weighted low-rank sparse decomposition algorithms (WNNM & WSNM) for SE to achieve low residual noise and speech distortion.

3. To perform a comprehensive evaluation of the proposed techniques and compare the

performance with existing algorithms that are suitable under low SNR conditions (< 0dB).

4. To assess and validate the performance of the proposed speech enhancement schemes with the specific goal of improving recognition accuracy by reducing the word error rate (WER) using state of art Automatic speech recognition (ASR) systems under noisy environments.

1.6 Speech Corpora Used in the Thesis

The Speech corpora used in most of the SE works are Noizeus, TIMIT and Libri data bases. Hence these databases are used for evaluating the performance of the proposed algorithms in this thesis.

1.7 Organization of Thesis

The thesis is organized into seven chapters. The following section gives a summary of the chapters.

Chapter 1: Introduction

The chapter introduces speech enhancement and a description of the low-rank approach for Speech Enhancement, motivation, problem statement, and objectives of the thesis work.

Chapter 2: Literature Survey

The chapter explains the limitations of existing Speech enhancement techniques. The literature also discusses a variety of noise estimating strategies and the scenario of low SNR handling. It presents the issues identified, and the different databases used.

Chapter 3: Speech Enhancement using Dictionary-based techniques

The chapter discusses the investigations carried out to test the sparsity of speech signals for SE using dictionary-based techniques under different noise conditions taken from NOIZEUS database.

Chapter 4: Speech Enhancement using Low-Rank sparse decomposition techniques under Low SNR Environments

The limitations of RPCA and SS-Godec methods are explained in this chapter. The implementation details of proposed weighted low-rank sparse decomposition methods for SE under low SNR conditions along with results are presented here.

Chapter 5: Unified speech enhancement Approach for Low distortion Under Low SNR Environments

The drawbacks of the proposed weighted low-rank sparse decomposition methods are discussed in this chapter. A novel SE method by combining Wavelet, weighted low-rank sparse decomposition algorithm, and gamma-tone filter bank under low SNR environments is proposed. The implementation details of the proposed SE method are explained along with the results.

Chapter 6: Validation of the Proposed Speech Enhancement system

This chapter presents the validation results of the proposed Speech enhancement techniques that are carried out by training Kaldi ASR to achieve low WER using different noises with SNRs ranging from -10dB to 10dB.

Chapter 7: Conclusions and Future Scope

The conclusions and scope for the future research is presented in the chapter.

Chapter 2

Literature Survey

The literature on Speech Enhancement and the low-Rank sparse decomposition techniques for Speech Enhancement is presented in this chapter. Recent strategies for post-filtering, noise estimation, and speech signal processing with low SNR are also discussed.

2.1 Introduction

In real-world situations such as Automatic Speech Recognition, hearing aids, and other communication systems, Speech Enhancement is a difficult task. It tries to improve the quality and intelligibility of speech signals degraded by a variety of noise situations, such as traffic & car, crowd, machine, street, reverberation effects, and speech signals from other speakers, etc [26]. A robust speech enhancement system should be able to function well in all noisy environments.

2.2 Enhancement Methods for Speech Processing Systems

There are numerous application areas for speech enhancement, such as telecommunication, electronic hearing aids, speech/speaker recognition, audio restoration, and surveillance, etc., In noisy environments, a speech enhancement stage can be placed as a front-end processor to reduce noise power and improve the quality and intelligibility of the signal.

Murnolo et al.[33] suggested a speech enhancement method to minimize the noise in wireless digital speech communications by analyzing the 2D spectral map and filling frame gaps using a heuristic rule. Rahmani et al.[34] suggested a dual-channel speech enhancement strategy based on the difference between the powers of the two received signals in near field situations, which worked well with adjacent microphones and non-stationary noise. Wee-Tong Lim[35] employed psychoacoustic signal processing based on the "missing fundamental phenomena" to amplify poor signal receptions at low frequencies to improve the performance of Singapore Armed Force's (SAF) Radio communication equipment. A polynomial-based nonlinear device was synthesized and spectrum-shaped. Beritelli and Rametta [36] provided the design and performance assessment of the dual stream solution for HD-VOIP transmission service in forensic scenarios wherein speech quality is critical. It reduced packet loss by 85% and improved speech quality by 0.8 (MOS). Srinivasan and Phandari pande[37] considered multi-microphone speech enhancement in an interference environment like hands-free voice communication and multi-party conference. When interference occurs, a directional microphone is utilized to measure its power spectral density (PSD). The PSD estimate is quantized and sent via a wireless network to a primary omnidirectional microphone, where the PSD of the spoken signal was measured. Optimal rates for encoding signal PSDs were investigated to decrease overall transmission power while maintaining MSE within a set limit. Nabi et al. [38] presented wavelet-based speech enhancement for mobile phones in stationary and non-stationary noise. Optimized filters were designed and used as mother wavelets by iteratively adjusting the cut-off frequency.

Speech recognition performance degrades in noisy or distorted environments [39]. The approach is to use a speech enhancement step as a pre-processor [40-42].

A vocal tract spectral limiting iterative wiener filtering based on line spectral pair transformation was reported by Hanson and Clements[43]. Athanaselis et al.[44] examined SVD-based and non-linear spectral subtraction speech enhancement algorithms for robust ASR when the input signal was contaminated by colored noise with variable SNR (SNR). Fine-tuning enhancement approach parameters is vital to ASR system performance. A robust speech recognition system relying on a de-Reverberation method and a spatial masking-based noise

filtering algorithm, where the threshold angle is first learned in several noise-only frames and then updated frame by frame, was presented by Qoc et al. [45] using binaural speech enhancement as a pre-processing step. A speech recognition system for living room noise was introduced by Delcroix et al. [46]. A multi-channel speech-noise separation technique and a single-channel enhancement algorithm were used by the recognition system's model-based speech enhancement pre-processor. The technology improves speech audibility and keyword identification accuracy at human levels. Li et al.[47] presented dynamic change enhancement and mean smoothing after estimating sub-band log energy to improve In-car Speech recognition.

A Deep Neural Network (DNN)-based feature enhancement method was proposed by Lee et al. [48] where the DNN inputs constitute pre-enhanced spectral features inferred by the Direction of Arrival restricted ICA preceded by Bayesian Feature Enhancement (BFE) using Hidden Markov Model (HMM), before reconstructing noise-resistant features for robust ASR even under mismatched noise conditions. The DNN learns to recover a clean spectral feature vector from corrupted, pre-enhanced, and noisy inputs. Cho et al.[49] developed an HMM-based BFE approach employing Independent Vector Analysis(IVA) and Reverberation parameter Re-estimation(RPR) in a multi-microphone system to enhance ASR WERs in additive noise and reverberant distortion situations. Chang Huai You and Bin MA[50] advocated employing spectral-domain speech enhancement to improve the ASR system by decreasing the feature distortion ratio, which comprised smoothing adaption to frame SNR and re-estimation of a priori SNR. Cho and Park[51] introduced an IVA-based feature improvement for robust speech recognition to address the low performance of ICA-based techniques in under-determined scenarios that cause erroneous noise spectrum estimation. Clean speech was estimated through Bayesian inference.

The World Health Organization’s (WHO) recent study [52] published global statistics on hearing loss. It claims that 430 million individuals, or greater than 5% of the worldwide people (432 million adults and 34 million children), have hearing loss and that almost one-third of persons over 60 years are also impacted by disabling hearing loss. Since these figures have increased over time, the issue requires highly effective intervention. The aging population and the exposure of both young and elderly people to noise are the main causes of the worrying increase. Hearing aids give persons with hearing loss a better quality of life and assist them in dealing with the disability.

The most crucial function of these devices is speech enhancement, which removes noise and adapts the signal for easy hearing. Over the years, a variety of speech enhancement algorithms for hearing aids have been suggested.

N. Whitmal et al. [53] employed implicitly filtered, shift-invariant wavelet packet basis vectors to improve the observed speech and reduce correlated noise while keeping low-level, high-frequency spectrum components' noise crucial for intelligibility. To improve speech quality and intelligibility for hearing aid users in noisy environments, L.Alvarez et al. [54] proposed a speech enhancement algorithm. This algorithm uses GMM fuelled by Genetic Algorithm(GA) to create an enhanced gain function and computes optimized parameters to maximize the Perceptual Evaluation of Speech Quality(PESQ) score. Thiemann. J et al. [55] proposed an SNR estimator that shifts between the output signals of a minimum variance distortion-less response beamformer and scaled reference microphone signals to enhance the target signal in binaural hearing aids. Reddy. C et al.[56] suggested a single microphone enhancement gain function that incorporates a trade-off parameter to optimize the super-Gaussian joint maximum a posteriori cost function and predict the clean speech magnitude spectrum. The level of noise reduction and speech distortion in the suggested approach is adjustable by smartphone users, allowing them to customize the system's performance to meet their individual needs. Nayan M et al.'s[57] approach proposes a multi-channel Wiener filter to improve speech for hearing aids that take into account a scalar combination of noise inputs to filter and speech correlations.

Restoration of historical and nostalgic audio recordings and their conversion to high-quality media have drawn increasing interest[58,59]. Wax cylinders, disc records, magnetic tape technologies, and even contemporary digital recordings all have deteriorated sources that need to be restored. Audio restoration seeks to eliminate any components from the recording that were not intended, such as background noise from the storage medium and background noise in general.

Vaseghi and Rayner suggested in [60] that when multiple copies are available, archival gramophone recordings can be restored. The signals are synchronized with noise rejection achieved by an adaptive Finite Impulse Response (FIR) filter interpolator based on least square error criteria.

Additionally, a detection estimation approach for removing impulsive noise is presented. The output of an autoregressive (AR) model driven by white noise stimulation is used to model audio signals. O'Shaughnessy et al.[61] enhanced the wiretap recording's intelligence. Applying the short-time spectral attenuation approach, Cape and Laroche[62] were able to restore music that had been ruined by background noise. Two methods for the restoration of movie soundtracks were described by Czyzewski et al. [63]. In the first, broad-band noise attenuation using a psychoacoustic model was employed, and in the second, sinusoidal components taken from the sound spectrum are used. A speech enhancement method was brought out by Xiao and Nickel in [64] for the offline restoration of old audio files for which there is a clear voice recording of the speaker. Based on its predicted probable features, the clean signal was re-synthesized. The technique improves the speech's naturalness and perceptual quality, but it has a significant processing cost and memory demand.

2.3 Speech Enhancement Approaches

The number of noise sources that are available limits the performance of a speech enhancement algorithm [1,65-68]. A single microphone input containing noisy speech is all that is needed for single-channel speech enhancement algorithms to produce enhanced speech [69–79]. The signal cannot be pre-processed since there is no noise signal reference, hence alternate speech and noise statistics are used for enhancement.

Single-channel SE systems traditionally make use of the Voice activity detection (VAD) stage to predict and update the noise statistics during noise-only segments. A well-designed VAD will improve the performance of the SE method in noisy environments concerning accuracy and speed, otherwise, it would degrade the system performance. In low SNR conditions, the current VAD approaches are imperfect. Furthermore, even if VAD is adequately built, alterations in the noise spectrum that occur in the midst of active speech segments are unable to affect the noise estimate on time. This would lead to an underestimation during long-spoken phrases where there are few noise-only portions. [80]. Several algorithms have been designed to improve speech quality by estimating and reducing background noise power spectral density (PSD) for stationary or slow-

varying noise signals with SNR above 0dB. Although these methods can enhance speech quality without any prior information about noise type, limited progress has been made to improve intelligibility under unseen non-stationary noise conditions that cannot guarantee a sufficient noise estimation for all scenarios.

Multichannel Speech Enhancement approaches function better in non-stationary noise circumstances because a reference channel is available [81–89]. To remove the undesired noise components, one of the channels might be phase-aligned. Multi-channel speech enhancement systems' cost of manufacturing and complexity is the main limitations.

In the majority of frequently used applications, such as mobile phones and electronic hearing aids, a single channel is utilized. Compared to their multichannel counterparts, single-channel enhancement methods are fairly simple to implement and less expensive.

Speech enhancement approaches can be categorized as either supervised or unsupervised. Supervised approaches reduce noise by considering a model for both speech and noise signals, which requires a training phase to predict the parameters. HMM-based methods[90-94], Gaussian Mixture models(GMM)[95-96], codebook-based algorithms[97-98], DNN-based approaches[99,100], and Nonnegative Matrix Factorization(NMF)-based methods[101-104] are examples of supervised methodologies.

A speech enhancement technique put forth by Philip Harding and Ben Milner[105] reproduces a clean speech signal employing a sinusoidal model and a set of acoustic speech features such as speech classification, fundamental frequency, and spectral envelope that are approximated from noisy speech using a single statistical model. By constraining the speech production model to generate the enhanced signal, the result is devoid of noise. Tian Gao et al. [106] developed a unified DNN method for reducing both background noise and speech interference in a speaker-dependent situation. The DNN system was trained to integrate speech enhancement and isolation. Speech interference signals are considered one form of noise. The unified system delivers superior performance under noise and speech interference mixed situations

compared to individual systems when just noise or speech interference is exists. Results indicate the efficiency of the ensemble approach in contexts with low SNR. In non-stationary noise situations, the performance of supervised techniques is dependent on prior knowledge provided to the system, which restricts its performance.

There are several unsupervised speech enhancement technologies for which no data is supplied. Estimating clean speech from noisy observations without prior knowledge of the noise type or speaker identification is difficult. Certain techniques need to be employed, these techniques include spectral subtraction algorithms, statistical model-based approaches, subspace methods, and low-rank sparse decomposition techniques[6].

Initiated by Boll[3], spectral subtraction is one of the first voice enhancement methods. To estimate the clean spectrum, an estimate of the noise spectrum is removed from the noisy speech spectrum. During speech absence frames, the spectrum of noise is determined. Major drawbacks include the generation of unpleasant musical noise following enhancement. Several changes, including perceptually inspired approaches (Petrovsky et al., [107]; Uderea et al.,[108]) and a geometric approach in complex planes, are proposed to minimize musical noise (Lu and Loizou[109]).

Statistical model-based approaches [110, 111] presume that the distributions of clean speech and noise spectra follow statistical models and employ the reduction of MSE or the Maximum aposteriori (MAP) estimation to get enhanced speech. Traditional approaches, such as the Wiener filter [112] and the MMSE estimate of the Short Time Spectral Amplitude (MMSE-STSA) by Ephraim and Malah[113], use Gaussian distribution models for both speech and noise spectra. Later, Martin[114,115] and Chen, Loizou [116] applied non-Gaussian distributions such as Laplacian and Gamma to model clean speech and improve speech enhancement. Trawicki and Johnson[117] constructed innovative perceptually-motivated MMSEWE and WCOSH cost functions and Chi distributions for speech prior to providing gains in all stages of enhancement. In addition to enhancing speech quality and intelligibility, it reduced background noise. Abutalebi and Rashidinejad[118] considered Laplacian speech modeling and β-order MMSE approach for

speech enhancement based on Mohammod et al's [119] model but chose to derive β-order LapMMSE estimator which is faster and less complex by applying some approximations for the Bessel function and the probability density function of the magnitude spectrum of the clean speech. The order of the cost function(β) is modified based on the frame SNR. The studies in this area suggested new optimal Laplacian distribution-based linear and non-linear estimators. The estimators are generated using an MMSE sense to reduce speech distortion under various scenarios.

Spectral subtraction [3] and Wiener [110,111] filtering has been used often for speech enhancement due to their simplicity and ease of implementation in single-channel systems, but one of their main drawbacks is the generation of musical noise following enhancement. Smoothing techniques such as the decision-directed method [113,120] or Wiener filtering based on prior SNR estimation [1-3] are frequently employed to minimize the intensity of musical noise. To eliminate residual noise, the majority of algorithms frequently generate a signal with significant distortion.

Linear algebra is the foundation for subspace methodologies. They map the noisy speech onto two orthogonal complement subspaces: a signal subspace including clean speech parts in addition to the noise part, and a noise subspace containing exclusively noise part. These subspaces are formed by Singular Value Decomposition (SVD) of a noise-corrupted speech data matrix [14] or Principal Component Analysis (PCA) utilizing Eigen Value Decomposition (EVD) of a noise-corrupted speech covariance matrix [15]. Speech enhancement is performed in two steps. In the first step, the noise subspace is eliminated by mapping the data onto a subset of the principal directions of the eigenvectors of the SVD or PCA analysis. In the second step, the contribution of noise to the signal subspace is diminished. In comparison to conventional approaches, signal subspace techniques have been shown to offer a superior compromise between signal distortion and residual noise generation.

It is hard to reduce noise without degrading speech because of the unpredictability of the noise and the inherent complexity of the speech signal. It has been a long-term objective to provide speech enhancement methods that guarantee a good balance between residual noise and output signal distortion[121]. In comparison to other current strategies, SSA[122,123] has been

demonstrated to provide a better balance between the two and in [124] a notable improvement of the output signal's SNR. The Karhunen-Loeve Transform (KLT) based techniques showed promise in improving the intelligibility of distorted speech, as demonstrated by Hu and Loizou[125]. Ephraim and Van Trees[15] used EVD, which employed KLT to project clean speech into the signal plus noise subspace.

In comparison to traditional SSA, it has been demonstrated that the Jabloun and Champagne method[126]'s of perceptual qualities in subspace reduces residual noise. The lack of an Eigen domain explanation for the hearing properties (i.e. masking effects) made it extremely difficult to include psychoacoustics into KLT-based approaches. In [126], the appropriate transformations for converting the masking threshold to the KLT domain and inversely are recommended. Ju and Lee[127] present an extended SVD-based method that is perceptually limited, incorporates the human auditory system's masking capabilities, and accurately and automatically determines the signal subspace dimension to render residual noise undetectable.

Source separation and dictionary learning [128] denoising methods were developed to overcome the above limitation. These methods include sparse coding like principal component analysis (PCA) [129], Independent component analysis (ICA) [130], K-SVD [131], and Non-negative Matrix Factorization (NMF) [132]. Dictionary learning (DL) techniques are effective in the context of SE based on compressive sensing, in which the prime data vectors are described by a sparse linear combination of basis elements. The DL process is hard due to insufficient data to create a noise dictionary. Even though PCA is particularly sensitive to noise, data corruption can cause estimates of the low-rank part to be inaccurate relative to the actual model. In the last few years, alternate and contemporary supervised SE approaches have made rapid advancements. They include Subspace methods like Nonnegative matrix factorization (NMF) [133], Hidden Markov Model [134], binary and soft mask estimation [135], etc. Recently, approaches such as supervised learning with a Gaussian mixture model or deep neural networks [136,137], and Deep denoising auto-encoders [138] have improved mask estimation performance. All of these solutions require either a specific characteristic or extensive initial training for supervised separation. However, accuracy is reduced due to the disparity between various real-world noises and training noises.

To solve this problem, another very elegant remedy called Robust Principal Component Analysis (RPCA) [139.140] was proposed. This is an unsupervised method solved via Principal Component Pursuit (PCP) [141] that decomposes noise-corrupted speech matrix into low-rank and sparse structures using convex optimization. The resulting sparse component contains speech-dominant features and a low-rank noise component. One obvious benefit of employing RPCA to improve noise robustness is that it requires no prior knowledge of the noise. Contemporarily, the convex relaxation of the rank minimization model and the nuclear norm minimization (NNM) problem has seen a lot of research interest in recent years. One can improve perceived audio quality and/or intelligibility with low signal distortion by utilizing the most successful machine learning algorithm. To estimate the noise spectrum from the input noisy speech spectrogram without any prior knowledge of speech and noise, D.L. Sun et al. [142] suggested a sparse and low-rank NMF with kullback-leibler divergence. This was achieved by separating the input noisy magnitude spectrogram into a sparse speech-like portion and a low-rank noise component.

Instrumental measures were inspired by BSS_Eval and perceptual metrics like the signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), source-to-artifact ratio (SAR), short-time objective intelligibility ratio (STOI), Perceptual evaluation of speech quality(PESQ), Signal distortion (SIG); background intrusiveness (BAK) and overall quality(OVL) have been developed since human listener testing is time-consuming and expensive. These metrics were designed to estimate how effectively new algorithms will perform by modeling human responses [143-145].

2.3.1 Low-Rank and Sparse Matrix Decomposition Methods

From the basic principle of RPCA, the noisy speech spectrum is decomposed into low-rank and sparse matrices using the Principal Component Pursuit (PCP) model. By using effective estimating techniques, the sparse and low-rank components may be estimated and recovered with a high probability. The low-rank matrix approximation (LRMA) method seeks to retrieve the underlying low-rank matrix by minimizing the rank of its relaxations from its corrupted observations of speech. Unfortunately, rank minimization is an NP-hard problem with no known efficient solution[146]. The nuclear norm, which contributes to NNM-based approaches [147], is

the best choice for substituting the rank function with its tightest convex relaxation. The classical Low-rank matrix Factorization (LRMF) method also known as the SVD technique, is capable of achieving the optimal rank-r approximation of input data matrix M by using a truncation operator on its singular value matrix with regard to F-norm fidelity loss. To suppress outliers mixed in data, a robust LRMA method called robust principal component analysis (RPCA) framework, based on nuclear norm minimization (NNM) is introduced. The NNM could be solved by the singular value thresholding algorithm [148] using the alternating direction method of multipliers (ADMM) [149] framework, which also belongs to the augmented Lagrange multipliers (ALM) framework. In the time-frequency (T–F) domain, noise signals present in distinct time-frames have similar spectral structures, and patterns are usually correlated with one another and can be captured with a few basis vectors. Therefore, the noise spectrogram is supposed to lie in a low-rank subspace. Further, as the spectral energy centralizes in a few T-F units, speech signals can be assumed to be relatively sparse in T–F domain [150]. The RPCA method is non-parametric and does not require any particular assumptions regarding the distribution of the spectral components of speech or noise. Because both speech and noise spectra can be recovered at the same time, therefore the procedure of VAD is unnecessary and irrelevant in this context. This method is superior to many conventional SE algorithms that depend on the performance of noise estimation algorithms [151,152]. The RPCA algorithm provides the benefits of a small number of tuning parameters and quick processing. Additionally, it can function effectively in strong noisy environments. This favours denoise speech through mask estimate on spectrogram via sparse and low-rank decomposition. Sharing similar principles several modifications have been investigated, to improve further the performance of low-rank and sparse models like the SS-GoDec [153] algorithm for the SE. Nuclear norm minimization (NNM), which can precisely retrieve the rank of the matrix with a few constrained and theoretical guarantee conditions, is the most notable effort.

2.4 Research Gaps Identified from the literature

The RPCA and SS-GODEC approaches based on Nuclear Norm Minimization (NNM) may, however, provide unfavourable results if prior information on the signal source is not used. The

standard NNM regularizes each singular value equally, resulting in simple calculation of the convex norm. This limits its flexibility and capacity to handle a variety of real-world problems where singular values have significant physical implications and need to be addressed as such. Also, these algorithms are limited due to the approximation of the original rank of noise through NNM and do not explicitly use low-rank property in optimization. As a result, NNM frequently tends to over-shrink the rank components, making it impossible for it to effectively approximate the matrix rank for many real-world applications. Recent developments have demonstrated that WNNM, which heuristically sets the weight as inverse to the singular values, outperforms NNM in terms of achieving a superior matrix rank approximation. It is proved that the recently proposed WNNM can replace the traditional nuclear norm, as an improved approximation to the rank of a matrix in computer vision applications [154]. As RPCA and SS-GODEC algorithms explicitly account for deviations of speech and noise time-frequency matrices from the idealistic sparse and low-rank model, there is a need to propose an alternate SE algorithm for speech and noise spectrogram separation by enforcing weighted low-rank and sparsity constraints. With the help of low rankness of WNNM, the efficacy of enhancement by using singular value decomposition, the ADMM, and the accelerated proximal gradient line search method can be improved. Therefore WNNM-based RPCA enhancement model can be used, which takes advantage of high correlation of the speech signals, adding excellence to NNM-based methods. Further study led to the invention of a new RPCA model, the weighted Schatten p-norm minimization model, to effectively perform low-rank regularization (WSNM). It is demonstrated in [155] that WSNM suppresses noise more effectively than state-of-the-art approaches and is better at modeling dynamic and complex situations. WSNM is a generalized version of WNNM whose performance in image denoising was analysed.

Though the standard RPCA-based approaches have proven to be useful for SE, there are a few potential drawbacks limiting the effectiveness. First, RPCA approaches are often approximated by spectrogram analysis using the short-time Fourier transform (STFT). However, due to segmentation and windowing operations, there is distortion in the STFT process [156,157]. Second, most of these algorithms optimize their cost function, which is based on Euclidean distance (ED). ED, can often lead to fairly significant reconstruction errors since it tends to overemphasize the accuracy of large values. As a result, the ED measure is not appropriate for

processing speech signals [158]. Third, the majority of existing SE approaches improve the STFT-based spectral magnitude while retaining the input noise-corrupted phase part unaltered, leading to distorting the recovered speech signals and reducing SE performance [159,160]. Fourth, the most well-known strategy used in the evaluation of RPCA is nuclear norm minimization (NNM) which precisely restores the rank of the matrix within specific constrained and theoretically guaranteed circumstances. In many cases, NNM fails to predict the rank of the matrix accurately [161]. These strategies are constrained by the estimate of the real rank of noise, and they do not fully utilize the low-rank characteristics in optimization.

2.5 Conclusion:

Extensive research for speech enhancement has been done using various methods. According to literature, reducing signal distortion and residual noise are main challenges in speech enhancement. Speech enhancement in low SNR situations is difficult. Low rank sparse decomposition models have demonstrated adequate performance in highly noisy situations, however they exhibit some signal distortion in low SNR (< 0 dB) conditions. Therefore, in order to improve performance, a novel speech enhancement strategy using low rank sparse decomposition models is required.

Chapter 3

Speech Enhancement using Dictionary-based Techniques

The main objective of this chapter is to address the issues identified from the literature survey. In this the use of the sparsity property in clean and noisy speech signals for single channel speech enhancement (SE) was investigated through i) fixed dictionaries like DFT, DCT, CEPSTRAL, etc., ii) adaptive dictionaries like KSVD, NMF, etc., and iii) low-rank sparse decomposition approaches like RPCA, SSGODEC.

3.1 Motivation

Sparse representation techniques have become more popular in recent years as a way to enhance speech. With the sparse representation technique, the most important information about a speech signal can be represented using a smaller number of classic spatial bases. The ability of a sparse decomposition method depends on the learned dictionary and how well the dictionary atoms match the signal features. The dictionary learning process and the sparse coding technique are the two main steps that lead to an over complete dictionary. In the dictionary selection step, a dictionary that has already been predetermined is used. Also, adaptive dictionary can be formed by a process of learning that is often based on alternating optimization strategies.

3.2 Introduction

Speech enhancement aims to improve the quality of noisy speech, typically by suppressing noise in such a way that the residual noise is not annoying to the listeners, and speech distortion introduced by the enhanced process is minimized.

Existing SE approaches are limited to stationary noise and lack robustness for non-stationary real-world noise. Compressive sensing (CS) promises to effectively recover a sparse signal from a few random samples [191-193]. A CS recovery method reconstructs structured speech while eliminating unstructured noise. Encouraged by the emerging technique, this investigation analyses the performance of CS in SE using fixed and adaptive dictionaries [194,195]. Various sparse domain and sensing matrices and combined transform domain (dictionary) combinations that satisfy incoherence criteria have been tested for speech enhancement. Despite being straightforward and having fast calculations, these non-adaptive(fixed) dictionaries are unable to effectively (sparsely) represent a particular class of signals. The last ten years have seen extensive research towards dictionary learning as a solution to the above problem [196,197]. In this approach, a dictionary is learnt from a certain class of training signals of interest. In several signal processing applications, such as image compression and enhancement [172] and classification tasks [197], it has been empirically demonstrated that these adaptive dictionaries perform better than non-adaptive ones. An adaptive dictionary requires prior knowledge of speech and noise for supervised learning. Under unseen non-stationary and strong noise environments, the adaptive dictionary approaches are ineffective. Therefore, real-world noisy speech enhancement requires unsupervised approaches to achieve better performance.

For supervised learning, an adaptive dictionary needs prior knowledge of speech and background noise. However, due to varying perspectives and circumstances, training and testing domains may differ during the designing of dictionaries. The adaptive dictionary techniques fail in contexts with significant noise and unknown non-stationarity. Therefore, unsupervised techniques are needed for noisy speech enhancement in real-world applications to attain higher performance.

The unique data analysis method known as Robust Principle Component Analysis (RPCA) has proven to be successful for data contaminated by noise. The benefit of RPCA is that it can be applied in unknown real-world noisy conditions since it doesn't require any prior information. Additionally, it can function well in strong noisy environments. As a result, RPCA as a low-rank and sparse matrix decomposition model has been often employed for unsupervised separation for robust speech enhancement. This chapter presents the results of simulation that were performed to

examine the suitability of low-rank sparse representations for SE under different NOIZEUS corpus settings employing objective and subjective measurements.

3.3 CS-BASED SPEECH ENHANCEMENT USING FIXED AND SPARSE DICTIONARIES

In recent past, the emergence of compressive sensing (CS) or sparse recovery, another method for sampling signals and lossless recovery based on sparse representation, was proposed [162-167]. The samples can be directly acquired from a high dimensional signal via linear mapping, and thus the process of measuring the entire signal is eliminated. The theory of CS assures an exact reconstruction of a low dimensional sparse signal, wherein the number of samples or measurements taken randomly is proportional to the sparsity level and a log factor of the signal dimension. An efficient reconstruction algorithm is needed to recover the signal from compressed samples [168]. Researchers have been attempting to apply the novel concept of CS to solve many of the signal processing problems [169-170] such as robust signal reconstruction, image processing, analogue to information conversion, radar data analysis, computational biology, etc. The property of sparsity is exploited to separate the non-sparse components in a noisy signal mixture. The important aspect of these applications of CS to real-world scenarios is that the sparsifying basis matrix must be known a priori. As the speech is sparse in the time-frequency domain, some of the CS-based speech enhancement methods proposed [171] use a random sampling matrix in the sensing scheme to extract the components of speech from a noisy signal. Often, the random sampling matrices are designed carefully using a mathematical model of the signal, which possesses Restricted Isometric Property (RIP) and incoherence with sparse matrix.

3.3.1 SPARSE DICTIONARY MODELING

Sparse modeling is a process of representing the input signal as a linear combination of fewer basis vectors or elements called atoms that are chosen from the dictionary (i.e., the whole

collection of the basis signals or code words) [172]. A sparse dictionary is based on a sparsity model of the dictionary atoms over a fixed base dictionary [173]. An over-complete (redundant) dictionary not only allows to represent multiple ways of the same signal but also improves the sparsity and flexibility of representation [174-175]. As natural signals are composed of impulsive and oscillatory transients, a transform matrix may exhibit sparsity to some of its components in one domain while others are sparse in some other domain. Therefore, in CS, searching for the best transform matrix is of great importance. However, in certain cases, a dictionary that is the best fit for the speech signal can significantly improve sparsity, which can be used for speech denoising [176]. This method avoids the use of unreliable frequency components to reconstruct. Therefore, investigating the CS-based speech enhancement techniques that optimally suppress noise is a critical task [177].

Sparse decomposition of a signal, however, depends on the degree of fitting the data and the dictionary. A predefined dictionary can be obtained by an analytical approach based on a mathematical transform matrix, such as a Discrete Fourier Transform (DFT), Discrete Cosine Transform(DCT), Discrete wavelet Transform (DWT)[178] matrices, cepstrum, curvelets[179] and contourlets [180]. The reconstructed signal is fairly good by selecting the K largest valued sparse transform coefficients. Once a matrix or a high dimensional vector is transformed to a sparse space, different recovery algorithms like Basis pursuit(L_1-minimisation) [181], Orthogonal matching pursuit (OMP) [182], CoSaMP [183], or fast non-iterative algorithms can be used to recover the signal.

The present enhancement approach is rooted in transforming the time-domain signals into a suitable feature space, and then this feature space is sparse-coded using signal models called dictionaries. Sparse coding can decompose the noisy mixture into its structured components and attenuate any unstructured component (i.e., noise). Ultimately, the estimate of clean speech is obtained by applying inverse transform from the feature space back into the time domain.

Many dictionary-based techniques have been proposed, However, there is a dearth of reasonable evaluation methods to determine whether a dictionary is good enough in representing a signal with reduced dimensionality subspaces. To solve this problem, a set of dictionary evaluation measures is defined. These measures not only address the sparseness and reconstruction error of signal representation but also consider denoising and separating performance[183]. Normally, the performance of the speech enhancement algorithm is measured by subjective listening tests with human listeners. Objective measures are designed to approximate subjective quality scores and intelligibility rates. The majority of objective measures figure out improvement by comparing the (unobserved) clean speech with distorted speech and enhanced speech in a perceptually meaningful way.

3.3.2 Noise Reduction through Compressive Sensing

The fundamental principle of Compressive Sensing is that if a signal has a sparse representation in one basis, it can be reconstructed from fewer numbers of projections on to the second basis [162,164,165]. Therefore, it is possible to recover a signal with very few measurements than the conventional Nyquist rate, providing compression due to low memory storage space and transmission bandwidth. The capability of compression assumes sparsity and incoherence. Sparsity refers to the rate of information which is much lower than the rate suggested by its bandwidth [164].

Sparsity is defined by considering an NXN matrix Ψ with columns forming an orthonormal basis. Speech signals are k-sparse while noise factors are not and therefore these two components are theoretically separable by the method of CS [184].

Thus, a k-sparse signal, $x(n) \in \mathbb{R}_N$ can be represented as

$$x(n) = \Psi\, \theta(n) \qquad (3.1)$$

Where $\theta(n) \in \mathbb{R}_N$ has k-non-zero entries. The CS Measurement vector can be expressed as:

$$y(n) = \Phi x(n) = \Phi \Psi \theta(n) = \Theta \theta \, , \qquad (3.2)$$

where x(n) is an NX1 vector and Φ is an MXN sensing matrix/a linear mapping matrix. i.e., $\Phi : \mathbb{R}^N \rightarrow \mathbb{R}^M$ and Θ is a MXN CS matrix.

Here it is noted that as $M \ll N$, the numbers of columns are more than rows, and the dimension of y(n) is smaller than x(n). If the sensing matrix adheres to the Restricted Isometry Property (RIP), It can be shown that a k-sparse signal x(n) can be reconstructed from fewer measured samples.

3.3.3 RECONSTRUCTION ALGORITHM

The impressive property of CS is that if a signal is K–sparse (or, not exactly K-sparse), the quality of the recovered signal $\hat{x}(n)$ is as good as selecting only the K largest values before the calculations, and measuring them directly [185]. This problem is undetermined and convex to solve, as there are more coefficients than variables. Different applications suggest various norms (L_p) to optimize parameters. The best solution is achieved by L_0-norm i.e. p=0, but it is an NP-hard problem. For p=1, various algorithms have been developed to solve this problem through linear programming. The computationally efficient L_2-norm i.e. p=2, a relaxed version, gives a solution but is not sparse. L_1-norm i.e. p=1 gives sparse solutions and good reconstruction probabilities. [186],[187-190].

The reconstruction algorithms are broadly classified for CS, namely Basis pursuit (BP) [182] and orthogonal matching pursuit (OMP)[183], and Compressive Sampling Matching Pursuit(CoSaMP) are the commonly used algorithms. BP with denoising (BPDN) seeks a better approach for findind a solution for the reconstruction of the noisy speech signal.

$$\hat{x}(n) = \text{Arg min} \, \| x \|_1 \quad \text{s.t. } y = \Phi x, \qquad (3.3)$$

$$\min \| x \|_1 \text{ such that } \| \Theta x - y \|_2 \leq \varepsilon \qquad (3.4)$$

This problem is equivalent to:

$$\text{Arg min } \| \Theta x - y \|_2 + \lambda \| x \| \qquad (3.5)$$

OMP is a greedy CS recovery algorithm i.e., it is simple and easy to implement solutions to complex, multi-step problems by deciding which next step will be a better choice and will provide the most benefit. To reconstruct the input signal 'x', OMP uses an iterative greedy algorithm to select a column in the sensing matrix ϕ that constitutes a part of y. In every iteration, a column of ϕ is chosen which is highly correlated with the residual of 'y'. Then the contribution of that column is subtracted from 'y' and the same procedure is repeated for the residual of 'y'. After 'K' iterations the algorithm would have identified the correct set of columns. The residual at the end is reconstructed signal 'x'. which can exactly recover k-sparse signals using: $M = c * k \log(N/K)$ measurements.

Needell and Tropp[190] proposed CoSaMP, which is a greedy algorithm based on OMP. In this method signal x is reconstructed by obtaining a proxy 'p', where 'p' is obtained by multiplying Φ^* with the observation vector 'y', Therefore, to identify the location of the significant K components of Φ, it is sufficient to identify the significant components of 'p'. In every iteration, the algorithm selects the largest 2K components of 'p' and includes an index in the set. Then by applying the least square method, an estimate 'e' is obtained. Finally, the signal 'x' is obtained by selecting significant k components of the estimate 'e'.

3.3.4 Different Sparsity basis for speech signal

Speech signals are more complex and consist of many harmonics which have uncorrelated frequency components. Therefore, a speech signal is considered less sparse in the observed time

domain. The speech signal is to be examined for the best sparsity property using different types of transform basis.

To exploit the sparsity in speech signals, a comparative analysis regarding these dictionaries is given as follows:

i) DFT Dictionary

DFT is the most basic transform used as a sparsifying basis. An experiment was conducted to test the sparse domain using a Fourier basis. It is observed that due to the presence of a large number of non-zero valued frequency components, the Fourier basis is not suitable for sparse representation of speech signals. It also results in a complex number and therefore they are less sparse in the frequency domain. Further, it has been observed that to increase the level of sparsity, k, discarding frequency components below a threshold results in a large reconstruction error.

ii) DCT Dictionary

DCT, a real-valued transform matrix was explored to test the sparsity in the speech frame. In this case, a frame length of 4267 speech samples was observed. Several experiments were conducted with a varied number of CS measurements on the original number of samples. The DCT gives a more satisfactory reconstruction than DFT. Further, we observe from experimental results that the DCT transform basis gives a smaller MSE for the same number of used measurements, compared to DFT.

iii) DWT Dictionary

DWT uses a set of orthogonal basis functions with scaling and translation properties to capture the characteristics of the signal in both time and frequency domains. The observation from

the results confirms that DWT provides better sparsity than DFT since the transform has better localization properties.

iv) Cepstral Dictionary

Cepstrum domain analysis corresponds to homomorphic mapping which is used to get information on speech formant and periodicity. Since speech signals are generally sparse in Time–Frequency domain and many types of noise are non-sparse, the noisy speech can be decomposed and reconstructed. The Table 3.1 shows the capacity of sparse transform of noisy speech for various SNR values, sparsity: K=517, and measurement samples: M= 4317

Sparse Basis / **Input SNR dB**			**Output SNR dB**			
	-10	-5	0	5	10	15
DCT	-11.8	-6.35	4.18	7.84	13.29	13.32
DWT	-9.27	-3.51	6.51	10.19	12.75	13.11
FFT	-15.65	-7.46	-3.43	-3.06	-2.84	-2.86
CEPSTRUM	-18.13	-9.04	-6.41	-2.41	1.48	1.87

Table 3.1 SNR output of various sparse basis considering the sparsity and measurement samples.

The key result observed is that the number of measurements M required for successful recovery is different under different solution criteria.

The following experimental analysis explores the capacity of sparsity of DCT and DWT for various speech and noise signals. The sparsity is calculated in Tables 3.2(a-d) by considering the K-sparse threshold transform domain coefficients that carry q % of energy here q is taken 99, 90, 80, and 70 by thresholding respectively.

Threshold= 99%	DCT			DWT		
Input SNR dB	K	M	Output SNR dB	K	M	Output SNR dB
-10	1526	5102	-12.68	3179	8914	-9.27
-5	1365	4837	-7.62	2573	7306	-3.51
0	1129	4072	4.29	2110	6975	4.26
5	909	3278	8.17	1718	5679	8.36
10	765	2759	15.50	731	2416	12.15
15	691	2492	15.39	711	2349	12.66

(a)

Threshold= 90%	DCT			DWT		
Input SNR dB	K	M	Output SNR dB	K	M	Output SNR dB
-10	524	1510	-15.82	983	8914	-12.72
-5	356	1373	-8.21	739	7306	-6.15
0	255	919	0.81	513	1695	0.24
5	198	714	0.86	349	1153	0.54
10	145	523	1.12	249	823	2.32
15	134	483	2.67	248	819	5.31

(b)

Threshold= 80%	DCT			DWT		
Input SNR dB	K	M	Output SNR dB	K	M	Output SNR dB
-10	427	580	-17.26	436	917	-14.20
-5	259	438	-9.81	397	835	-7.01
0	103	371	-0.29	212	700	-0.93
5	79	284	-0.11	164	541	-0.85
10	58	209	0.65	129	425	0.27
15	54	94	1.82	131	432	0.7

c)

Threshold= 70%	DCT			DWT		
Input SNR dB	**K**	**M**	**Output SNR dB**	**K**	**M**	**Output SNR dB**
-10	74	301	-21.24	362	592	-18.07
-5	56	238	-11.37	239	483	-9.47
0	47	169	-1.28	110	363	-1.03
5	38	137	-0.82	89	293	1.26
10	29	104	-0.44	73	241	1.4
15	27	97	-0.35	74	244	2.86

(d)

Tables 3.2(a-d) Effect of threshold transform domain coefficients at a) 99% b) 90% c) 80% d) 70% on K, M, and output SNR of the DCT and DWT dictionaries.

From the simulation results, it is observed that the performance of the fixed dictionaries degrades at low (< 0 dB) input SNR for all threshold scenarios. This demonstrate that the capacity of sparse transformation on strong noisy speech influences the sparsity.

DCT gives high output SNR values than DWT while reconstructing. Though DWT gives the best sparsity value at low SNR, poor reconstruction is observed. Therefore, DCT basis is chosen in all the test cases. Figure 3.1 shows the results of applying DCT to sparsify the noisy speech signal and using a random sensing matrix, the noisy components are discarded. An L1-minimization recovery algorithm is used in reconstructing, the output of the sensing matrix.

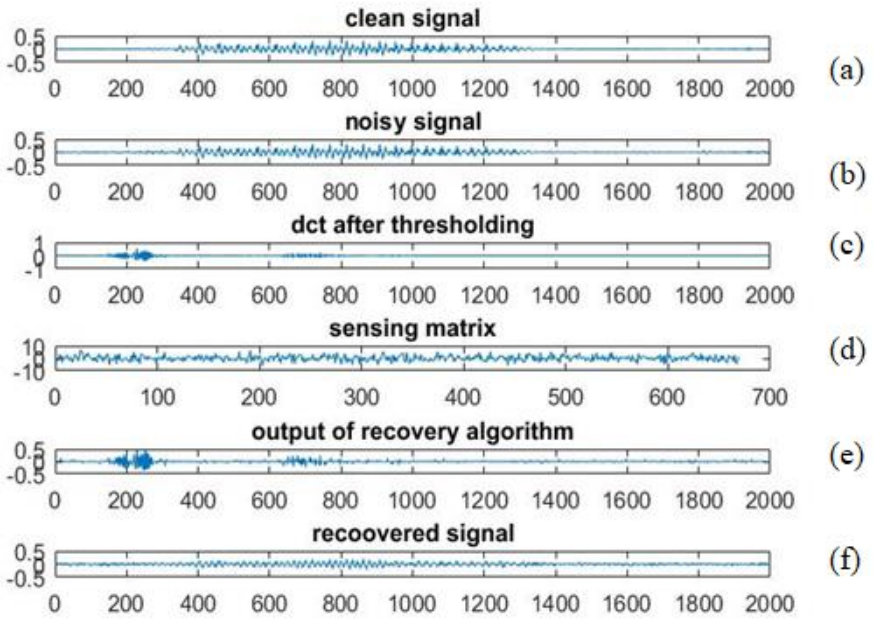


Figure 3.1(a-f) Results of applying DCT a) clean signal b) noisy signal c) DCT after thresholding d) Sensing matrix, e) Output of recovery algorithm f) recovered signal.

3.3.5 COMBINED BASIS

Further, speech enhancement based on compressive sensing by considering a combined basis is attempted and its performance is compared by using PESQ and MSE measures. Linear prediction (LP) is a commonly used technique that extracts the sparse LP coefficients and conveys large information about pitch and formants present in the speech signal. The speech signal is separated into voiced and unvoiced parts by using a zero-crossing rate (ZCR). Using a hamming window, the signal is divided into frames and then ZCR for each frame is calculated and frames are classified into voiced (V) and unvoiced (UV) then de-framing is then done so that voiced and unvoiced parts are separated. Four different pairs of the basis for both parts of the signal forming a combined basis scheme are tested here for analysis. Table 3.3 shows a comparison of experimental results of PESQ and MSE (Average values) for the frames in the speech signal corresponding to male and female speakers using combined Sparse basis pairs.

SIGNAL M = 100	SPARSITY BASIS	OMP		BP (L1- MIN)	
	Voiced(V), Un-Voiced (UV)	PESQ	MSE	PESQ	MSE
SP01(male)	DCT(V), LPC(UV)	1.832	1.58E-04	1.35	1.22E-03
	LPC(V), DCT(UV)	1.911	1.40E+05	2.09	3.59E-00
	LPC(V), FFT(UV)	1.457	6.20E+04	2.03	2.29E-00
	FFT(V), LPC(UV)	1.30	5.14E+04	1.83	4.30E-03
SP05(male)	DCT(V), LPC(UV)	1.370	4.18E+02	1.61	0.63E-00
	LPC(V), DCT(UV)	1.524	10.4E+02	2.32	8.25E-06
	LPC(V), FFT(UV)	1.439	2.84E+02	1.85	0.002E-00
	FFT(V), LPC(UV)	1.376	4.18E+02	2.22	0.002E-00
SP13(Female)	DCT(V), LPC(UV)	1.318	5.00E+02	1.76	6.54E-06
	LPC(V), DCT(UV)	1.875	4.71E+02	2.015	5.98E-07
	LPC(V), FFT(UV)	1.515	1.39E+02	1.817	4.98E-06
	FFT(V), LPC(UV)	1.317	5.00E+02	1.595	3.129E-05
SP15(Female)	DCT(V), LPC(UV)	1.205	1.91E+02	1.66	0.45E-01
	LPC(V), DCT(UV)	1.599	3.67E+03	2.24	5.82E-07
	LPC(V), FFT(UV)	1.199	3.67E+03	1.37	6.72E-05
	FFT(V), LPC(UV)	1.784	1.91E+03	1.98	2.98E-04

Table 3.3 : Comparison of experimental results of PESQ and MSE(Average values).

From Table 3.3, it is observed that LPC(V), and DCT(UV) pair give better results when compared to other combinations for both reconstruction algorithms with a fixed number of measurements. As the number of measurements increases, in both reconstructions, the SNR increases while MSE decreases. Finally, L1-Minimization gives the best reconstruction compared to OMP from all combinations.

The test methodology adopts a DCT dictionary with L_1 minimization (frame-by-frame analysis). From the simulation results it is observed that, though COSAMP is faster than OMP, the reconstruction through the OMP method is accurate. But when speech is corrupted by noise, COSAMP shows better reconstruction than OMP. For low SNRs, in all cases, L_1 minimization shows good performance.

3.4 Adaptive Dictionaries

Fixed-dictionary performance degrades with low (< 0 dB) input SNR. To optimize sparse signal representation, train a generic dictionary. Fixed (predefined) dictionaries are suitably structured for generic signals and are relatively easy to implement. As input SNR increases, the PESQ score also increases. The implicit DCT dictionary model performs best on sparse signals with noise. Since the signal dictionary is fixed, we can't denoise it entirely in non-stationary settings. Further, when the input SNR is low, the performance of fixed dictionaries degrades. The alternate optimization strategy is to train a generic dictionary for sparse signal representation. To increase performance over predefined dictionaries, efficient and flexible dictionary learning approaches like KSVD, and NMF are examined.

A dictionary learning technique finds a dictionary in such a way that all training signals have a suitable sparse representation in them by employing a training data matrix, which contains signals from a specific class of signals. A typical dictionary learning method specifically addresses the following issue: $\min \| Y-DX\|^2_F$

$$D \in D, \quad X \in x \quad \ldots\ldots\ldots \quad (3.6)$$

where D is the dictionary and x coefficient matrix are acceptable sets of data, respectively, and $\|.\|_F$ is the Frobenius norm. D is often described as the set of all dictionaries with unit column-norms 'x' guarantees that the coefficient matrix has sparse columns.

Notably, the aforementioned issue is not convex with regard to the pair (D, X). Most dictionary learning algorithms approach this issue by repeatedly carrying out the following two-step process: i) Sparse representation and ii) Dictionary update. The sparse representations of each training signal are generated using the current dictionary in Stage 1, which is just a regular sparse coding issue. For this step, a variety of sparse coding techniques have been suggested [210]. Stage 2, in which the dictionary is updated to decrease the representation error of stage 1, is the primary distinction between several dictionary learning techniques.

Another well-known and very effective approach is K-Singular Value Decomposition (K-SVD) [211]. Only one atom is modified at a time during the dictionary update step. The non-zero items in the corresponding row vector are likewise updated together with each atom's change. This results in an issue of matrix rank-1 approximation, which is subsequently resolved by using an SVD procedure. Though K-SVD is sequential like K-means, it fails to simplify to K-means by destroying the structure in sparse coefficients". This is due to performing SVD in K-SVD, which (unlike K-means) forces the atom norms to be 1, and the resulting coefficients are not necessarily 0 or 1.

3.4.1 Nonnegative Matrix Factorization (NMF)

In the presence of nonstationary noise, NMF could be employed to denoise speech. Using the technique of nonnegative matrix factorization, the matrix equation $V \approx WH$ is solved by determining the W and H which are locally optimum. As a result, it is possible to break down a signal into a convex set of nonnegative constituent parts. Smaragdis [25] demonstrated how NMF may be utilized to distinguish single-channel mixtures of sounds where the signal, V, is a spectrogram and the building blocks (basis vectors), W, are a collection of certain spectral shapes. This is done by associating various sets of building blocks with various sound sources. In Smaragdis' formulation, H takes the shape of the building components' time-varying activation levels. Each source is represented by a set of building blocks in W, and since H permits activations to change over time, this decomposition is well suited to modelling nonstationary sounds.

NMF is effective in separating sounds when the constituent parts of several sources are entirely diverse. When the constituent parts of several sources are entirely diverse, NMF is effective in separating sounds. The building blocks for one source will not be very useful in characterizing the other, for instance, if one source, like a flute, only produces harmonic sounds while another source, like a snare drum, only produces nonharmonic sounds. However, there is substantially less distinction between sets of building components in many situations of practical significance. Human speech, for instance, comprises harmonic sounds (potentially at various fundamental frequencies at various times) and nonharmonic sounds, and it can have energy throughout a large frequency range. For these reasons, the speech building blocks can at least partially reflect numerous interfering sounds. NMF is entirely incapable of reconstructing the individual sources since the basis functions for speech and noise are identical. NMF is entirely incapable of representing the individual sources since the basis functions for speech and noise are identical.

Under entirely non-stationary noisy environments such as Traffic noise, machine noise, babble noise, crowd noise, etc., both KSVD and NMF approaches produce poor results. The performance of techniques based on Non-negative Matrix Factorization (NMF) is better than K-SVD [131] for non-stationary noise. NMF is a supervised method and has the problem that it needs a priori knowledge of speech and noise. When unseen real-world noise is encountered, the NMF technique fails.

3.5 Robust Principle Component Analysis (RPCA)

Robust Principle Component Analysis is a unique data analysis technique that has been shown to be effective for noise-corrupted data. The advantage of RPCA is that it doesn't require any previous knowledge of noise and hence can be used in unseen real-world noisy environments. Moreover, it can perform well under strong noisy conditions. Therefore, RPCA has been widely used for robust speech representation algorithms. In [30], RPCA is applied to the spectrogram of speech signals, and the resulting sparse component is shown to contain low levels of noise and thus be noise-robust.

The low-rank and sparse matrix decomposition model which is employed as the core of the speech enhancement method called Robust Principal Component Analysis (RPCA) is also employed for unsupervised separation [198,199]. Furthermore, because it accounts for deviations of speech and noise time-frequency (T-F) matrices from the idealistic sparse and low-rank model based on bilateral random projections, the Semi-Soft Go Decomposition (SS-GODEC) [200] will be treated as the matrix decomposition at the core of the enhancement process. In this paper, we perform analysis of the low-rank and sparse matrix decomposition techniques to separate speech signal from the noisy speech in an unsupervised way. The suggested method separates the noisy speech spectrogram Y into three submatrices, Y = L+S+E, where L, S, and E represent noise, speech, and residual noise matrices respectively. This method differs significantly from RPCA [201,202], which assumes Y = L+ S and can be applied to a wide range of problems.

Speech signals were recovered from noisy recordings by decomposing the spectrogram of the input signal into a low-rank noise estimate and a sparse activation matrix of a dictionary of target speech templates, using SS-GoDec as its centerpiece. For enhancement of speech signals, this work assesses the method of [202] with RPCA [199] and Semi-Soft GoDec[200] as its basis. In doing so, special attention is paid to the STFT parametrization within this algorithm and its impact on the enhancement method's outcomes. Log-sigmoid soft masks [203] were considered and tested in addition to binary T-F masks, which were considered for the masking step of the enhancement algorithm in [201,202]. Test signals were also studied to determine if the low-rank and sparse model is suitable for speech and noise signals. Finally, the SE method's performance with RPCA and SS-GoDec was compared.

3.5.1 RPCA MODEL FOR SPEECH ENHANCEMENT FRAMEWORK

RPCA is concerned about the following problem: Given a matrix $M \in \mathrm{R}^{n1 \times n2}$,and decompose it as the sum of a low-rank matrix $L_0 \in \mathrm{R}^{n1 \times n2}$, with rank(L_0) < min(n_1, n_2), and a sparse matrix $S_0 \in \mathrm{R}^{n1 \times n2}$,

$$M = L_0 + S_0 \qquad \ldots\ldots \quad (3.7)$$

The aforementioned convex problem cannot be solved. Furthermore, it is an NP-hard problem since information on the low-rank and sparse components is not known in advance. A technique called Principal Component Pursuit (PCP) is suggested in [197] as a solution to the aforementioned RPCA separation problem. The estimations for L_0 and S_0 obtained via PCP are shown to be precise under a variety of circumstances. PCP successfully distinguished between low-rank and sparse components with great reliability and accuracy. Additionally, PCP has demonstrated encouraging outcomes when used for low-rank and sparse matrix decompositions in video surveillance and facial recognition tasks. This motivated the present study to examine the potential of RPCA for use in signal-processing applications.

PCP accomplishes low-rank and sparse matrix decomposition by solving the convex optimization problem:

$$\arg\min \left(||L||_* + \lambda\, ||S||_1\right) \text{ under the constraint } M=L+S \qquad \ldots\ldots \quad (3.8)$$

where λ is a scalar. The parameter λ is used to balance the two opposing minimizations' priorities.

The approximated values of $\hat{L}, \hat{S}$ are calculated using Eq (3.9) as follows:

$$\hat{L}, \hat{S} = arg\, min_{L,S} \left||L|\right|_* + \lambda\left||S|\right|_1$$
$$s.t \quad M - L - S = 0 \qquad \ldots(3.9)$$

By using the Lagrange method, a Lagrange multiplier Y is associated to produce an unconstrained function. The optimum values of L and S are found in an iteration using Y value

from the last iteration. Thus, in this way, the values of L, S, and Y are updated to reach the global optimum.

The higher the value of λ, the sparser $\hat{S}$ will be at the cost of low rank estimates $\hat{L}$. The smaller the value of λ is chosen, the better the low-rank estimates $\hat{L}$ will be at the expense of less sparse $\hat{S}$. Solving this optimization problem yields estimates of $\hat{L}$ for L_0 and $\hat{S}$ for S_0. Thus, in this way, the values of L, S, and Y are updated to reach the global optimum.

The decomposed components, however, must be nonnegative to preserve the physical interpretation of a spectrogram, hence designers provide the following Nonnegative RPCA(NRPCA) model:

$$\hat{\mathrm{L}}, \hat{\mathrm{S}} \;= arg\, min_{L,S} \left|\left|L\right|\right|_{*} + \lambda \left|\left|\mathrm{S}\right|\right|_{1}$$

$$s.t \quad M = L + S\,, L \geq 0, S \geq 0 \qquad \ldots(3.10)$$

Figure 3.2 Illustrates an overview of RPCA based SE approach that decomposes noisy input speech into sparse 'S' and low rank ' L' components.

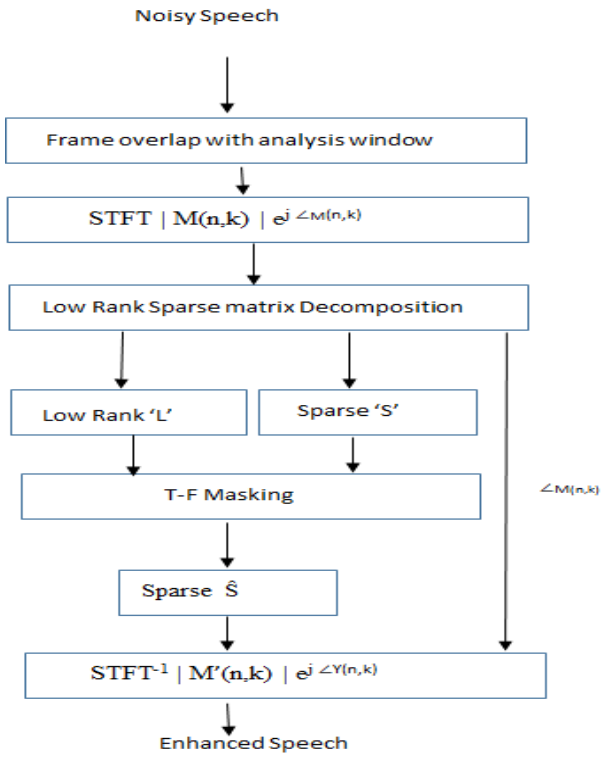


Figure 3.2 Overview of RPCA based SE frame work

The algorithm for the RPCA based SE is as follows:

Algorithm 1: Speech Enhancement by NRPCA

Input: Observation data M

1: Initialize $\mu_0 > 0$, $\lambda > 0$, $\rho > 1$, $\theta > 0$, $k = 0$, $L_0 = M$, $Y_0 = 0$;

2: do

3: $S_{k+1} = \arg\min_S \lambda \|S_k\|_1 + \mu_k / 2 \| M + \mu_k^{-1} + Y_k - L_k - S_k\|_F^2$;

4: $L_{k+1} = \arg\min_L \|L\|_* + \mu_k /2 \|M + \mu_k^{-1} Y_k - S_{k+1} - L\|_F^2$;

5: $Y_{k+1} = Y_k + \mu_k (M - L_{k+1} - S_{k+1})$;

6: Update $\mu_{k+1} = \rho * \mu_k$;

7: $k = k + 1$;

8: while $\|M - L_{k+1} - S_{k+1}\|_F / \|M\|_F > \theta$

9: Output: Matrix $L = L_{k+1}$ and $S = S_{k+1}$;

3.5.2 Speech Enhancement Method Using SS GODEC

GODEC: Another matrix decomposition that will be investigated and compared to NRPCA in terms of applicability for SE techniques is the Go Decomposition [200]. Although NRPCA and GoDec have similar principles, there are a few ways in which GoDec differs from NRPCA.

According to preliminary testing, the original RPCA technique [201], which decomposes noisy speech spectrogram into two submatrices, is not robust or effective enough to extract the formant structure of clean speech. As a result, we enhance the original RPCA approach by making the noisy speech spectrogram Y the superposition of L, S, and E; Y = L+S+ E. We assume that L is in a low-rank subspace, that the speech structure is sparse, and that R is a noise term that perturbs the ideal rank and sparse character. The assumption is based on the fact that the noise spectral pattern is always repeated, whereas the speech signal has more diversity and is relatively sparse within the noise. Under the perturbation of E, the aim is to recover the low-rank matrix L and sparse matrix S from Y.

The original Go-Dec and Semi-Soft Go-Dec are two separate variants of the Go-Dec. For the original Go-Dec [200] technique to find approximations L and S with these features that minimize noise power, the rank r of L and the cardinality c of the support set of S must be predefined. It is difficult to decide on suitable values for c and r in advance for a use of Go-Dec in a practical application, such as the core of an SE algorithm. If the choice is too large, some of the noise R can leak into L or S. Choosing them too low could result in the noise term being incorrectly allocated to portions of L or S. Further, poor decisions for c and r could result in L's components leaking into S, or vice versa. Therefore, it is necessary to investigate which values of c and r are appropriate choices for each situation that Go-Dec should be used in.

Furthermore, suboptimal choices for '*c'* and '*r'* could cause parts of L to leak into S or vice versa. Hence, for each respective context that GoDec is used in, it is necessary to explore which values of 'm' and 'r' are suitable choices. The Semi-Soft GoDec algorithm just needs r as an input and automatically determines c. Semi-Soft Godec offers two significant advantages over traditional

Go Decomposition. The possibility of selecting c erroneously is first and foremost eliminated by obtaining a suitable c automatically. Second, compared to original GoDec, the Semi-Soft Godec time cost is substantially less [204]. Another key difference between RPCA and GoDec is that for PCP, neither 'c' nor 'r' needs to be predetermined, and the estimates for L_0 and S_0 are completely unconstrained.

3.5.3 Optimization in GoDec

The original GoDec optimization problem is written as :

$$\arg min_{L,S} \| M - L - S\|^2 \; s.t \; rank(L) \leq r \quad and \; card(S) \leq m \qquad \ldots\ldots (3.11)$$

and results in low-rank and sparse estimates $\hat{L}$ and $\hat{S}$.

So, L and S have to be chosen such that they meet the predefined conditions on their rank and cardinality of their support set while the noise power $\|E\|^2 = \| M - L - S\|_F^2$ is minimized.

As the cardinality S is hard to estimate, by using soft threshold λ for matrix decomposition, the optimization problem is formulated in Eq(3.5) as follows:

$$arg \; min \; \|M - L - S\|^2 + \lambda \|S\|_1 \; s.t \; .rank \; (L) \leq r \qquad \ldots(3.12)$$

Additionally, it generates the sparse and low-rank estimates Lˆ and Sˆ.

The objective function in this case is the sum of the noise power and the l_1-norm of S, scaled by a balancing factor λ. The l_1-norm is employed as a measure of the number of large entries in S and hence as a measure of sparsity S to some extent.

The number of relevant entries of S in the objective function is traded off with one another by the parameter λ, as it is necessary to reduce the noise power. According to the trend, the higher

the value of λ, the sparser the value of $\hat{S}$, and vice versa. This enables the self-determination of a number of significant components in S, which is useful when deciding on a good value for λ. In Section 3.9, the value of λ will be explored to see if it is appropriate to utilize Semi-Soft GoDec in a speech enhancement method. There will also be a search for the best option for *r*.

Although the restrictions of (3.11) are rather irregular and difficult to formulate, the algorithm that will be explained in the following part will be able to solve the SSGoDec optimization problem.

Figure 3.3 illustrates an overview of SS-GODEC based SE approach that decomposes noisy input speech into sparse 'S', low rank ' L' and a residual noise 'E' components.

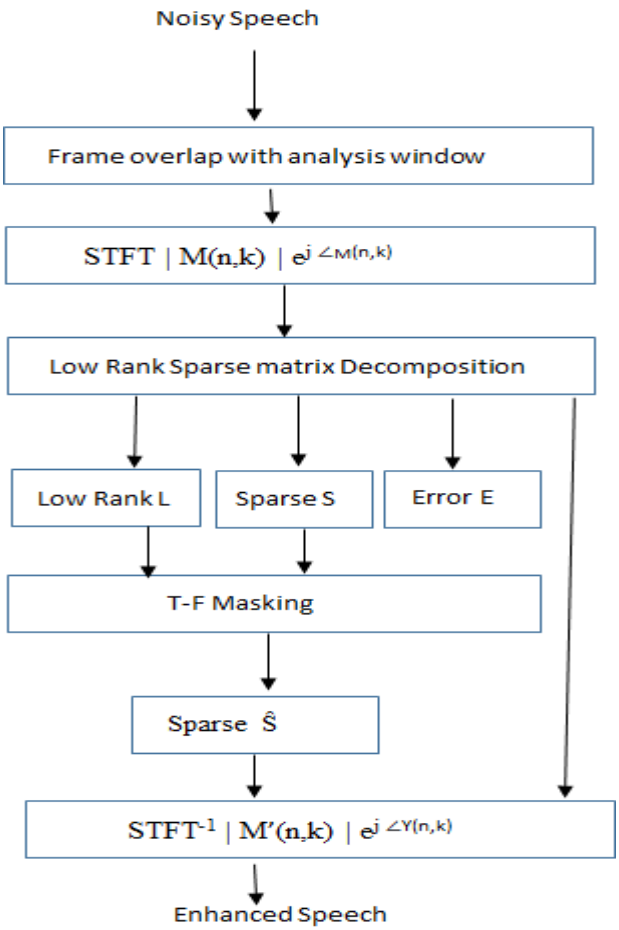


Figure 3.3 Overview of SS-GODEC based SE framework

3.5.4 Fast GoDec Algorithms

The algorithm given in [200] solves the problem (3.6) by addressing the two subproblems alternately.

$$L_{k+1} = \arg\min \| M - L - S_k \|_F^2 \text{ and } S_{k+1} = \arg\min \| M - L_{k+1} - S \|_F^2 \quad \text{s.t } \operatorname{rank}(L) \leq r \ \operatorname{card}(S) \leq c \quad (3.13)$$

In particular, Singular value hard thresholding of $(M\text{-}S_K)$ is used to tackle the first subproblem. This involves computing a singular value decomposition $(M\text{-}S_k) = U \Sigma V^T$. Then, by setting all but the r largest singular values in Σ to zero, determines, $\tilde{\Sigma}$. Finally, $Lk_{+1} = U \tilde{\Sigma} V^T$ is derived as the updated estimate of L_k. This gives a rank r approximation of $(Y\text{-}S_K)$. Similarly, the second subproblem is addressed using $(M - L_{k+1})$ entry-wise hard thresholding. This means that the updated estimate S_{k+1} is $(M\text{-}L_{r+1})$, with all entries saving the m biggest set to zero.

An algorithmic framework for this strategy looks like this:

Algorithm2 for Semi-Soft GoDec

Input M, r, λ; Initialize S_0;

1.for k = 0, 1, 2, . . .

2. L_{k+1} = singular value hard thresholding of

$(M - S_k)$;$S_{k+1} = S_\lambda(M - L_{k+1})$;

3. if the final convergence test is satisfied,

STOP with final estimates L_{k+1} and S_{k+1};

4. end (for)

In this study, S_0 is initialized as a matrix with all zeros. The convergence test is satisfied when

$$|\mathrm{M} - L_{k+1} - S_{k+1}|_F \leq 10^{-2} \quad \ldots \quad (3.14)$$

holds.

An SVD of a $n_1 \times n_2$ matrix requires min $(n_1^2 n_2, n_1 n_2^2)$ flops which is expensive for large matrices. To decrease the computations, a method based on bilateral random projections is proposed to calculate a rank 'r' approximation of a matrix, and this method is used to approximate $(\mathrm{M} - S_k)$ in eq(3.14) in place of the singular valuehard thresholding[200]. This low-rank approximation requires $r^2(n_1 + 3n_2 + 4r) + 8n_1 n_2 r$ flops per iteration. This approach's computational cost scales linearly with the matrix's dimensions, which saves time when the matrix is huge. Semi-Soft GoDec decomposes matrices using the bilateral random projection-based low-rank approximation approach.

Because the PCP method in section 3.4.1 requires a singular value decomposition, the Semi-Soft GoDec technique will require significantly fewer flops per iteration for big matrices than PCP approach. The sizes of the matrices that have to be decomposed for the speech enhancement experiments in this work are presented in section 3.10. The number of flops required for one singular value decomposition in the RPCA-based speech enhancement algorithm and one low-rank approximation in SS GoDec-based speech enhancement algorithm will be compared. Also, the costs of computing one recovered speech signal for both strategies will be shown and compared.

3.5.5 Matrix Decomposition for Speech Denoising

The recovery of speech signals from noisy recordings is difficult in general, but it becomes considerably harder if only one microphone is being used to record both the speech and the noise. Additionally, there is no spatial information available that could be utilized to distinguish between the two signals. In this monoaural situation, it is essential to take into account the distinct tonal characteristics of both speech and noise that enable separation between speech-like and noise-like components of the mixed signal. The SE algorithms tested in this work were unsupervised and

untrained, and based on a very innovative model for the speech and noise components. It will be assumed that the noise contribution is approximately low-rank and that the speech contribution is approximately sparse in its T-F representation. The low-rank and sparse matrix decompositions that were looked at in the previous section should be able to distinguish between speech and noise if this model is adequate

3.5.6 Considerations of the Time-Frequency Settings

STFT settings impact the effectiveness of matrix decomposition-based SE. The works of [201,202,205] missed this fact. The findings of these experiments show that parameterizing the STFT is crucial for SE. The effect of transform windows on the spectrogram that arises in low-rank or sparse approximation was investigated. The success of SE technique for calculating the T-F matrix differs slightly for rectangular, Hann, and Blackman windows with a fixed length N. The results of experimental evaluations on rectangular, Hann and Blackman windows for a fixed length of N have revealed that the success of the SE algorithm for calculating the T-F matrix differs slightly.

The length N of STFT influences the T-F frequency resolution. Longer N provides more frequency bins. Both T-F matrices contain speech and noise information. If an SE technique based on the low rank and sparse structure of noise and speech signals is used, a high resolution is needed to capture those qualities in the spectrogram. Through a low-rank and sparse matrix decomposition, the resolution shows how well speech and noise spectrogram components can be separated. STFT lengths of 512, 1024, 1536, and 2048 were used to assess the enhancement algorithm. As N increases, performance generally improves. The best STFT length was M = 1536, and the worst was N = 512. N = 2048 decreases the SE algorithm's performance. Hop size determines spectrogram resolution. It should be set high enough to differentiate noise and speech accurately in the masking step. After enhancement, experimental observations indicated variations in recovered speech signals for different hop sizes.

Developing and applying T-F masks improves the enhanced speech signal's SNR. Every T-F bin in a binary mask is either speech or noise. If the suggested SE technique only includes a few frequency bins per time step, binary masks won't work [8,15]. As speech and noise bands may overlap, log-sigmoid masks were therefore considered.

The mask g was defined using the results of the matrix decomposition S and L. This matrix was used to estimate speech and noise in the T-F representation of the original noisy speech spectrogram.

$$|\hat{M}_{\text{speech}}(m,n)| = |\,M_{NS}\,(m,n)\,|\,g(m,n), \quad \ldots \quad (3.15)$$

$$|\hat{M}_{\text{noise}}\,(m,n)| = |\,M_{NS}\,(m,n)\,|(1-g(m,n)). \quad .. \quad (3.16)$$

The method used to calculate the masks from L and S is different for binary and log-sigmoid masking.

3.5.6.1 Binary Masking

Binary mask entries are 0 or 1. When the matrix is applied to |MNS|, each element of |MNS| is either fully assigned to speech estimate $\hat{M}_{\text{speech}}$ or noise estimate $\hat{M}_{\text{noise}}$. If two signals' T-F representations don't overlap, an IBM can isolate them from a noisy input signal. This condition is called W-disjoint orthogonality or WDO in short [206]. Binary masks won't accomplish perfect separation in the proposed SE technique as speech and noise don't meet the WDO requirement. While [206] reveals two or more speech signals exhibit W-disjoint orthogonality, noise and speech do not. Many noises excite the same frequency bands as speech, causing T-F matrices to overlap. As the IBM is unknown, Mask M is estimated as the best mask. This estimate assumes noise interference is already suppressed in the matrix decomposition's sparse output S. When (m, n) in |S| is larger than (m, n) in |L|, the entry of $|\,M_{NS}\,|$ is assigned to the speech estimate $|\hat{M}_{\text{speech}}|$. Otherwise, $|\hat{M}_{\text{speech}}|$ speech's T-F bin is set to zero. Since matrix decomposition

does not ensure a clean pre-separation of the mixed spectrogram, mask estimate M deviates from IBM.

3.5.6.2 Log-Sigmoid Masking

In the case of mask uncertainties, overlapping speech, and noise T-F representations, a binary hard decision mask may not be the optimal solution. [203] Suggests using log-sigmoid soft masks in a similar circumstance to decompose a combination of speech signals. [203] shows that log-sigmoid masks outperformed binary masks in ASR experiments. Log-sigmoid masks were used in the speech enhancement process for this work. Section 3.7.2.1 compares the method's performance with no mask, binary masking, and log-sigmoid masking. The results show variations between the three alternatives. Log-sigmoid mask entries are in the range [0; 1]. However, the transition between 0 and 1 is a slope whose steepness is regulated by α. In preliminary trials, $\alpha = 1.6$ proved highly suitable for the SE algorithm and was employed throughout this work.

3.6 Simulations and Results

This section evaluates SE algorithms from experiments. These studies have considerable practical significance for assessing enhancement's potential and limitations. They calculate the influence of variables considered in sections 3.4 and 3.5. Due to uncertainty and complex interactions, the previous theoretical discussion was insufficient to predict the speech recovery technique's performance. The experimental results show how well the algorithms performed in testing, as well as validating the assumption that they will perform similarly in comparable scenarios.

3.6.1 Framework

The speech corpus [207], a collection of male and female speakers who utter columns of figures, was utilized to record the speech signals. The signals are provided as Wav files

with a 20 kHz sampling rate. For the tests, 20 of these files—10 with male speakers and 10 with female speakers were selected. The noise collection was used to extract the noise signals [208]. Additionally, the signals were available in Wav- format with a 20 kHz sampling rate. Five noise recordings were chosen for evaluation: a car driven in traffic, wind, a machine, a bubbling stream of water, and a crowd of people cheering.

After that, each of the five noise signals was combined with a corresponding speech signal at speech-to-noise energy ratios of -10dB, -5dB, 0dB, 5dB, and 10dB. This resulted in an overall number of $5 \cdot 20 \cdot 5 = 500$ mixed test signals, which were all about 3 seconds long each. It should be noted that all settings were kept unchanged, apart from those that were altered to assess their impact on the effectiveness of the speech enhancement approach. Particularly, the RPCA mask's settings for $\lambda = 1$, gain = 1, and power = 1 remained the same. In this section, the experiments that were conducted for this work are described precisely to make the results obtained meaningful.

3.6.2 Evaluation of the RPCA-Based Speech Denoising Method

This section presents the assessment results for various parameter settings for the RPCA-based SE technique that were discussed in Sections 3.4 .The results were analyzed to determine the factors that affect the performance of SE algorithms, the best values for these factors, and whether some noise types help speech recovery more than others. The effects of T-F masks will be examined first. The section that follows will explore the results of different STFT parameter selections. To determine if the results of the first two parts can accurately predict how well the RPCA-based SE algorithm will perform for various noise kinds, the results of the first two sections will be coupled to the analysis of the noise and speech signals with m_l and m_s.

3.6.2.1 Influence of Binary and Log-Sigmoid Time-Frequency Masking

The SE technique with RPCA matrix decomposition was applied to all 500 test signals. To evaluate and compare the various effects of the masks experimentally, the enhancement strategy

was tested with binary masks, log-sigmoid masks, and without masks using fixed STFT settings of M = 1024, hop-size = 256, and Hann windowing. Table 3.4 is a list of the bss-eval parameter's findings [209]. The average SDR was provided for all five types of noise, all five input SNRs, and all three types of masks. It was obtained by averaging the SDR values of all 20 speakers (no mask, binary mask, and log-sigmoid mask).

For all noises and all masks, the approach with the aforementioned settings produces an SDR in the speech estimate that is noticeably higher than the SNR of the input signal. Most entries with input SNR of -10dB and all entries with input SNR of -5dB, 0dB, 5dB, and 10dB exhibit this behaviour. The enhancement strategy only failed to further enhance speech quality for a small number of entries at the already high input SNR of 10dB and instead lowered it. The SE approach based on RPCA performed as expected by reducing background noise effectively even when the noises were unpredictable like the sound of a cheering crowd or a stream of bubbling water. The graph shows that all three masks provide results that are quite similar for low-input SNR values, while the output SDR values become slightly more dispersed for higher-input SNR values.

All 500 test signals were decomposed using RPCA. The enhancement technique was tried using binary masks, log-sigmoid masks, and without masks using M = 1024 with a hop-size of 256 and Hann windowing. The average SDR plots for all five noise classes are displayed in Fig.3.4 for input SNRs ranging from -10 dB to 10 dB using three different masks. The above settings result in an estimate of SDR that is greater than the input signal SNR for all noises and masks. This tendency is seen in all -5dB, 0dB, 5dB, and 10dB entries as well as the majority of -10dB entries. At 10dB input SNR, the enhancement method only slightly decreased speech quality for a few instances. Even in the presence of demanding or unexpected stimuli, such as a cheering crowd or gushing water, the SE approach based on RPCA can reduce background noise. With low input SNR, all three masks yield similar results; however, for high input SNR, the output SDR values become more dispersed.

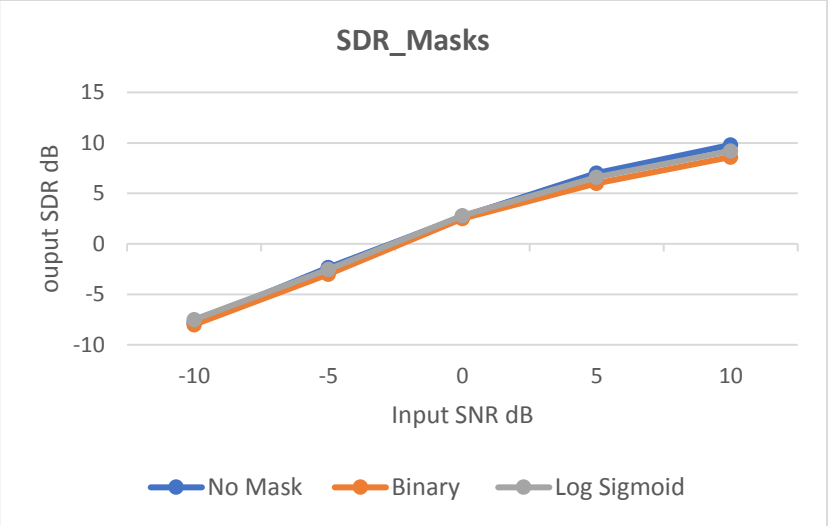


Fig.3.4 Influence of Masks on SDR (averaged for all five noise classes)

The SE approach uses no mask that performs best for high SNR values, followed by binary masking in the next place and log-sigmoid masking in the third. This demonstrates that the masks that were taken into consideration in this work cannot improve the RPCA decomposition further, which already accomplished a fair separation of speech and noise. Instead, the masks lead to unwanted changes that degrade the outcomes. The results are very similar at low SNR levels, with log-sigmoid masking performing the best. As a consequence, masking can significantly enhance the outcomes of speech recovery in extremely noisy environments.

Furthermore, Figure 3.4 demonstrates that binary masking consistently performs around 0.5 dB worse than log-sigmoid masking. The figure clearly shows that the three different mask types respond similarly to each of the different noise kinds, deviating relatively little from the mean of all noise signals. This indicates that none of the noise signals being used has a particular type of mask that is specifically suited for a particular noise type and performs noticeably better for this noise than for others and noticeably better than the other masks do for this noise. Finally, Figure 3.5 demonstrates how the SE technique frequently increases SNR. The SDR is increased by around 2.2–2.6 dB when SE uses the RPCA method without masking in -10dB to 5dB range of the input SNR.

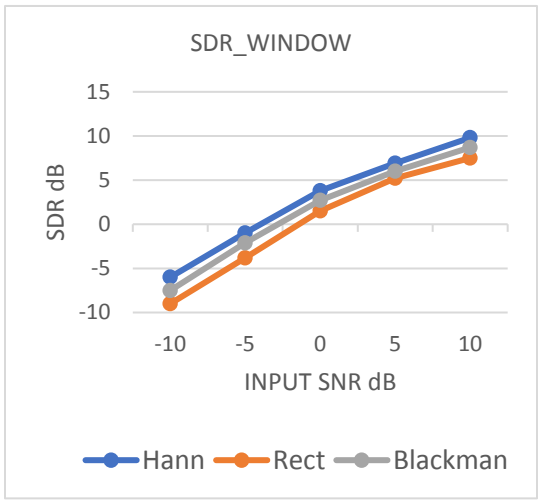
To assess the effectiveness of RPCA-based SE technique, the empirical perception that might be obtained by hearing the recovered speech signals. This is because a straightforward assessment of speech quality, like the bss-eval [209] SDR, cannot fully capture all relevant information. Again, the most crucial element is that the noise truly seems to have been significantly reduced in speech estimations. The improvement method works with all input SNR levels and all five different types of noise. The residual noise in speech estimations just sounds considerably softer without masking. Noise can be successfully eliminated by using binary and log-sigmoid masking. For any input SNR level or noise type, there was no noticeable change in the speech-to-residual noise ratio between the three mask variants. The results would have shown that at high input SNR levels i.e. above 5 dB, the speech-to-residual noise ratio would be better without masking than that achieved with masking, but the auditory tests were unable to confirm the data. Another false impression that could be drawn from the data is that at already high input SNR levels, the SE technique is unable to further reduce the noise. It is incorrect. The noisy speech input signal is believed to be less noisy than the speech estimates produced under these circumstances. It is obvious that the improvement is less noticeable than for input SNR levels with lower levels. The remaining noise is decreased and its sound is altered using binary and log-sigmoid masking. It ceases being unique and is no longer as clearly recognizable as, for instance, a crowd of people cheering or a car moving in traffic. It sounds strange and artificial. The recovered speech signals are therefore unpleasant to hear compared to speech signals that were recovered without masking that sound more natural.

3.6.2.2 Effect of the Short-Time Fourier Transform Parametrization

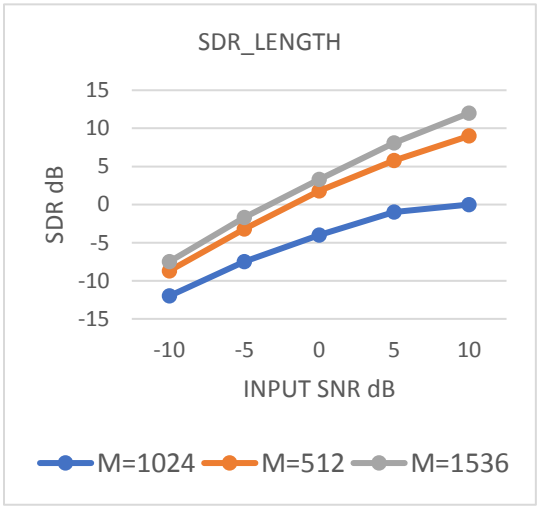
The following sections examine how STFT settings affected SE algorithm's performance. STFT window type, hop size, and length M were considered. For each of the three parameters, the tests, bss-eval findings [209], and auditory impressions of recovered speech signals were reported. The significance of the test results is addressed as a conclusion.

To evaluate the STFT window, M = 1024 and hop-size = 256 were fixed. The SE approach

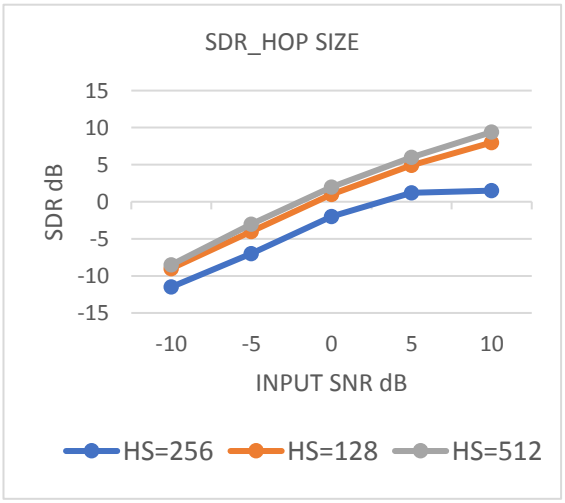
with log-sigmoid masking was tested for Hann, Blackman, and rectangular windowing using all 500 test signals. Figure 3.5(a-c) illustrates the average values of all 20 speakers for the five types of noise vs input SNR. Figure 3.5 (a) illustrates SE algorithm results utilizing three different STFT windows. Despite all three window types providing T-F representations that incorporate full input signal information, the SDR levels for different types of noise were still high.



(a)



(b)



(c)

Figure 3.5 a) Influence of the STFT window type on the output SDR as a function of the input SNR
b) Influence of the STFT length M on the output SDR as a function of the input SNR.
c) Influence of the STFT Hop Size on the output SDR as a function of the input SNR

3.6.2.3 Influence of the STFT window

Rectangular windowing, which is basically no windowing, provides the best performance and is, on average, 1.45–2.3 dB better than Blackman windowing, which results in the worst results. SDR results for bss-eval are consistently 0.6dB lower for Hann windowing than for rectangular windowing [209]. Rectangular windowing provides the best results on average, while Blackman windowing yields the poorest results. Figure 3.5(a) illustrates that Blackman windowing is ineffective at increasing input SNR values (a). At an input SNR level of 10 dB, the technique reduces SDR on average by more than 1.86 dB, while the enhancement strategy using Blackman windowing already fails to improve the SDR at an SNR of 5 dB. It must be remembered that the Blackman windowing enhancement technique did not produce substantial improvement for various noise types even at the input SNR level of 0dB. (Machine and water noise).

The major point from the figure applies to windows and is comparable to the one that was determined for T-F masks. The behavior of the three window types for different noise kinds does not significantly differ from the average of all noises. This indicates that none of the used noise signals has a window type that is particularly appropriate for this particular noise type and performs for it both significantly better than for the others and better than other masks perform for it. The auditive tests also highlight several elements that bss-eval SDR value was unable to capture.

The listening tests support the ratings of the three window types that were indicated by the numbers: Rectangular windowing produces recovered speech signals with the best noise suppression. Blackman windowing seems to generate more noise in the speech signals overall than the other two windows together. However, there aren't many variations, between Blackman windowing and rectangular windowing as the figures have indicated.

In contrast, the energy ratio was noticeably worsened by the bss-eval SDR values, especially for Blackman window. Listening to the recovered speech signals that were generated via rectangular windowing revealed the buzzing sound in the speech estimates as one important characteristic. All recovered speech signals that were generated via rectangular windowing included this sound, which is

about as loud as the background noises and a little unpleasant to hear. Hann windowing, which ultimately made the best impression, was be employed.

The numerical ratings for the three window types were supported by the listening tests. Speech signals without noise were produced using rectangular windowing. The speech noise coming from Blackman Windowing looked louder than the other two. Blackman windowing and rectangular windowing were not as distinct as the data suggests. As demonstrated by the masking experiment, the approach reduces noise for all three window types at a 10dB input SNR level. The energy ratio was made worse by bss-eval SDR values, notably for the Blackman window. The buzzing sound in the speech estimates was audible when listening to reconstructed rectangular windowed speech signals. All rectangle windowed recovered speech signals contained this irritating sound. Following will use Hann windowing made the best impression.

3.6.2.4 Influence of the STFT Length

The effect of STFT length M on the outcomes of the SE algorithm is examined now. This was accomplished by decomposing the test signals using the enhancement program's Hann windowing and log-sigmoid time-frequency masking during the STFT calculation.

The hop size of ¼ M was determined in accordance with the STFT length. The values of M = 512, 1024, and 1536 were utilized for the STFT length itself.

The results of these tests' bss-eval [16] SDR tests are once again presented in table 3.4 in the following format:

Noise	dB	M=512	M=1024	M=1536	Noise	dB	M=512	M=1024	M=1536
Crowd	-10	-12.576	-9.379	-8.353	Water	-10	-13.795	-8.358	-6.868
	-5	-6.567	-3.185	-2.497		-5	-7.984	-2.703	-1.355
	0	-1.740	2.077	2.871		0	-3.052	2.284	3.766
	5	1.125	6.292	7.742		5	0.389	6.322	8.386
	10	2.460	9.205	11.827		10	2.127	9.156	12.192
Wind	-10	10.590	-8.022	-7.462	Machine	-10	-10.172	-6.649	-5.876
	-5	-5.485	-2.493	-1.906		-5	-4.509	-0.759	-0.191
	0	-1.351	2.436	3.270		0	-0.615	3.958	4.893
	5	1.248	6.387	7.897		5	1.602	7.558	9.323
	10	2.391	9.121	11.671		10	2.583	9.859	12.848
Traffic & Car	-10	-8.253	-6.350	-5.716					
	-5	-2.933	-0.535	0.117					
	0	0.544	4.257	5.163					
	5	2.215	7.823	9.475					
	10	2.848	10.061	12.862					

Table 3.4: SDR-levels for different noise types after the SE algorithm with different STFT lengths

The average values across all 20 speakers and each of the five forms of noise are presented as a function of input SNR. M = 1536 comes first, M = 1024 comes second, and M = 512 comes third for all entries and averages. None of the noise signals employed has an STFT length that is particularly suited for this sort of noise and performs much better for it.

3.6.2.5 Influence of the Hop-Size

In this series of tests, an STFT with a length of M = 1024, and log-sigmoid masking were employed. Then, the hop size was changed between 128,256 and 512, i.e., 1/8M, 1/4M, and 1/2M.

Yet again, the recovered speech signals were assessed using the bss-eval SDR, PESQ, and STOI values [16] and subjectively compared by listening to them. Figure 3.8 (c) reveals that the algorithms with different hop sizes for the individual noise types did not deviate much from the average overall noises.

Finally, the performance of the SE method could be tested with an unlimited number of values for each parameter and an endless number of parameter combinations. The evaluation in this research is more intended to demonstrate that, in this type of speech enhancement approach, the STFT parameters should be taken into consideration because they do affect the enhancement method's outcomes.

3.7 Evaluation of the Test Signals with m_l and m_s

RPCA-based SE algorithm's SDR performance depends on STFT settings. Another tendency in the data that has not yet been addressed is the dependence of SDR results on different types of noise. Certain noises, like wind and traffic and car, are more easily hidden than others. The findings show how SE performs differently depending on noise kind. This figure is the average of each of the noise's STFT values and input SNR levels.

Test signals were analyzed using the metrics m*l* and m_s, which suggest the feasibility of the low-rank and sparse model for speech and noise. This strategy was chosen since the results could only differ if the SE algorithm's PCP could separate speech and noise accurately in some T-F representations or from some other noises. Since PCP is a low-rank and sparse matrix decomposition approach, it was hypothesized that inconsistencies were caused by the model's unsuitability for speech and noise.

The m_l and m_s values for test signals were evaluated and compared to enhancement test results to find probable relationships. The table shows the noise types' m_l and m_s values in different T-F formats. The rectangular, Hanning, and Blackman were calculated using the respective

window functions with an STFT length of M = 1024, and a hop size of 256. The findings in columns "1536" and "512" were obtained with respective STFT lengths, Hann windowing, and ¼ M hop-size. A hop size of 1/8 M and ½ M with M = 1024 and Hann windowing led to the "one-eighth" and "one-half" columns. In Table 3.5, the values are averaged over all the noise types.

Noise	**Rectangular**		**Hanning**		**Blackman**		**512**		**1536**		**One eighth**		**One Half**	
	m_l	m_s	m_l	ms	m_l	ms	m_l	ms	m_l	ms	m_l	ms	m_l	ms
Crowd	0.61	2.13	0.52	2.78	0.46	2.91	0.63	2.82	0.53	2.65	0.48	2.77	0.56	2.85
Wind	3.26	0.65	3.22	0.34	3.16	0.35	4.67	0.35	2.11	0.34	2.76	0.35	2.36	0.34
Water	0.54	2.81	0.56	3.83	0.55	3.82	0.46	3.85	0.57	3.85	0.42	4.01	0.73	3.98
Machine	0.51	2.64	0.44	3.73	0.46	2.01	0.48	3.65	0.48	3.52	0.35	3.68	0.56	3.72
Traffic & Car	1.86	1.93	2.58	1.87	1.46	1.93	1.65	1.82	0.92	1.76	1.34	1.95	1.13	1.98

Table 3.5: Average m_l *and* m_s values for the test signals

It is interesting to see from m_s values in Table 3.5 that the sparsity assumption for the speech in comparison to the noise signals was fulfilled by all test signals apart from the wind noise. This can be recognized because $m_s > 1$ means that on average the 20 speech signals can better be modeled as sparse than the noise signal. On the contrary, table 3.5 reveals that many of the noise signals do not exhibit a more predominant low-rank character than the speech signals. Table 3.5 showed that for the machine noise for example and with an STFT length of M = 1536, the recovered speech signals had an SDR that was about 2.8–4.8 dB higher than the respective input SNR levels, which means that the SE algorithm was successful. However, Table 3.5 now reveals that the machine noise in the STFT representation with M = 1536 can be distinguished from speech signals by its low rankness because on average, the speech signals are better modeled as low-rank than the noise. Still, the enhancement method based on a low-rank and sparse decomposition performed very well for machine noise. Further, on average the recovery of the speech signals from the machine noise yielded the second-best results despite the low average ml value in Table 3.5.

Similar observations can be made for the wind noise with the difference that for that noise type the m_l values are good while m_s values were too low for the low-rank and sparse model to be fit well.

The conclusions that arise from this are as follows: The sparse model for speech and low-rank model for noise is not the best choice. Despite the fact that speech is often far sparser than the noise signals, it appears that the assumption that noise signals are well represented as low-rank is a little over optimistic. According to the experiments, speech signals tend to be less mainly low-rank than noise signals.

However, the effectiveness of the enhancing technique cannot be denied. Additionally, it suppresses noise types with poor m_l or m_s values. And this is also true in cases where improvement is only made possible by PCP without masking. This indicates that PCP has favorable qualities for the purpose to which it is applied in this study, i.e., beneficial qualities that go beyond its capacity to produce a reliable low-rank and sparse matrix decomposition

The m_l *and* m_s Tables 3.4 and 3.5 further reveal that there does not appear to be any connection between m_l *and* m_s and the typical success of the enhancement method. It might be assumed that the Traffic and car noise, where $m_l > 1$ and m_l *and* $m_s > 1$ hold, and which has the greatest overall bss-eval enhancement findings [209], demonstrate the relationship between good m_l *and* m_s values and speech recovery. However, there is a very evident exception to this trend: Although the average m_l *and* m_s values for wind noise are better than those for traffic and car noise, the average SDR results are substantially poorer than those for car noise. This raises the concern of whether it is possible to forecast how well the enhancement method will work for every signal based on the values of m_l *and* m_s. In summary, some noise patterns allow for improved speech recovery, but they are difficult to distinguish only from m_l and m_s measurements.

Tables 3.4 show that different STFT parametrizations have a significant impact on how sparse or low-rank the speech and noise signals appear in their T-F representations. This confirms the discussion and assumptions of Section 3.4.

There is a trend in the data that can be recognized when looking at Table 3.4 for the SDR results in dependency of STFT window type and Table 3.5 for m_l values in dependence of the window type. It appears that rectangular windowing led to the best SDR results of all three windows for all noise types and all input SNR levels. Fittingly, rectangular windowing led to the highest m_l values of all three windows for all noise types.

The same trend is not true for m_s where rectangular windowing led to the worst values of all three window types except for the wind noise. But because the m_s values tend to be better in comparison with m_l, m_l might be the more critical and more influential figure.

A similar trend was not found for different STFT lengths or different hop- sizes. This could have to do with the fact that for different STFT length and hop-size settings the T-F matrices differ in their sizes while for different windows, the sizes stayed the same. Possibly the size of the T-F matrix influences the performance of PCP step.

3.8 Determination of an Optimal GoDec Parametrization

Studies to determine the parameters for SS-GoDec-based SE algorithm show that, on average, this algorithm gives the best results across 500 test signals. SS GoDec low-rank, sparse, and noisy matrix decomposition requires r and λ. r specifies the rank of the low-rank component, while it trades cardinality for noise energy. Higher λ makes sparse components sparser and vice versa.

The two parameters need to be tuned. If λ is too small, the noise will leak into the speech estimate because the necessity to minimize the cardinality of the sparse component is not strong enough to remove all substantial noise contributions. If λ is too large, the desire to lower the cardinality of the sparse component is so strong that portions of the speech will be lost. All 500 test signals were decomposed with SS GoDec-based SE algorithm for different combinations of r and λ. All 500 speech signals were examined with r ∈ [1, 2, 3, 5, 8] and λ ∈ [0 to 1.5]. This

evaluation involved 5. 16. 500 = 40000 decompositions. The SE technique uses an STFT with Hann windowing, M = 1536, ¼ M hop-size, and log-sigmoid masks. Figure 3.9 shows the average SDR over 500 test signals for every (r, λ) combination.

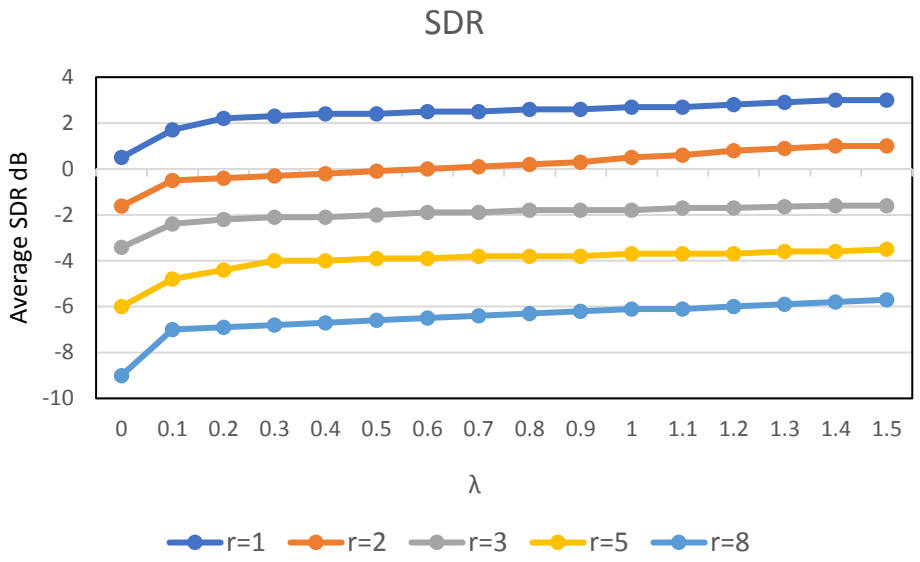


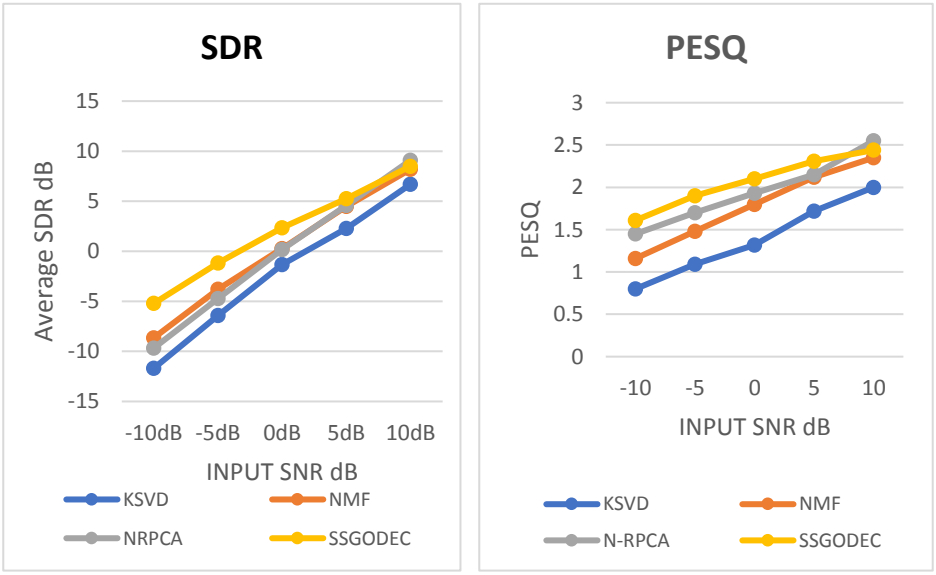
Figure 3.6: Influence of the NNM-RPCA parametrization on the average output SDR.

Figure 3.6 shows that setting r = 1 and λ = 1 yields an average SDR of 2.92 dB. This parameter will be used to compare SS-GoDec with RPCA in SE. SS-GoDec with r = 1 no longer performs low-rank decomposition. r = 1 is the lowest possible rank of any matrix, hence almost none of the noise signals are absorbed by the low-rank component but most parts of the noise signals will be assigned to matrix R in the SS-GoDec decomposition. SSGoDec with r = 1 prefers sparse and noisy component decomposition over low-rank and sparse.

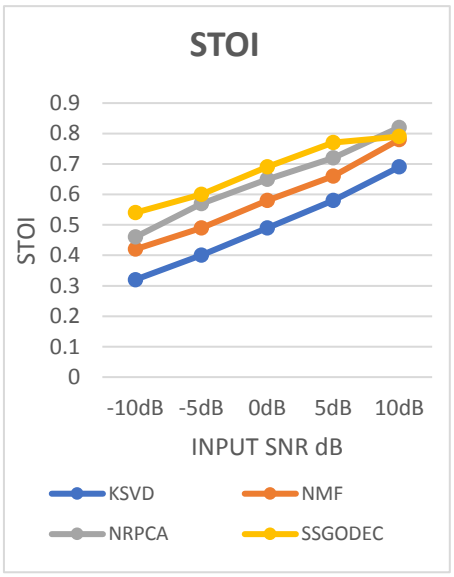
A higher value of *r* leads to a rapid decrease in the SDR level of the recovered speech signal as can be seen in Figure 3.6. This is in accordance with Table 3.4 and the concerns raised in Section 3.5.1 about the validity of using a low-rank and sparse decomposition to improve speech.

3.9 Comparison of KSVD, NMF, RPCA, and GoDec-Based Speech Enhancement

In this section, the SS-Godec SE algorithm is compared to RPCA and other baseline SE methods using bss-eval SDR metrics [209] and the subjective listening impression. Both RPCA and SSGODEC computation times for generating an improved speech signal were calculated. All approaches employed an STFT with Hann windowing, M = 1536, and 384 hops. No T-F masks were used. For the SS-Godec algorithm, log-sigmoid masks were utilized. Section 3.5.2's evaluation recommended setting r and λ to 1. All 500 noisy input signals were decomposed using both approaches, and Figure 3.7(a-c) depicts the average results for 20 speakers at five SNR levels.



(a) (b)



(c)

Figure 3.7(a-c): Comparison of the performances of RPCA and Semi-Soft GoDec with existing baseline SE methods in terms of a) SDR b) PESQ c) STOI

SSGODEC-based SE yields better SDR, PESQ, and STOI values than baseline techniques at lower input SNR levels, while RPCA is superior at higher input SNR levels. GoDec produced a 4.5–6.45dB improvement under noisy and very noisy conditions. Listening to the test signals gave the subjective impression that GoDec suppressed noise better than RPCA at 0dB, 5dB, and 10dB input SNR levels for all noise classes. For low input SNR levels, crowd, water, and machine noise, it was unclear which method suppressed noise best in enhanced speech signals. For the other two noise categories, GoDec enhancement had a considerable benefit at low input SNR.

The SS-GoDec SE approach performed well for traffic and wind noise. At higher input SNR levels, just traces of noise remained in recovered speech signals, and speech was minimally affected. At input SNR levels of -10dB and -5dB, noise bled into voice signals, and speech was lost, but the results sounded convincing and marvelous. With RPCA Enhancement, the noise remained in all speech channels. The SE approach employing SS-GoDec performed better than RPCA-based algorithm. Faster GoDec. GoDec computed one enhanced speech signal in 1–1.5 seconds, while RPCA needed 15–20 seconds. A test signal T-F matrix with the above STFT values is 769 x 161. RPCA needs a minimum of (769^2. 161, 769. 161^2) = 19933249 flops for one SVD but SS GoDec just needs 12 (769 + 3.161 + 4.1) + 8 .769 161. 1 = 991728 flops for one low-rank approximation. The SE technique using GoDec matrix decomposition at its foundation generated better and faster-improved speech signals.

3.9.1 Observations

Sparse, low-rank RPCA was presented to estimate the noise spectrum from a noisy speech spectrogram. With the right settings, the proposed SE technique suppresses noise in noisy speech signals, even in unsteady noise. Experiments on the Noizeus database showed sparse and low-rank minimization's effectiveness under low SNR. Parameterizing the STFT in the SE algorithm affects its success and should be carefully considered. For the test corpus signals, M = 1536 with 384 hops, and Hann windowing worked well. An appropriate parametrization for the test signal corpus was found, and a comparison between conventional baseline approaches and RPCA-based SE algorithms demonstrated that SS-GODEC outperformed in terms of enhanced speech signal quality. The proposed approaches performed better on PESQ scores without a prior noise or speech dictionary than state-of-the-art algorithms. No T-F masking led to the best enhancement outcomes because binary and log-sigmoid masks degraded SDR performance and had unpleasant sonic side effects. Noise types affect how successfully the denoising algorithm separates speech. This could not be explained by the newly added m_l and m_s measurements of low-rank and sparse model appropriateness for noise and speech signals. These measurements

demonstrated that the low-rank and sparse model for noise and speech is not suitable since in most cases, speech signals had a more prevalent low-rank character than noise signals.

The explanation in this part did not give any precise directions like" utilize an STFT length of M = 1024, a hop-size of 512, and a Hann window to achieve the best results". It's hard to predict how RPCA and the masking step will react to spectrograms with different parameter combinations. Even more so, since algorithm noise signals can vary. For one noise and speech signal, one STFT value may be best, while another may be better for another. Predicting the optimum parameter selections is particularly tough because it's not enough to find a T-F representation with sparse speech or low-rank noise. A T-F representation is needed in which speech is sparse and noise isn't, and noise is low-rank and speech isn't. A T-F representation that separates the two components based on these properties is desired. Future research could examine the impact of STFT parameters on weighted Nuclear norm minimization. Then, even more, extensive test signal corpora and more elaborate objective speech quality measures might be applied. It should also be determined which noise types the SE approach performs well for and which are difficult. Good RPCA-based findings for particular noise types could encourage real-time implementation of this technique, which could be useful for hands-free mobile communication in cars or hearing aids.

3.10 Summary

Existing SE methods are confined to stationary noise and are ineffective for non-stationary noise experienced in the physical realm. Compressive sensing (CS) claims to successfully recover a sparse signal from a few random samples. Structured speech is restored using CS recovery techniques, which also remove unstructured background noise. This study, which was motivated by the newly developed method, examines how well CS performs in SE utilizing both fixed and adaptive dictionaries. The effectiveness of various sparse domain, sensing, and combined transform domain (dictionary) combinations for improving speech is investigated.

These non-adaptive dictionaries are unable to efficiently (sparsely) represent a particular class of signals while being simple and having fast computations. An in-depth study on dictionary learning as a solution has been conducted during the past decade. This method involves learning a dictionary from a particular class of useful training signals. It has been empirically proven that these adaptive dictionaries outperform non-adaptive ones in several signal processing applications, including image compression and enhancement and classification tasks. An adaptive dictionary requires prior knowledge of speech and background noise for supervised learning. However, while developing dictionaries, training and testing domains may vary owing to different viewpoints and situations. In environments with considerable noise and unidentified non-stationarity, the adaptive dictionary strategies are ineffective. Therefore, to get better performance for noisy speech enhancement in real-world applications, unsupervised approaches are required. For data that have been corrupted by noise, the novel data analysis technique known as Robust Principal Component Analysis (RPCA) is effective. Since RPCA doesn't require any prior knowledge, it may be used in unknown real-world noisy situations. It can also work effectively in very noisy surroundings. As a result, unsupervised separation for robust speech enhancement frequently uses a low-rank and sparse matrix decomposition model. The experimental tests that were conducted to determine if low-rank sparse representations were appropriate for SE under various NOIZEUS corpus conditions using both objective and subjective assessments are presented in this chapter.

Chapter 4

Speech Enhancement using Low-Rank Sparse decomposition techniques under Low SNR Environments

In this chapter, a novel speech enhancement Approach based on Low-Rank sparse decomposition techniques under Low SNR (< 0dB) Environments is proposed to reduce speech distortion and residual noise under low SNR conditions.

4.1 Motivation

Noise estimation is a crucial stage in speech enhancement (SE), and it commonly necessitates the use of prior models for speech, noise, or both. Prior models, on the other hand, can be ineffective in dealing with unseen nonstationary noise, especially at low signal-to-noise (SNR) levels as they produce residual noise, i.e., the noise remaining after the enhancement process. In addition to the requirement for minimal distortion of original speech, which was discussed in Chapter 3, it is important that the residual noise does not sound unnatural. Although the Adaptive dictionary-based methods such as KSVD and NMF methods can eliminate interferers, under low SNR conditions, part of the recovered speech formant structure is lost during the matrix decomposition process, resulting in speech distortion. Therefore, there is a great need to minimize noise and produce a speech output, reducing listener fatigue and improving intelligibility. Low-rank sparse decomposition techniques prevent musical noise by estimating precisely the matrix rank using the Nuclear Norm Minimization (NNM) approach.

4.2 Introduction

Musical noise results due to the existence of randomly spaced peaks in the spectrum of the reconstructed signal because of the overestimation and underestimation of clean signal in adjacent spectral groupings, resulting sometimes from rough estimation of the noisy signal power spectrum. Low-rank and sparse matrix decomposition (LRSD) like Robust Principle Component Analysis (RPCA) method is used to estimate noise and speech when neither is available beforehand by decomposing the input noisy spectrum into a low-rank noise component and a sparse speech component. Due to the approximation of the actual rank of noise, these techniques are constrained, and they do not directly exploit the low-rank property in optimization. Nuclear norm minimization (NNM) is the most well-known approach, as it can precisely recover the matrix's rank under certain restricted and theoretically guaranteed conditions. NNM, on the other hand, is unable to reliably estimate the matrix rank in many situations. The source of musical noise in LRSD method is the inaccurate estimation of the rank which produces fluctuating tone-like components. Musical noise and distortion can be reduced by making use of low-rank and sparse decomposition models with a different objective function than conventional RPCA approaches. For each noisy input, all the regularization parameters are automatically modified and updated. Consequently, to alleviate speech distortion, it is intended to build a novel low-rank and sparse matrix decomposition model by placing appropriate constraints on the sparse part.

4.3 Low-rank and Sparse Matrix Decomposition

The principal Component Pursuit (PCP) model decomposes noisy speech spectrum into low-rank and sparse matrices from RPCA. Efficient estimation methods may approximate and extract sparse and low-rank components with high probability. The low-rank matrix approximation (LRMA) approach minimizes the rank of its relaxations using corrupted speech observations to recover the low-rank matrix. Rank minimization is NP-hard with no efficient

solution. The nuclear norm, which contributes to NNM-based techniques [212], is optimal for replacing the rank function with its tightest convex relaxation.

The traditional Low-rank matrix Factorization (LRMF) approach, commonly known as SVD, uses a truncation operator on its singular value matrix to achieve the best rank-r approximation of input data matrix M in terms of F-norm fidelity loss. The robust principal component analysis (RPCA) approach, based on nuclear norm minimization (NNM), suppresses outliers in data. Singular value thresholding [148] and the alternate direction method of multipliers (ADMM) [149] framework can solve NNM. In the time-frequency (T–F) domain, noise signals in different time-frames exhibit similar spectral structures and patterns that are frequently associated and may be represented using a few basis vectors. Therefore, the noise spectrogram is supposed to be in a low-rank subspace. Speech signals are sparse in the T–F domain since spectral energy centralizes in a few T-F units [150]. The non-parametric RPCA approach requires no assumptions regarding speech or noise spectral component distributions. VAD is unnecessary since speech and noise spectra may be reconstructed concurrently. This strategy outperforms noise estimation-based SE methods [151, 152]. RPCA is quick and has minimal tuning parameters. It also works well in high noise. Mask estimation on spectrogram using sparse and low-rank decomposition helps denoise voice. Similar improvements have been attempted to enhance low-rank and sparse models like Semi-Soft Go-Decomposition (SS-GoDec) [153] technique for the SE.

Nuclear norm minimization (NNM) can retrieve the matrix rank under certain theoretically guaranteed conditions. When the signal source is unknown, Nuclear Norm Minimization (NNM)-based RPCA and SS-GODEC approaches may provide undesirable results. The conventional NNM attempts to balance singular values equally, making convex norm computation easy. This limits its capacity to tackle a wide range of practical difficulties where individual values have physical significance and should be treated properly. Due to NNM's estimate of noise rank, these methods cannot directly exploit the low-rank characteristic in optimization. Since NNM over-shrinks rank components, it cannot correctly approximate matrix rank in many real-world applications. Weighted nuclear norm minimization (WNNM) achieves a superior matrix rank approximation than NNM, which heuristically sets the weight

as inverse to the singular values. In computer vision applications, the recently suggested WNNM can replace the nuclear norm as an improved rank approximation [154]. We propose a weighted low-rank and sparsity SE method for speech and noise spectrogram separation since RPCA and SS-GODEC explicitly account for deviations of the speech and noise time-frequency

matrices from the idealistic sparse and low-rank model. The low rank of WNNM improves singular value decomposition, ADMM, and accelerated proximal gradient line search methods. Thus, the WNNM-based RPCA enhancement model makes use of speech signal correlation and outperforms NNM-based techniques. The weighted Schatten p-norm minimization (WSNM) RPCA model was developed to accomplish low-rank regularisation. WSNM suppresses noise better than state-of-the-art methods and models dynamic and complicated situations better [155]. WSNM, a generalized WNNM, was tested for image denoising.

4.4 SPEECH ENHANCEMENT USING RPCA-BASED WEIGHTED NUCLEAR NORM MINIMIZATION (WNNM)

The goal of NNM decomposition is to recover the underlying low-rank matrix L from its degraded observation matrix M, by minimizing $\|L\|*$. But the main problem with the above formulation of NNM-RPCA is that the optimization function is non-convex and the problem falls under NP-hard problems, which are computationally expensive. Moreover, the technique assigns equal weights to all the singular values or rank components resulting in a biased estimate of low-rank and sparse components, restricting its flexibility in practical applications. The singular values of a matrix in the context of speech processing are closely associated with the physical properties of the speech signal. Large singular values account for prominent features of speech such as short-term zero crossing and energy, while smaller ones correspond to noise components. Therefore, large singular values must be treated differently from the smaller ones and must be preserved to reproduce high-quality speech. To improve the performance of NNM, in the last few years, numerous applications based on NNM have been proposed, such as video enhancement, background extraction, and subspace clustering. However, the nuclear norm is generally adopted as a convex surrogate for matrix rank. The singular value thresholding (SVT) model for NNM treats different rank components equally,

leading to over shrinking the rank components, and hence the estimation of the matrix rank becomes inaccurate. As a result, it is obvious that the traditional NNM model, as well as the accompanying SVT approaches, are insufficient to deal with such problems. The methods such as truncated nuclear norm regularization (TNNR) and partial sum minimization (PSM] among N singular values, keep the largest 'r' (rank of the matrix) singular values unchanged and only minimize the smallest (N-r) ones. TNNR and PSM, on the other hand, are not flexible enough because they make a binary decision on whether or not to regularize a particular singular value or not. This could produce an over-fitting solution due to the noise effects.

Inspired by the singular values that have distinct physical implications, the weighted nuclear norm minimization (WNNM) model has been proposed. WNNM generalizes NNM and improves the flexibility of NNM significantly. To improve the flexibility of the nuclear norm, we propose to investigate the weighted nuclear norm and evaluate its minimization strategy. The weighted nuclear norm of a matrix M is defined in Eq(4.6) as follows:

$$|| M ||_{w,*} = (\sum || w_i \, \sigma_i(\mathrm{M}) ||_1 \qquad (4.1)$$

Where vector $w = [w_1, w_2 ..., w_n]$ and $w_i \geq 0$ is a non-negative weight assigned to σ_i. The rational weights rules for weighting can be specified depending on prior knowledge and understanding of the problem, which will greatly improve the representation capability of the original data from the corrupted input. From prior knowledge, it is understood that higher singular values of M are more essential than the smaller ones in natural speech because they indicate the energy of the major components of M. The larger the individual values are, the less they should be shrunk while denoising. As a result, it's a natural assumption that the weight given to $\sigma_i(M)$, i^{th} singular value of M, should be inversely proportional to σ_i (M). WNNM is a non-convex problem that is more complex to solve than NNM. So far, the WNNM problem has got very little attention in this work. We investigate in-depth the WNNM problem using F-norm data fidelity. The solutions are examined under various weight conditions.

SE aims to estimate the hidden clean speech from its noisy observation. As a classical and fundamental problem in low SNR conditions, SE has been extensively explored for many years; however, it remains a prominent research area since enhancement is an ideal testbed for investigating and evaluating statistical speech modeling techniques. The use of speech Nonlocal self-similarity (NSS) has improved significantly the SE performance in recent years. The NSS prior refers to the fact that for a given local frame in a natural speech, one can find many similar frames to it across speech signal. The nonlocal similar frame vector is stacked into a matrix, which must be a low-rank matrix with sparse singular values. As a result, enhancement algorithms can be designed using low-rank matrix approximation method.

4.4.1 Model formulation for WNNM Model

RPCA attempts to identify a low-rank version and a sparse version from a single matrix and has a wide range of applications. In this section, we propose reformulating eq (4.1) using the weighted nuclear norm, resulting in WNNM-based RPCA (WNNM-RPCA) model represented in Eq(4.2) as follows:

$$\arg\min \left(\left|\left|L\right|\right|_{w,*} + \lambda \left|\left|S\right|\right|_1 \right)$$
$$under\ the\ constraint\ M = L + S \qquad \ldots\ldots\ldots \qquad (4.2)$$

ADMM is used to solve the WNNM-RPCA problem, just like it is in NNM-RPCA. By using ALM method, a Lagrange multiplier Y is associated to produce an unconstrained function represented in Eq(4.3) as follows:

$$arg\ min_{L,S} ||L||_{w,*} + \lambda\, ||S||_1 + < Y, M - L - S > + \frac{\mu}{2} ||M - L - S||_F^2 \qquad \ldots\ldots\ldots (4.3)$$

Where μ=1/2k

The optimum values of L and S are found in an iteration using the Y value from the last iteration. Then again, the value of Y is updated in the current iteration with the new optimum L and S values.

L_k, Y_k, and S_k are local variables and represent the local optimum in the kth iteration represented in Eq(4.4) and Eq(4.5) as follows:

$$S_{k+1} = argmin\, f(S, L_k, Y_k)$$
$$= argmin_S\, \lambda|| S ||_1 + \frac{\mu}{2}|| M + (Y_k/\mu) - L_k - S ||_F^2 \quad \ldots\ldots (4.4)$$

Similarly

$$L_{k+1} = argmin_L\, \lambda|| L ||_{w,*} + \frac{\mu}{2}|| M + (Y_k/\mu) - L - S_{k+1} ||_F^2 \quad \ldots\ldots\ldots (4.5)$$

For the weight w_i of each group M_i, large singular values of each frame group mj in M usually offer significant information, and vice versa, inspired by singular values that have clear physical implications. As a result, we usually shrink large singular values less and smaller singular values more. In other words, the weight w_i of each group m_j in M is set to be inverse to the singular values, and so as in [154], the weight is heuristically set as:

$$w_{i,j} = c / (\sigma_{i,j} + \epsilon), \text{ where c and } \epsilon \text{ are small constants. } \ldots\ldots(4.6)$$

Solving the above equation, we obtain Eq(4.7) as follows:

$$Y_{k+1} = Y_k + \mu_k(M - L_{k+1} - S_{k+1}) \quad \ldots\ldots (4.7)$$

Thus, in this way, the values of L, S, and Y are updated to reach the global optimum. The algorithm1 and the flow chart (shown in figure 4.1) for WNNM-RPCA is shown below

Algorithm 1 SE by WNNM-RPCA

Input: Noisy speech data M, weight vector w

1: Initialize $\mu_0 > 0$, $\lambda > 0$, $\rho > 1$, $\theta > 0$, k = 0, $L_0 = M$, $Y_0 = 0$;

2: do

3:$S_{k+1} = \ argmin_S \lambda|| S ||_1 + \frac{\mu}{2}|| M + \ (Y_k/\mu) - \ L_k - S ||_F^2$

4. for each frame mj in M do

5: Find a similar frame group Mj

6: Estimate weight vector w

7: $L_{k+1} = argmin_L \lambda|| L ||_{w,*} + \frac{\mu}{2}||M + \left(\frac{Y_k}{\mu}\right) - L_k - S_{k+1} ||_F^2$

8: $Y_{k+1} = Y_k + \mu_k(M - \ L_{k+1} - S_{k+1})$;

9: Update $\mu_{k+1} = \rho * \mu_k$;

10: k = k + 1;

11: while $||M - L_{k+1} - S_{k+1}||_F / ||M||_F > \theta$

12: Output Matrix L = L_{k+1} and S = S_{k+1};

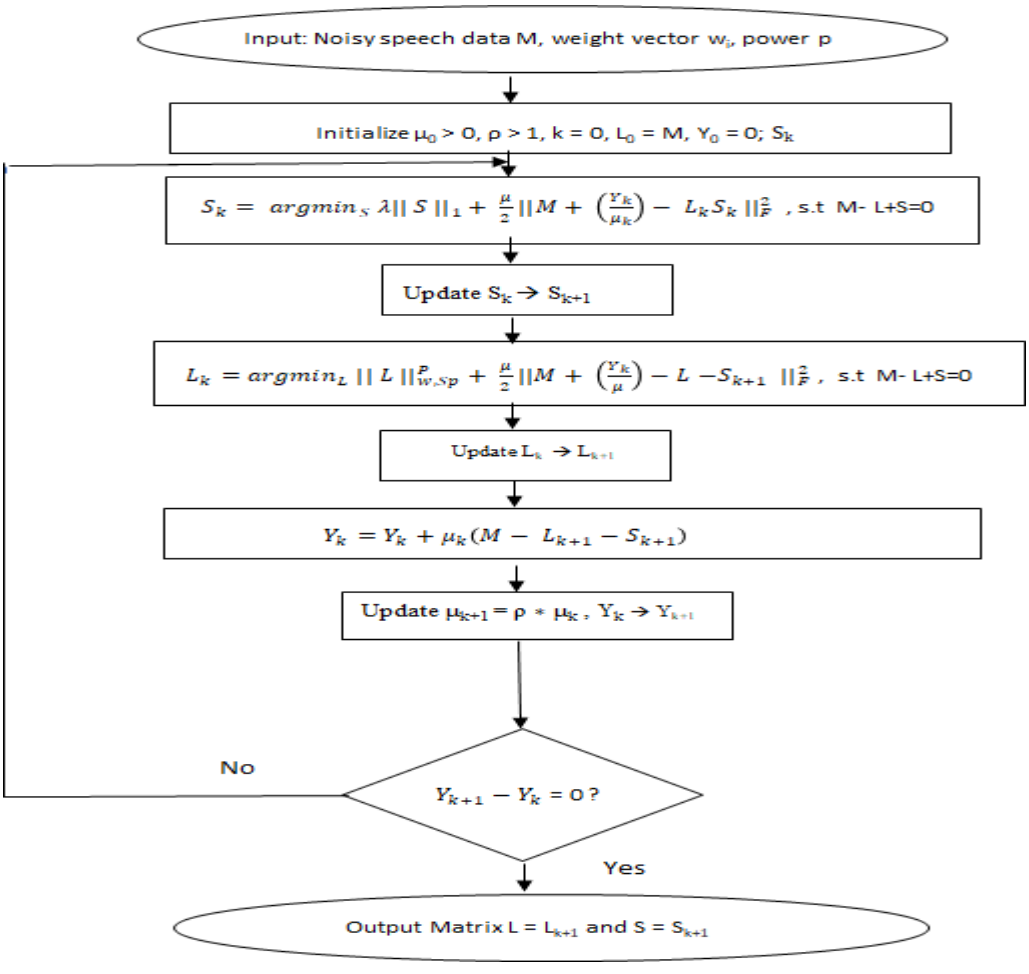


Figure 4.1 Flow chart representation of the SE using WNNM-RPCA

4.4.2 Model formulation for WSNM Model

WSNM is a generalized variant of Weighted Nuclear Norm Minimization, whose image-denoising performance has been studied in [155],[213]. WSNM Low-rank approximation tends to carry out low-rank regularization effectively wherein we employ the loss function expressed in Eq (4.8) as follows:

$$arg\ min_{S,L} \|\ S\ \|_1 + \|\ L\ \|^P_{w,S_p}\ s.t\ \ M = L + S \tag{4.8}$$

Using the Augmented Lagrangian function in Eq (4.8) we get :

$$L(L,S,Z,\mu)\ =\ \|\ S\ \|_1 + \|\ L\ \|^P_{w,S_p} + <\ Y, M - L - S\ > + \frac{\mu}{2}\ ||M\ -L\ -S||^F_2 \tag{4.9}$$

Where Y is a Lagrangian multiplier, μ is a positive scalar. The values of the weighted vectors are defined in Eq (4.9) as follows:

$$w_i = C\sqrt{(mxn)}/(\sigma_i(M) + \epsilon) \quad \ldots. \tag{4.10}$$

Thus, in this way, the values of L, S, and Y are updated to reach the global optimum. The algorithm2 for WSNM-RPCA is shown below

Algorithm 2 SE by WSNM-RPCA
Input: Noisy speech data M, weight vector w, power p
1: Initialize $\mu_0 > 0$, $\rho > 1$, $k = 0$, $L_0 = M$, $Y_0 = 0$;
2: do
3: $S_{k+1} = \ argmin_S\ \lambda\|\ S\ \|_1 + \frac{\mu}{2}\|M + \left(\frac{Y_k}{\mu_k}\right) - \ L_k S_k\ \|^2_F$
4: $L_{k+1} = argmin_L\ \|\ L\ \|^P_{w,Sp} + \frac{\mu}{2}\|M + \left(\frac{Y_k}{\mu}\right) - L - S_{k+1}\ \|^2_F$;
5: $Y_{k+1} = Y_k + \mu_k(M - \ L_{k+1} - S_{k+1})$;
6: Update $\mu_{k+1} = \rho * \mu_k$;
7: k = k + 1;
8: while $\|M - L_{k+1} - S_{k+1}\|_F$ / $\|M\|_F$ not converged
9: Output Matrix L = L_{k+1} and S = S_{k+1};

4.5 Simulations and Methodology

This section provides details of the experimental setup and methods for evaluating the suggested noise reduction methods' performance and suitability. The results of these experiments are of high value for assessing the possibilities and limitations of the enhancement method. They also allow for estimating the influence of the parameters that were considered in Sections 4.4 and 4.5. Because of many uncertainties and complicated relations, the theoretical discussion in the previous section did not suffice to make a reliable prediction of the performance of the speech recovery procedure. On the contrary, the results of the experiments show how well the algorithms have already performed in tests, and the assumption is justified that they will perform similarly in identical situations. Therefore, the following section contains very valuable information about the potential of the SE method in practical use. Even more so, as the number of test signals that were used is rather high.

The standard Noizeus corpus [145] was used in studies. The speech signals are available as Wav files with a sampling rate of 8 kHz. A total of 20 clean sentences were chosen for this study. The noisy stimuli were created by adding clean phrases with five different signal-to-noise ratio levels, including -10, -5, 0, 5, and 10 dB. The noise signals obtained from a noise collection were available as waveforms with a sampling rate of 8 kHz as well. Five noise recordings: The cheering of a crowd of people, a bubbling stream of water, wind, machine, and car driving in traffic, were selected for evaluation. AWGN was simulated and added to the clean speech. This resulted in an overall number of $5 \cdot 20 \cdot 6 = 600$ mixed test signals, which are all about 3 seconds long.

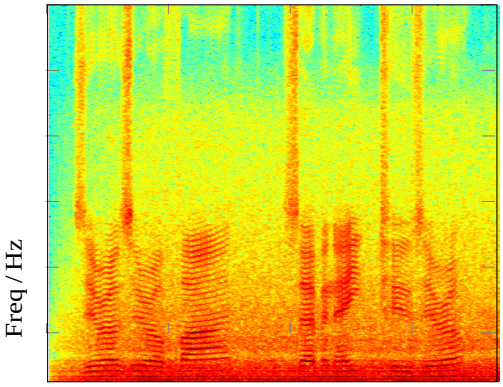
Low-rank, sparse, and noise matrix decomposition algorithms need to be given the parameters r and λ. r determines the rank of the low-rank component while λ value is used to trade off the desire to minimize the cardinality of the sparse component against the desire to minimize the energy of the noise component. It is important to tune the two parameters. If λ value is chosen too small, then parts of the noise will leak into the speech estimate because the urge to minimize the cardinality of the sparse component is not high enough to eliminate all relevant noise contributions from the sparse component. If λ value is too big on the other hand, the urge to minimize the cardinality of the sparse component is so dominant that parts of the speech will be eliminated from speech estimate which is counterproductive. It should be pointed

out that apart from the parameters that were changed in order to evaluate their influence on the performance of the SE method, all settings were left as they were. The best value in this investigation is an average output SDR of 2.89 dB which was achieved by setting r = 1 and $\lambda = 1$. Therefore, this will be the setting that will be used in the following comparison of the performances in speech-denoising methods.

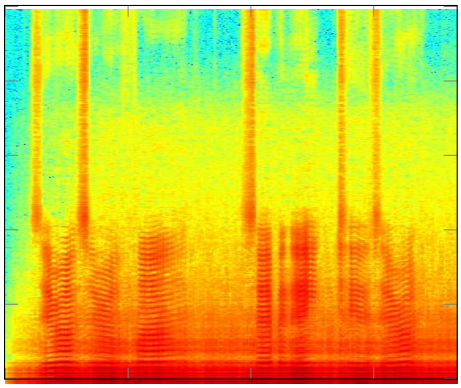
4.5.1 Influence of Binary and Log-Sigmoid Time-Frequency Masking

To illustrate the time-frequency masking step and the influence of different masks, Figure 4.1 contains plots of all matrices that are relevant for example masking step. The spectrogram of the noisy speech input signal is shown in Figure 4.2(a) Figures 4.2(b) and 4.2(c) show the low-rank and sparse components that decomposed the input spectrogram by WNNM-RPCA algorithm. Figure 4.2(d) depicts the binary mask derived from the low-rank and sparse components, while Figure 4.2(e) displays the final speech estimate after applying the binary mask. Figure 4.2(f) shows the log-sigmoid mask calculated from the low-rank and sparse components and Figure 4.2(g) is the final voice estimate after applying the log-sigmoid mask. Figure 4.2(h) shows the final speech estimate spectrogram when no mask is applied. The SE algorithm with WNNM-RPCA matrix decomposition was applied to all 600 test signals. With fixed STFT settings of M = 1024, and hop-size = 256 using Hanning windowing, the enhancement methods were tested with binary masks, log-sigmoid masks, and without masks to experimentally evaluate and compare the influences of masks.

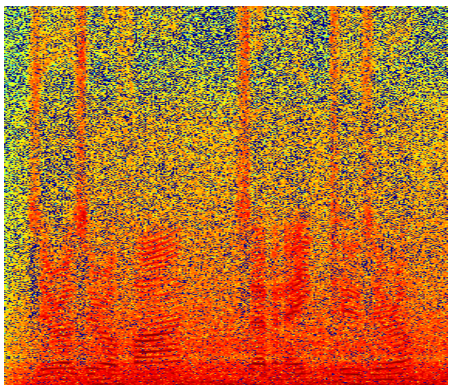
For all five noise types, all five speech-to-noise energy ratios in the input signal, and all three different masks (no mask, binary mask, and log-sigmoid mask), the mean speech-to-distortion energy ratio was obtained by averaging the resulting SDR values of all 20 speakers. The most obvious and most important thing that can be learned from these results is that for all noises and all masks, the approach with the settings specified above achieves a speech-to-distortion energy ratio (SDR) in the speech estimate that is considerably higher than the speech to noise energy ratio in the input signal.


Freq / Hz
Time / sec

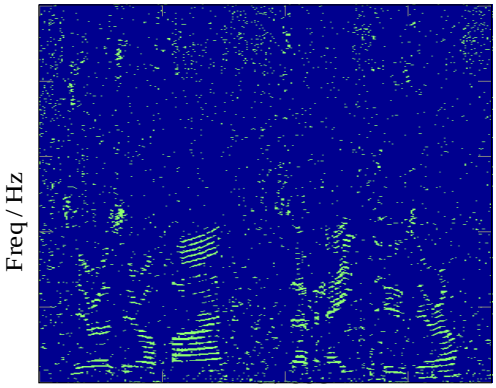
(a)


Time / sec

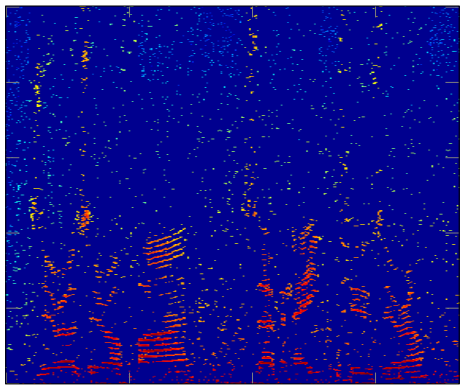
(b)


Freq / Hz
Time / sec

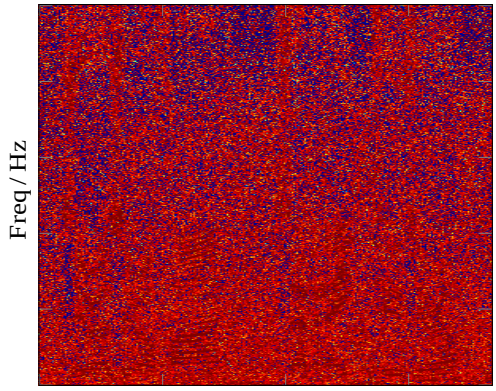
(c)


Freq / Hz
Time / sec

d)


Freq / Hz
Time / sec

(e)


Freq / Hz
Time / sec

(f)

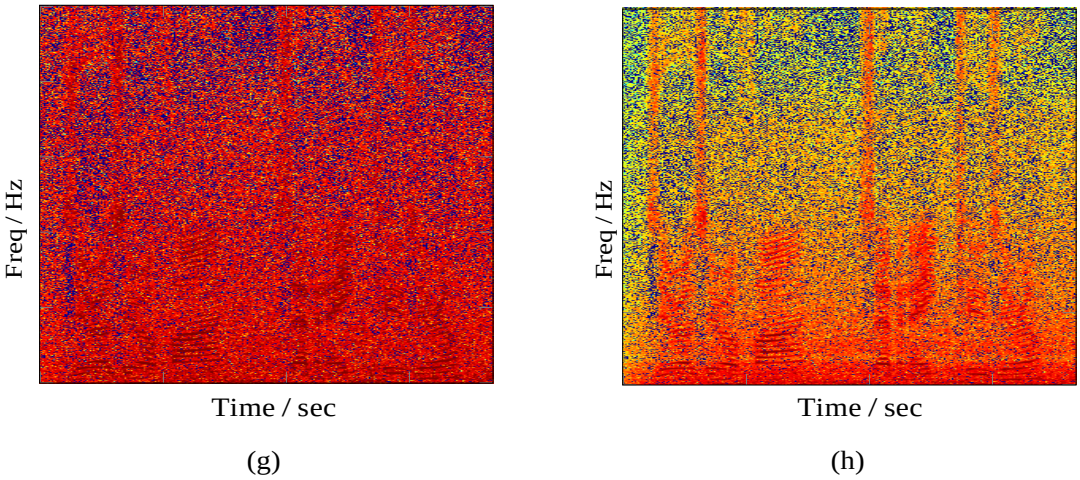


Figure 4.2. Plots of relevant matrices for the time-frequency masking step using WNNM-RPCA Based SE algorithm: a) Spectrogram of noisy speech signal b) Low-rank component. c) Sparse component d) Binary mask e) Speech estimate after binary masking f) Log-sigmoid mask g) Speech estimate after Log-sigmoid mask h) Speech estimate without masking.

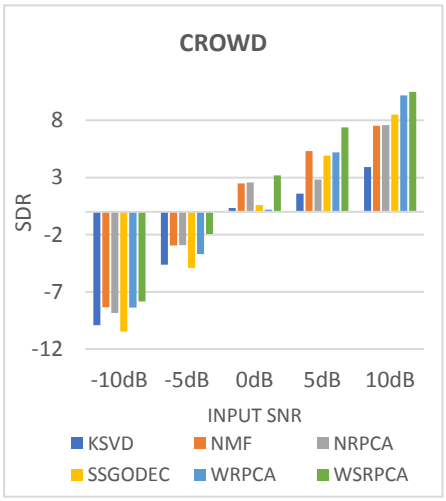
The most obvious and most important thing that can be learned from these results is that for all noises and all masks, the approach with the settings specified above achieves a speech-to-distortion energy ratio (SDR) in the speech estimate that is considerably higher than the speech to noise energy ratio in the input signal. This is true for all entries with input SNR levels of -5dB, 0dB, and 5dB, and most of the entries with input SNR of -10dB. Only for some entries at the already high input SNR of 10dB, did the enhancement method fail to produce a further increase in the speech quality and decreased it instead. From the results, it is observed that all three masks achieve very similar results for low values of the input SNR and that the output SDR values become slightly more spread out for higher values of the input SNR. For high SNR values, the SE algorithm without any mask performs best, log-sigmoid masking is second best and binary masking is last with a performance that is about 1dB worse than that without masking. This suggests that the WNNM-RPCA decomposition already achieves a good separation of speech and noise, which cannot be further enhanced with the masks used here. Instead, the masks cause undesired alterations that deteriorate the results.

For low values of SNR on the other hand, the results are closer together with log-sigmoid masking performing best. So, in very noisy conditions, masking can help improve the outcome of speech recovery a little bit. Another aspect that results reveal is that log-sigmoid masking does constantly perform about 0.4dB better than binary masking. This is not only true on average but can also be verified by comparing corresponding individual entries. It can be

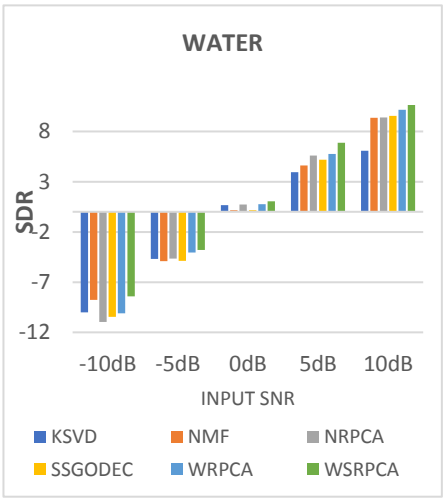
realized that the behavior of the three different mask types for the individual noise types does not deviate significantly from the average of overall noise signals. This means that none of the tested noise signals has a mask type that is particularly suitable and performs significantly better than all other mask types.

4.6 Evaluation of the Weighted Low Rank and Sparse decomposition models for SE

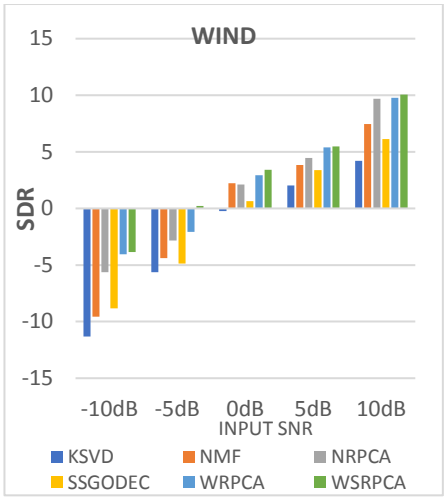
This section contains the evaluation results that were achieved with WNNM-RPCA (WRPCA) and WSNM-RPCA (WSRPCA) based enhancement procedure for different settings of the parameters which were discussed in Sections 4.3 to 4.6. The 600 test signals were decomposed with the SE algorithm that has NNM-RPCA at its core. The results indicate how well the WSRPCA-based SE algorithm will perform for different noise types. The suggested SE algorithms were evaluated and validated against the baseline state-of-the-art SE algorithms using objective evaluation metrics such as SDR, PESQ, STOI, SIG, and BAK. The results of experiments revealed that WSRPCA outperforms state-of-the-art enhancement algorithms not only in terms of PESQ and STOI index but also in local structure preservation, leading to listening being more pleasant.



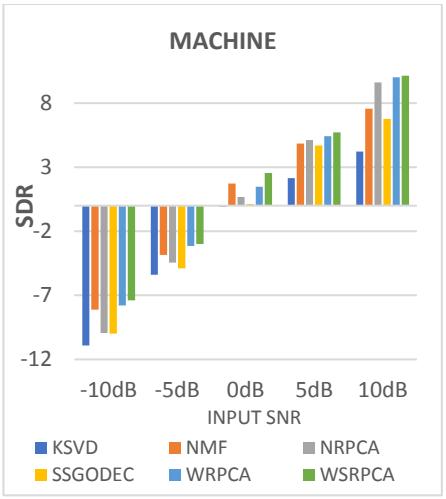
(a)



(b)

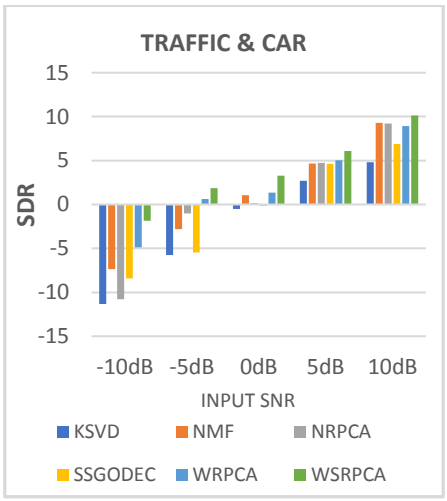


(c)

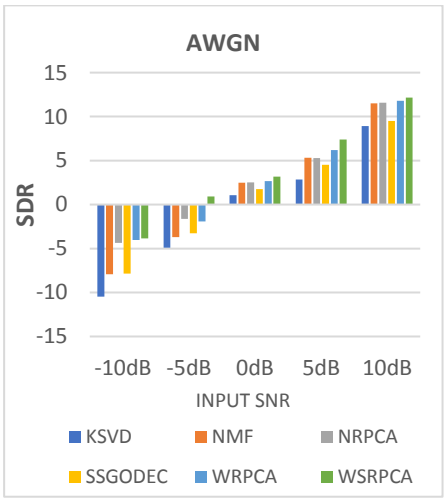


(d)

Figure 4.3(a-d) Performance comparison of the proposed SE algorithms with baseline methods in terms of SDR values using the standard NOIZEUS database for a) Crowd b) Water c) Wind d) Machine



(e)

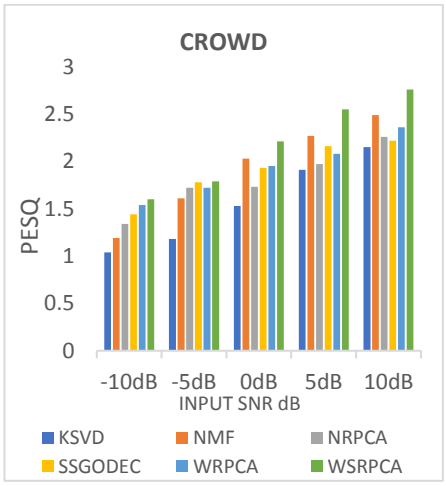


(f)

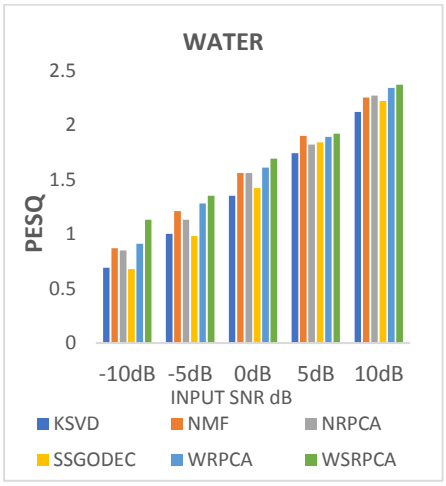
Figure 4.3(e, f) Performance comparison of the proposed SE algorithms with baseline methods in terms of SDR values using the standard NOIZEUS database e) Traffic & Car f) AWGN noise.

At -10dB, using the suggested WSRPCA approach, an improvement of 8.14 dB and 6.17 dB in SDR was observed with Traffic & Car and Wind noise, respectively. The weighted low-rank and sparse models have shown improvements in all SNR levels and noise environments. The proposed methods are also examined with AWGN as a stationary noise case.

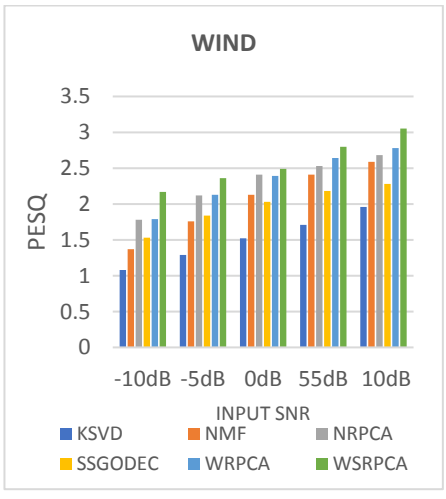
The performance study of the proposed algorithms versus KSVD, NMF, RPCA, and SS-GODEC in terms of PESQ [214] for all SNR levels is depicted in Figure 4.4(a-f). PESQ was improved the most in a noisy unprocessed speech at -10 dB traffic and car noise (ΔPESQ = 0.49) and the least with 10 dB AWGN (ΔPESQ = 0.27). When compared to the baseline techniques, the suggested speech enhancement algorithms showed a considerable improvement in PESQ at all SNR levels and noise situations. At -10 dB noise levels, the greatest PESQ scores were obtained in traffic and car noise, wind noise, and AWGN, with PESQ = 2.51, 2.27, and 2.43, respectively.



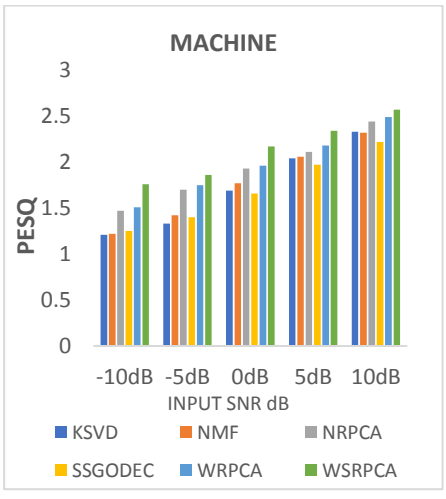
(a)



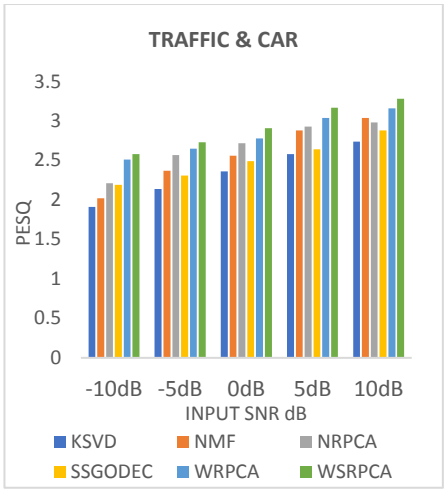
(b)



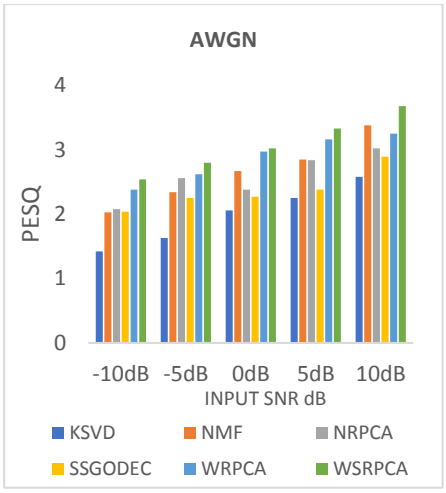
(c)



(d)



(e)

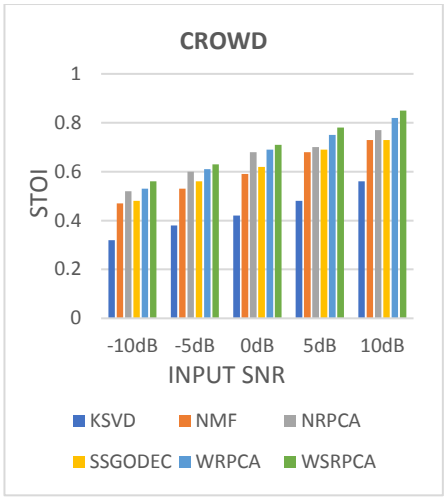


(f)

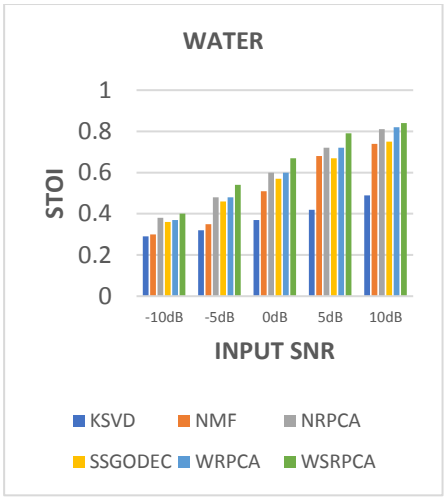
Figure 4.4(a-f). Performance comparison of the proposed SE algorithms with existing methods in terms of PESQ values using the standard NOIZEUS database.

According to the results of the previous investigations on proposed algorithms, using a binary T-F mask improved speech intelligibility in strong noisy conditions (-10 to 0 dB). Figure 4.5(a-f) demonstrates the improved speech intelligibility with a binary mask using STOI [215] measure. For SNR = 10 dB, all noise sources resulted in the highest intelligibility scores (STOI > 0.86).

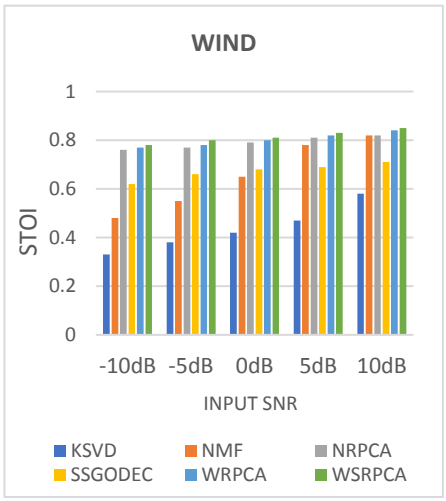
To determine the speech distortion and background residual noise introduced by the recommended algorithms, measures like SIG, BAK, and OVL must be taken into account. For the proposed algorithm's speech distortion (SIG), residual noise (BAK) and Overall quality were measured and are shown in Figures 4.6 (a-c). The strategies offered provide low residual noise and consistently produce high BAK and OVL values at all SNR levels and noise conditions. In all noise situations, the algorithm performed effectively in low SNR levels (-10 dB) and significantly decreased residual noise when compared to baseline techniques. The proposed approach produces high SIG values at all SNR levels and noise situations, demonstrating its usefulness in terms of speech content preservation.



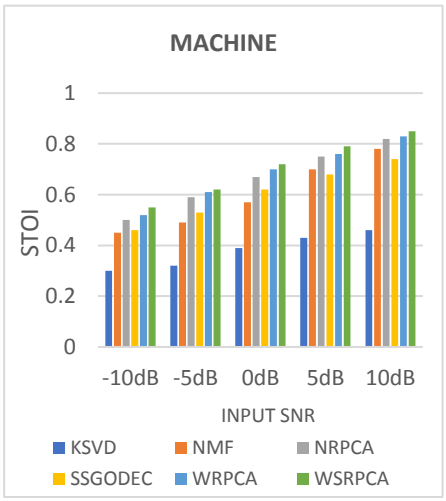
(a)



(b)

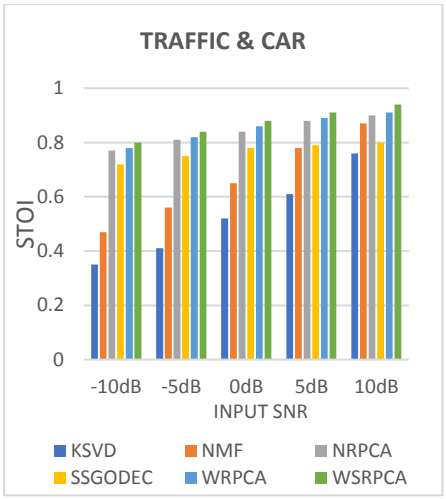


(c)

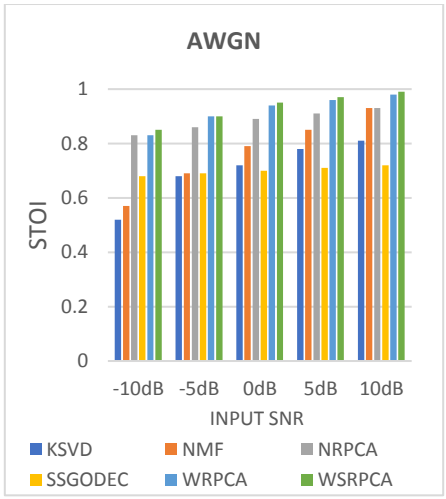


(d)

Figure 4.5(a-d) Performance comparison of the proposed SE algorithms with existing methods

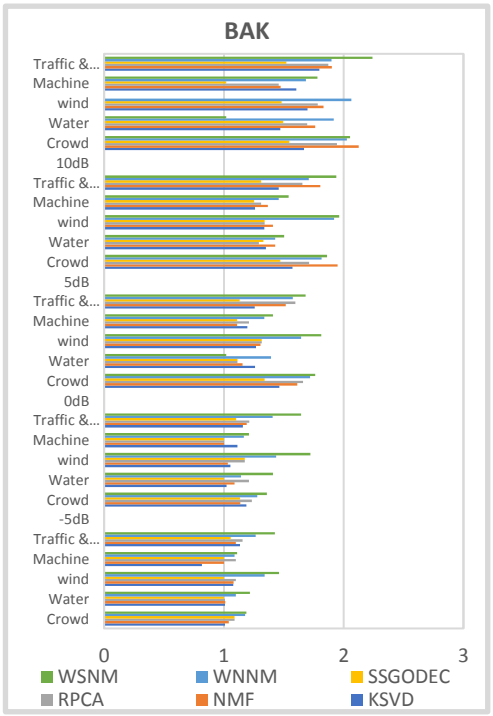


(e)

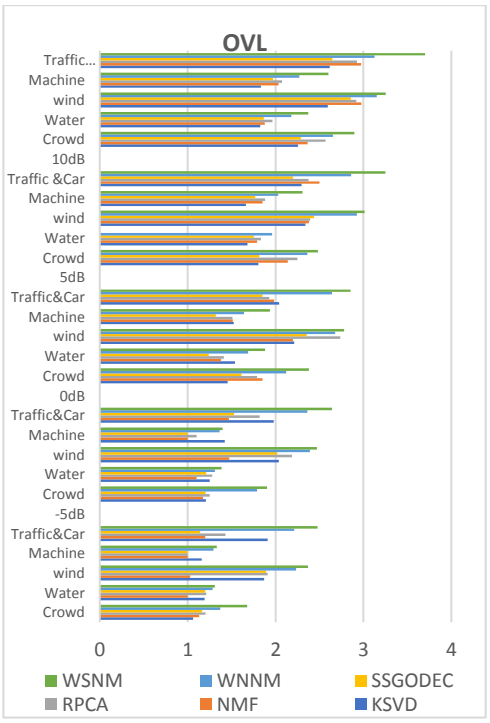


(f)

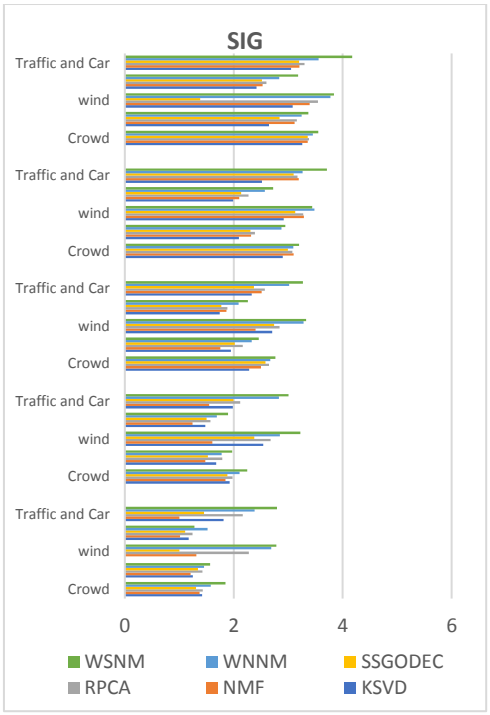
Figure 4.5(e-f) Performance comparison of the proposed SE algorithms with existing methods in terms of STOI using standard NOIZEUS data base



(a)



(b)



(c)

Figure 4.6(a-c) Performance comparison of the proposed SE algorithms with existing methods in terms of Objective metrics: a) BAK b) OVL c) SIG

When compared to baseline techniques, in low SNR situations (< 0dB) the proposed approach produced the greatest SIG values, demonstrating its usefulness in terms of speech content preservation. The approach outperformed in Traffic & car noise at all SNR levels by introducing low speech distortion and also very little residual noise in all noise settings.

4.7 Summary

The work proposes two convex optimization-based speech enhancement approaches that don't require any prior knowledge of speech or noise. Using a low-rank sparse matrix decomposition model, the approach decomposes input noisy speech magnitude spectra into low-rank noise and sparse speech components. Due to better characteristics, researchers believe that the algorithms recommended can provide a new and feasible approach to the SE problem in low SNR situations. The suggested methods are non-parametric strategies that do not require any assumptions about the spectral component distribution in speech or noise. In T-F domain, it only requires low-rank noise and sparse speech. The VAD approach is irrelevant and

unnecessary in SE framework since speech and noise components can be obtained simultaneously.

The contribution of this chapter is to provide an unsupervised speech-denoising strategy under diverse, strong, and unseen real-world nonstationary noisy settings that use low-rank and sparse decomposition models with a different objective function than conventional RPCA approaches. For each noisy input, all the regularization parameters are automatically modified and updated. Although existing methods such as KSVD and NMF methods can eliminate most interferers, under low SNR conditions (< 0dB) part of the recovered speech formant structures are lost during the matrix decomposition process, resulting in speech distortion. To alleviate speech distortion, we intend to build a novel low-rank and sparse matrix decomposition model by placing appropriate constraints on the sparse part. The present study assessed several objective measures widely used for evaluating speech quality. The performance metrics of RPCA, SS-GODEC, WNNM, and WSNM were evaluated and these were compared with KSVD and NMF in a wide range of acoustic conditions. The test conditions included speech signals from Noizeus databases and five real-world noises at five SNR levels (-10 dB, -5dB, 0 dB, 5, and 10 dB). Acoustic conditions with stationary noise at various SNR levels were included in our experiments and they reported excellent performance. With the proposed model, promising results were obtained in our experiments in terms of better objective measures like SDR, PESQ, SIG, BAK, OVL, and STOI values when compared with baseline methods such as KSVD, NMF, RPCA, and SS-GoDEC.

The proposed SE methods, however, were unable to completely remove background noise since the convex optimization techniques were inaccurate in estimating exact low rank. The problem of developing robust speech enhancement algorithms that can effectively remove background noise while maintaining good quality and intelligibility with low distortion in highly nonstationary and adversely noisy situations has yet to be solved. For Superior performance, models to estimate exact low-rank and noise type are to be explored.

Chapter 5

Unified Speech Enhancement Approach for Low Distortion Under Low SNR Environments

This chapter discusses the shortcomings of the suggested weighted low-rank sparse decomposition techniques that were previously developed. A novel SE approach is presented to address these issues in low SNR situations by integrating Wavelet, weighted low-rank sparse decomposition, and gamma-tone filter banks. Along with the results, the proposed SE method's implementation details are presented. The implementation details of the proposed SE method are explained along with the results.

5.1 Motivation

In low SNR and nonstationary noise situations, estimation of noise-related parameters in unsupervised speech enhancement (SE) approach is difficult. In recent SE approaches, the best results were achieved by partitioning noisy speech spectrogram into low-rank noise and sparse speech parts. Although the standard RPCA-based methods have proven useful for SE, there are a few potential drawbacks limiting its effectiveness. First, RPCA approaches are often approximated by spectrogram analysis using short-time Fourier transform (STFT). However, due to segmentation and windowing operations, there is distortion included in STFT process [141]. Second, most of these algorithms optimize their cost function, which is based on Euclidean distance (ED). The ED, on the other hand, can contribute to fairly significant reconstruction errors since it tends to overemphasize the accuracy of large values. As a result, the ED measure is not appropriate for processing speech signals [142]. Third, the majority of existing SE approaches improve STFT-based spectral magnitude while retaining the input noise-corrupted phase part unaltered, leading to distortion of the recovered speech signals and

reducing SE performance [156,157]. Fourth, the most well-known strategy used in the evaluation of RPCA is nuclear norm minimization (NNM), which precisely restores the rank of the matrix within specific constrained and theoretically guaranteed circumstances. In many cases, NNM fails to predict the rank of the matrix accurately [212]. These strategies are constrained by the estimate of the real rank of noise, and they do not fully use the low-rank characteristics in optimization.

As a result, discrete wavelet packet transform (DWPT)-based SE techniques [157] have been developed, to overcome previously mentioned distortion problems due to the use of noisy phase, by directly processing signals in the temporal domain and achieving acceptable SE performance. According to a recent study, monaural mixed audio can be separated more effectively on a cochleagram than on a spectrogram [158]. Because the cochleagram is created from non-uniform time-frequency (T-F) transform that simulates the human ear, the T-F units in the sensitive low-frequency portions have greater resolution than those in the high-frequency regions. In reality, speech and noise respond differently on the cochleagram. As a result, using a sparse and low-rank decomposition model to improve speech through mask estimation on the cochleagram appears promising. The Frobenius norm measure is commonly used in the cost function since it has been examined in studies of sparse and low-rank models without additional constraints to regularize the decomposed components. In addition, the decomposed speech involved should be nonnegative. Based on these findings, we chose to employ a Non-negative RPCA (NRPCA) scheme with DWPT that improves RPCA-based SE in an unsupervised manner to avoid the aforementioned shortcomings.

5.2 Introduction

In the literature, different cost functions, including Euclidean distance (ED), Kullback-Leibler divergence (KLD), and Itakura-Saito divergence (ISD), have been used as indicators of the quality of decomposition. Experimental evidence has shown that KLD works better than the squared ED or ISD for categorizing musical instruments [160]. As a result, the KLD measure has been used in this work to perform sparse and low-rank decomposition with

NRPCA (KLNRPCA) for SE [159]. However, the NNM-based NRPCA technique does not effectively utilize auditory perceptual features due to an improper calculation of matrix rank. Convex optimization models like weighted nuclear norm minimization (WNNM) and weighted Schatten p-norm minimization (WSNM), which overcome NNM limitations and obtain a better matrix rank approximation than NNM, have demonstrated their effectiveness in computer vision and machine learning domains. As a result, we introduced the SE framework in our earlier work [161,1*] by dividing a noisy spectrogram using weighted low-rank background noise and a sparse speech component, which yielded useful findings. Therefore, utilizing a weighted low-rank sparse decomposition methodology, discrete wavelet packet transform (DWPT), and the KL Divergence, under a variety of noise environments, we present in this study a novel approach for separating speech and noise cochleagram that enhances speech enhancement performance in various noise environments. Using the Noizeus data set, we evaluate and compare the effectiveness of DWPT-KL nonnegative RPCA (WKLNRPCA) with DWPT-KLWNNM (WKLWNNM) and DWPT-KLWSNM (WKLWSNM) for unsupervised speech enhancement. The outcomes demonstrate that in terms of speech quality and intelligibility, our approach greatly exceeds the conventional STFT-based SE approach. According to experimental results, the suggested approach achieves lower residual noise and less speech distortion in low-SNR settings than several of the most widely used SE approaches.

When compared to baseline approaches, our investigations using the suggested model in low SNR conditions revealed promising results in terms of enhanced objective measurements, including SDR, PESQ, STOI, SIG, BAK, and OVL values. The findings demonstrate that for all types of noise levels, the proposed technique offers an output SDR that is much greater than the input SDR. With traffic and car noise and wind noise, respectively, the WKLWSNM model shows improvements in output SNR of 11.95 dB and 5.46 dB at -10dB input SNR. For similar scenarios, PESQ values of 2.36 and 2.14 were obtained. For input noise between -10 and 0 dB, it was noticed that the suggested WKLWSNM approach with a binary T-F mask improved STOI scores.

5.3 DISCRETE WAVELET PACKET TRANSFORM

For DWPT/IDWPT, a series of clearly-defined low and high-pass filters, as well as a factor-2 down/up-sampling procedure, were used to provide distortion-free analysis/synthesis for an arbitrary signal. The structure of DWPT/IDWPT with a 2-level analysis/synthesis (l = 2) is shown in Figure 5.1. A full-band time signal m(n) is first divided into d_0^1 and d_1^1 sub-band signals, where d stands for the set of all level-l sub-band signals denoted as d_n^l , which provide information for the low and high frequencies, respectively, on the left side. Each of the two sub-band signals is subjected to the decomposition procedure again to produce four sub-band signals. By framing the sequence, a signal space is produced for each sub-band signal.

The DWPT process can be formulated as:

$$d_n^l = DWPT_n^l\{\text{m}\}, \text{n} = 1, 2, 3 \ldots 2^l \quad \ldots. (5.1)$$

where $\{d_n^l\}_{n=1}^{2^l}$, denotes the nth sub-band signal from a level-l DWPT.

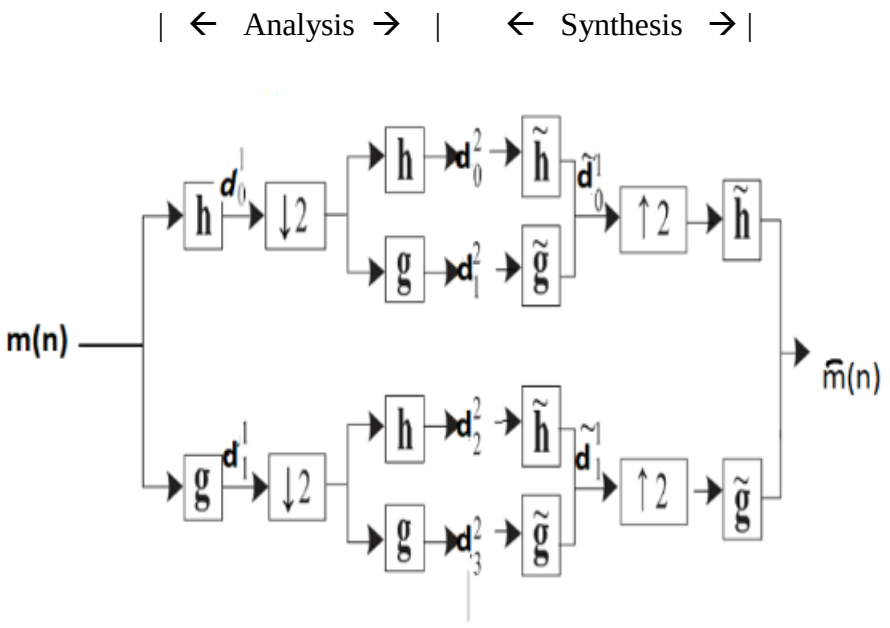


Figure 5.1: Structure of DWPT/IDWT two-level decompositions

The frequency responses of the low and high pass decomposition filters are denoted by h and g, respectively. The IDWPT reconstructs a full-band time signal by combining the sub-band signals. As a result, IDWPT can be written as $\widehat{m}$= IDWPT^{l} $\{d\}$. The input signal will be the same as the reconstructed signal, namely $\widehat{m} = IDWPT_n^l$= m, assuming a well-defined filter set.

5.3.1 Gammatone filter bank

The proposed SE technique decomposes the input noisy speech signal into a T-F representation using a Gammatone filter bank [159], also referred to as cochlear filtering method. The filter bank's impulse response is provided as follows:

$$GFB(\omega, t) = \begin{cases} t^{n-1} e^{-2\pi\mu t} \cos(2\pi\omega t), & t \geq 0 \\ 0 \quad Otherwise \end{cases} \qquad \text{---(5.2)}$$

Where ‘n’ refers to the filter's order, μ stands for the rectangular bandwidth that increases as the centre frequency ω increases. The response is shifted backward by (k-1)/2 πμ to account for filter delay. The output of each filter channel is used to produce time frames with a 50% overlap between them. The T-F spectra of each filter output are combined to create the cochleagram.

5.3.2 SE Using Weighted low-rank sparse models

A SE technique for speech and noise spectrogram separation by applying weighted low rank and sparsity requirements, as RPCA and SS-GODEC explicitly account for deviations of the speech and noise time-frequency matrices from the idealised sparse and low-rank model. The effectiveness of enhancement using singular value decomposition, the ADMM, and the accelerated proximal gradient line search method is enhanced thanks to the WNNM's low rankness. Therefore, a WNNM-based RPCA improvement model is suggested here that

outperforms NNM-based techniques by taking advantage of the high correlation of speech signals. In order to efficiently carry out low-rank regularization, extensive studies resulted in the development of a new RPCA model, the weighted Schatten p-norm minimization (WSNM) model.

5.3.2.1 SE Method Using NRPCA

Considering that a noisy speech signal is composed of clean speech signal *s(n)* and an additive uncorrelated noise signal *l(n)*, which is represented as follows:

$$m(n) = l(n) + s(n) \quad \ldots (5.3)$$

Based on the sparse and low-rank hypothesis for speech and noises, it is possible to decompose the T-F representation of m(n), i.e., $M \in R^{kxn}$, into two terms, represented as follows

$$M = L_0 + S_0 \quad \ldots (5.4)$$

where S_0 is a component for sparse speech and L_0 is a component for low-rank noise [147,200]. The RPCA approach finds a low-rank version and the sparse version of a noisy speech data matrix using Eq(5.5) as follows

$$\arg\min_{L,S} (\|L\|_* + \lambda \|S\|_1) \text{ s.t } M=L+S \quad \ldots (5.5)$$

Where $\|.\|_1, \|.\|_*$ symbolize l_1 matrix norm and nuclear norm, a positive constant λ regulates the relative weight between rank minimization and l_0-norm respectively. To preserve the physical significance of the cochleagram, the decomposed terms must be nonnegative, hence the following nonnegative robust principal component analysis (NRPCA) model is proposed:

$$\arg\min_{L,S} (\|L\|_* + \lambda \|S\|_1)$$

$$\text{s.t.} \quad M = L + S, L \geq 0, S \geq 0 \qquad \ldots \quad (5.6)$$

A Lagrange multiplier Y is used in the Lagrange method to create an unconstrained function. The Y value from the previous iteration is used to determine the optimal values of L and S. In order to reach the global optimum, the values of L, S, and Y are therefore updated in this manner.

According to preliminary experiments, the NRPCA model in Eq. (5.5) introduces musical noise into the reconstructed speech signal due to non-negativity restriction. Therefore, the formant structure of the original speech cannot be effectively and robustly extracted. This changes the NRPCA method to use the GO-Dec approach [153]. The cochleagram of the noisy speech M is thus represented as the summation of L, S, and D, that is, M = L+ S + D, where L and S are low-rank and sparse elements and D is a noise component that perturbs the ideal low-rank and sparse feature. The optimization of the objective function is therefore expressed as follows in Eq (5.7):

$$\arg\min_{L,S} \|M - L - S\|_F^2$$
$$\text{s.t } \operatorname{rank}(L) \leq r \quad \text{and } \operatorname{card}(S) \leq c \qquad (5.7)$$

Where r and *c* define the rank of L and cardinality of S, respectively. In order to achieve the predefined requirements for rank and cardinality of their support set, L and S must be chosen while minimizing the noise power defined as:

$$\|E\|^2 = \|M - L - S\|_F^2 \quad \cdots \qquad (5.8)$$

is minimized.

The aim of NNM decomposition in RPCA is to minimize $\|L\|_*$, so that the hidden low-rank matrix L can be recovered from the corrupted observation matrix M. However, in many realistic situations, NNM is unable to estimate the rank of the matrix accurately since the rank

components are frequently over-shrunk. Furthermore, the technique's practical utility is also constrained by the fact that it gives uniform weights to all rank components or singular values, which leads to a biased estimation of low-rank and sparse elements. In response to singular values which have different meanings, the weighted nuclear norm minimization (WNNM) method was introduced. WNNM greatly enhances flexibility and generalizes NNM. So, in this research, we propose to investigate WNNM and assess its minimization approach.

The original Low-Rank Matrix Approximation (LRMA) problem can be approximated more precisely using WSNM since it can accommodate a wide range of rank components. WNNM is a generalized variant known as WSNM. WSNM changes into WNNM when power p = 1. A Weighted Schatten p-Norm exists for the matrix $M \in R^{m \times n}$ with power p. The p-norm low-ranking approximation is used by WSNM in place of the nuclear norm flow-ranking approximation. WSNM Low-rank approximation typically performs low-rank regularization efficiently wherein we use Augmented Lagrangian function, represented in Eq(5.9)

$$X(L_w,S,Y,\mu) = \arg\min_{L,S} ||L||^{p}_{w,SP} + \lambda ||S||_1 + \langle Y, M-L-S \rangle + \mu/2 \, ||M-L-S||_F^2 \quad \cdots (5.9)$$

where Eq(5.9) expresses the weighted vector values as follows:

$$\omega i = C\sqrt{(mxn)}/(\sigma i(M)+\epsilon) \quad \ldots. (5.10)$$

Where the vector $w_i = [w_1, w_2 ..., w_n]$ and $w_i \geq 0$ is a non-negative weight given to σi (M). Based on prior knowledge and understanding of the issue, the rational weights criterion for weighting may be selected, significantly enhancing the capacity to represent the original data from the noisy input. Since they represent the energy of the main components of M, the large singular values of M are known to be highly significant than smaller ones in natural speech. The higher the individual values, the smaller they need to be shrunk during denoising. It implies that the weight assigned to σ_i(M), the i^{th} singular value of M, should be inversely proportional to σ_i (M).

The experiments have shown that a generalized KL cost function performed best in speech and noise separation tasks. Therefore, a refined model with L_0 and S_0 constraints, the KL divergence (KLD) cost function has been introduced. Similar to the human auditory system indicated by the formula in Eq(5.11), the KL scale invariant metric considers both low and high values while performing calculations.

$$KLD(M||Y) = \sum_{i=1}^{N} P_{Mi} log \frac{P_{Mi}}{P_{Yi}} - \sum_{i=1}^{N} P_{Yi} log \frac{P_{Mi}}{P_{Yi}} \approx 0 \quad .. (5.11)$$

The cost function of KLD is to minimize the KL measure between Y and its sparse and low-rank sum M, where L and S are constraints that need to be nonnegative of elements in the cochleagram. Since the cochlear magnitude values of speech signals are nonnegative, the decomposed S component follows the non-negativity criterion. Negative components might cause additional unpleasant residual noise if the non-negativity criterion is not imposed. The temporal gradient operator (β) is employed to avoid negative components. As a result, the KL-NRPCA model is formulated as:

$$\text{Arg min}_{L,S} \, KLD(\| M, S + L\|) + \lambda\|S\|_1 + \beta\|L\|_* \quad (5.12)$$

$$\text{s, t } \mathbf{L} \geq 0, \mathbf{S} \geq 0$$

where β is the optimization parameter.

5.3.2.2 Optimization Algorithm using ADMM

Further, by including the auxiliary variables M, L_+, and S_+ , KLWLRS model in Eq(5.12) can be represented as

$$\arg\min_{L, S, L_+, S_+, M} \; KLD(\| M, Y \|) + \lambda\|S_+\|_1 + \beta\|L_+\|^p_{w,*} \quad \text{-------} \quad (5.13)$$

$$\text{s, t} \quad M=L+S, S_+=S, L_+ = L, L_+ \geq 0, S_+ \geq 0$$

The Augmented Lagrangian function for Eq (51.3) is

$$X_\Omega = \Omega/2 \, \|M-L-S+ \Psi_M\|_F^2 + \Omega/2 \, \|S - S_+ + \Psi s\|_F^2$$

$$+ \Omega/2 \, \|L - L_+ + \Psi_L\|_F^2 + \lambda \|S_+\|_1 + \beta \|L_+\|^p_{w,*} \quad \text{------} \quad (5.14)$$

Where the scaled dual variables are Ψ_M, Ψs, Ψ_L, and the scaling parameter is Ω. As the objective function in Eq(5.12) is separable, ADMM algorithm can be used to solve it. By solving related issues within ADMM framework, all the variables in Eq(5.12) are updated separately and alternatively. The gradient descent approach can be used to minimize X_Ω for each of the two primal variables, two auxiliary variables, and three dual variables.

.

Contrary to Eq(5.13), it is necessary to address the sub-problem of updating M as indicated in Eq(5.15).

$$M = \arg\min_{M \geq 0} \; KLD(M,Y) + \Omega/2 \, ||M\text{-}L\text{-}S + \Psi_M||_F^2 \qquad (5.15)$$

The goal of WKLNRPCA method is to lower KL distance between Y and the sum of sparse and low-rank components (expressed by M in the equation above), with both components in the cochleagram being nonnegative. So, as a way to improve the performance, a novel SE algorithm is proposed by cascading stages of DWPT, KLD and NRPCA leading to a novel SE Framework denoted by "WKLNRPCA". The algorithm for WKLNRPCA is shown below

Algorithm 1 for SE using Wavelet-KLD-NRPCA

Input: Noisy speech data matrix M - Sub-band wise

1: Initialize: $\lambda > 0$, k = 0, $L_+^{(0)} = S_+^{(0)} = 0$, $\Omega = 1$,
$\Psi_M^{(0)} = \Psi_L^{(0)} = \Psi_S^{(0)} = 0$, $\theta > 0$, $\mu = 1$, $\eta = 0.95$

2: While $k \leq \theta$ do

3: $\arg min_{X \geq 0}$ KLD(M,Y) + $\Omega_{/2} \parallel M - L^{(k)}\text{-}S^{(k)} + \Psi_M^{(k)} \parallel_F^2$

4: $S^{(k+1)} = (M^{(k+1)} - L^{(k)} + \Psi_M^{(k)} - \Psi_S^{(k)} + S_+^k) / 2$

5: $L^{(k+1)} = (M^{(k+1)} - S^{(k+1)} + \Psi_M^{(k)} - \Psi_L^{(k)} + L_+^k) / 2$

6: $S_+^{(k+1)} = S_{+\,\lambda/\Omega}(S^{(k+1)} + \Psi_S^{(k)})$

7: $U \sum V = svd\,(L^{(k+1)} + \Psi_L^{(k)})$

8: $L_+^{(k+1)} := US_{+\,\beta/\Omega}(\sum) V$

9: $\Psi_M^{(k+1)} = \Psi_M^{(k)} + \mu\,(M - L^{(k+1)}\text{-}S^{(k+1)})$

10: $\Psi_L^{(k+1)} = \Psi_L^{(k)} + \mu\,(L^{(k+1)} - L_+^{(k+1)})$

11: $\Psi_S^{(k+1)} = \Psi_S^{(k)} + \mu\,(S^{(k+1)} - S_+^{(k+1)})$

12: $\mu = \eta * \mu$, k = k+1

13: end While

14: Output Matrix ($\hat{L} = L_+^{(k)}$, $\hat{S} = S_+^{(k)}$)

Similarly, Cascading DWPT, KLD, and WNNM models lead to another novel SE Framework denoted as WKLWNNM. The algorithm for WKLWNNM is shown below

Algorithm 2 for SE using Wavelet-KLD-WNNM

Input: Noisy speech data Matrix, Sub-band -wise

1: Initialize : $\lambda > 0$, k = 0, $L_+^{(0)} = S_+^{(0)} = 0$,M, Y0 = 0; $\Omega = 1$, $\Psi_M^{(0)} = \Psi_L^{(0)} = \Psi_S^{(0)} = 0$, $\theta > 0$, $\mu = 1$,$\eta = 0.95$, p=1

2: While $k \leq \theta$ do

3: $\arg min_{Y \geq 0}\ KLD(M,Y) + \Omega_{/_2} \parallel M - L^{(k)} - S^{(k)} + \Psi_M^{(k)} \parallel_F^2$

4: $S^{(k+1)} = (M^{(k+1)} - L^{(k)} + \Psi_M^{(k)} - \Psi_S^{(k)} + S_+^k) / 2$

5: $L^{(k+1)} = (M^{(k+1)} - S^{(k+1)} + \Psi_M^{(k)} - \Psi_L^{(k)} + L_+^k) / 2$

6: $S_+^{(k+1)} = S_{+\ \lambda /_\Omega}(S^{(k+1)} + \Psi_S^{(k)})$

7: for each frame mj in M do

8: Search for a similar frame-group Mj

9: Predict weight vector w_i

10: $U \sum V = svd(L_w^{(k+1)} + \Psi_L^{(k)})$

11: $L_+^{(k+1)} := U S_{+\ \beta /_\Omega}\ (\sum_w) V$

12: $\Psi_M^{(k+1)} = \Psi_M^{(k)} + \mu\ (M - L^{(k+1)} - S^{(k+1)})$

13 : $\Psi_L^{(k+1)} = \Psi_L^{(k)} + \mu\ (L^{(k+1)} - L_+^{(k+1)})$

14: $\Psi_S^{(k+1)} = \Psi_S^{(k)} + \mu\ (S^{(k+1)} - S_+^{(k+1)})$

15: $\mu = \eta * \mu$, k= k+1

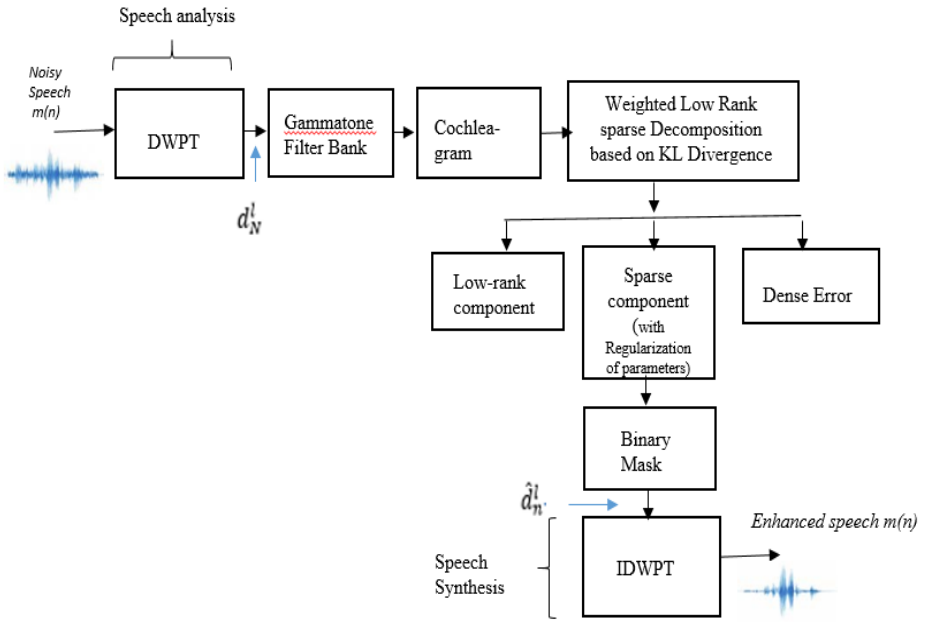
16: end While

17: Output Matrix ($\hat{L} = L_+^{(k)}$, $\hat{S} = S_+^{(k)}$)

5.4 DWPT- Weighted low-rank sparse model-based SE SYSTEM

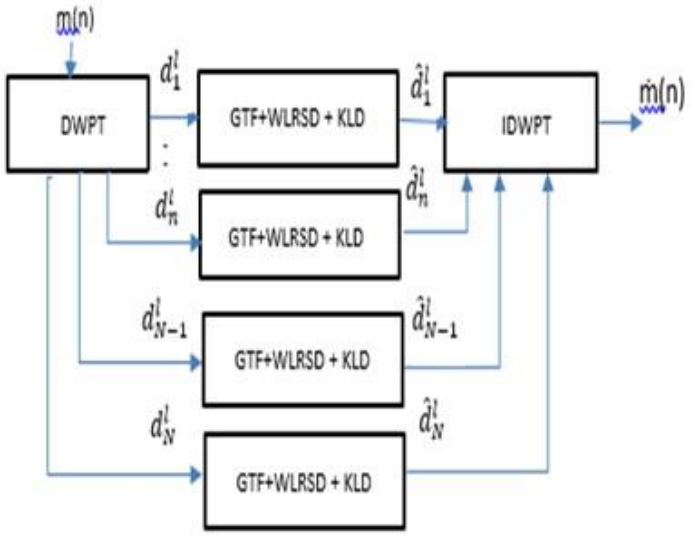
The proposed method uses Wavelet-based Weighted Low-Rank Sparse decomposition with KLD (WKLWLRSD) model-wise enhancement to each DWPT- sub-band time signal. The block diagram of WKLWLRSD model has decomposition, enhancement, and reconstruction

components, as shown in Fig. 5.2 (a), while the detailed sub-band-based WLRSD SE is shown in Fig. 5.2 (b). The overlap-add method is then employed to retrieve the gain function and enhance the associated sub-band signal for the virtual gain subspace. Finally, all enhanced sub-band sequences are subjected to inverse DWPT.



(a)

Fig: 5.2 a) Block diagram of Overall methodology of proposed SE framework



(b)

Fig: 5.2 b) Block diagram of DWPT-based SE Framework.

In the decomposition stage, the DWPT achieved in Eq. (5.1) was used to break a noise-corrupted signal m(n) into d_N^l-sub-band signals in the decomposition stage. Then, using WKLWLRSD model-wise enhancement method, each of the sub-band signals was separately enhanced. Finally, the enhanced signal $\widehat{m}$(n) is constructed by joining these updated sub-band signals using the IDWPT. The entire WKLWLRSD SE system, including the stages of decomposition, enhancement, and reconstruction, operate on the signals in time domain only.

5.5 SIMULATIONS AND RESULTS

This section presents the experiments and results for assessing the proposed noise reduction techniques' feasibility and effectiveness. Additionally, they enable calculating the impact of the variables that were considered in Sections 5.3 and 5.4.

The studies made use of the standard Noizeus corpus. The speech signals are available in Wav-file format with an 8 kHz sample rate. For this investigation, a total of 20 clear sentences were used. Clean sentences were mixed with five different SNR levels, such as -10, -5, 0, 5, and 10 dB, to produce the noisy speech. Waveforms with an 8 kHz sampling rate were also available for noise signals obtained from Noizeus corpus. For evaluation, five noise recordings were used, including a car driving in traffic, wind, a machine, a stream of bubbling water, and the cheering of a crowd of people. Simulated AWGN was also used with clean speech. As a result, there were $5 \cdot 20 \cdot 6 = 600$ test signals generated, each lasting about 3 seconds.

In order to illustrate the T-F representation of cochleagrams, Figure 5.3 shows the plots of all matrices that are related to decomposition. The cochleagram of the original clean speech is presented in Figure 5.1a. Figures 5.1b, 5.1c, and 5.1d display the cochleagrams of noisy speech, sparse and low-rank components respectively that decompose the input cochleogram by WKLWLRSD algorithm for 10dB of crowd noise case.

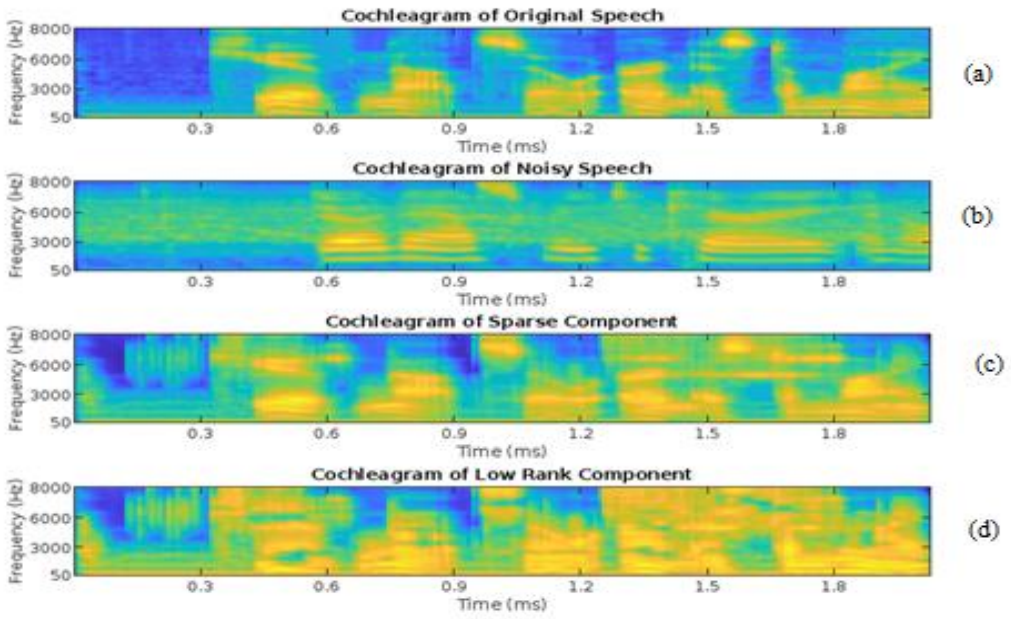


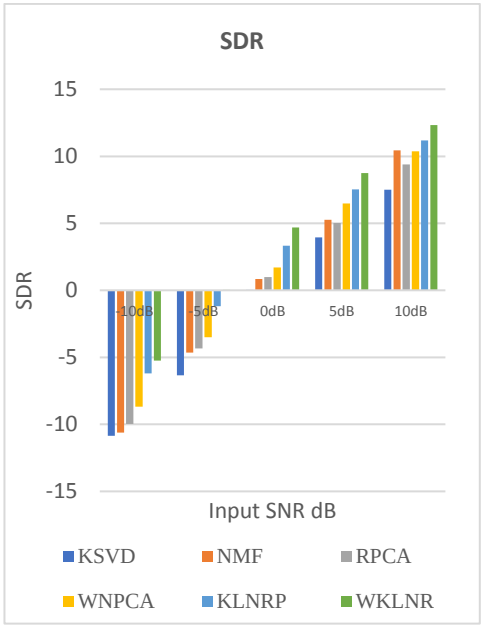
Fig. 5.3(a-d) Cochleagrams of a) original speech b)Noisy speech c) Sparse component d) Low-rank component

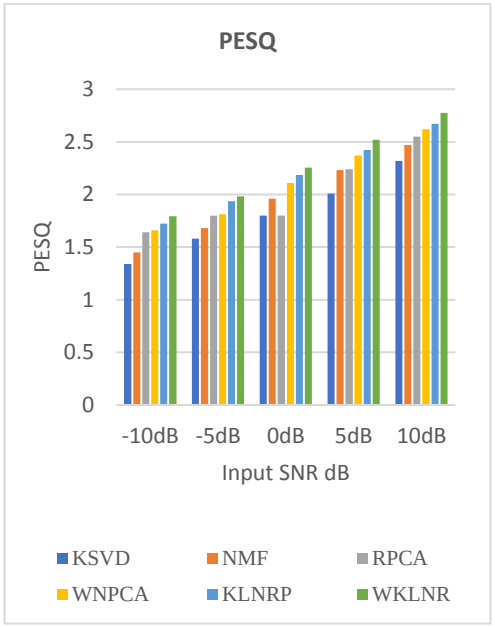
The regularization parameter selection process is an essential component of WLRSD approaches. A better general rule of thumb is suggested in the studies of [29] as $\lambda = (\max(e,f)^{-\frac{1}{2}})$. For WKLWLRSD models, λ could be empirically tuned. We select $\lambda = \mu(\max(m,n))^{-\frac{1}{2}}$. WKLWLRS problem formulation is comparable to that of PCP, which generalizes RPCA to have three decomposed terms. We intuitively assume that λβ= c is right, based on the choice of regularization settings in PCP. Through extensive experiments, we determine u and c. We select μ= 0.5 to achieve the desired balance between speech quality and noise suppression. We select c = 0.035 after fixing μ. We also observe that the Binary mask provides significantly better-increased speech quality.

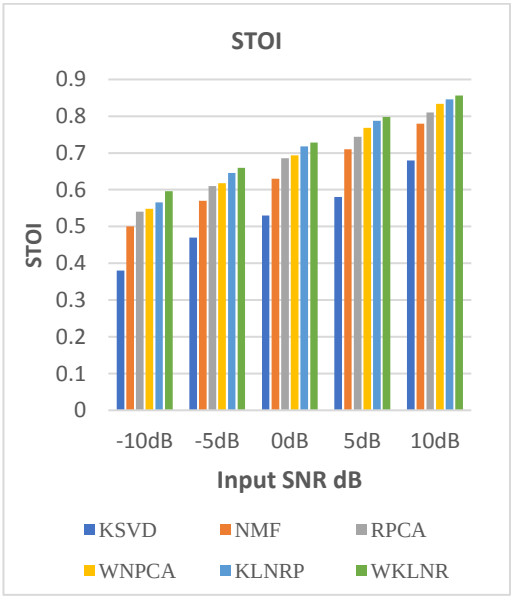
5.6 Evaluation of the Wavelet-based Weighted Low Rank and sparse decomposition models for SE

SE may also be impacted by the number of DWPT levels. To examine how the level of DWPT affects the outcome to enhance speech, we used several resolution levels (l = 2, 3, 4, and 5). As a result, we observed that when $l > 4$, the enhancing influence is significantly reduced. This is feasible because the frequency range of the filter narrows as DWPT decomposition carries on, containing low speech information in the frequency band. As a result, the total enhancing effect is diminished because WLRSD models are unable to enhance the speech of each band. Additionally, if $l < 3$, DWPT will not be able to provide the signal with more information, lowering the impact of enhancement, particularly in the case of a low SNR. The amount of computation doubles and the speech quality is improved when $l = 4$, but the noise in speech is not decreased. Therefore, we selected $l = 3$ to enhance speech.

For purposes of comparison, five baseline SE algorithms: KSVD, NMF, NRPCA, WAVELET-NRPCA(WNPCA), KL-NRPCA(KLNRP), and WAVELET-KL-NRPCA(WKLNR) were utilized as competing algorithms. The histograms in Figs 5.4(a-c) present the performance of these approaches in terms of averaged scores of SDR, PESQ, and STOI for five types of noise under various SNR levels. The proposed WKLNR SE technique demonstrates that it outperforms the other five techniques. When compared to KSVD, NMF, and NRPCA average SDR improvements in speech quality were 4.12 %. When compared to NRPCA, WNPCA, and KLNRP, the proposed WKLNR method improves PESQ and STOI by 10.6% and 2.7 %, respectively.





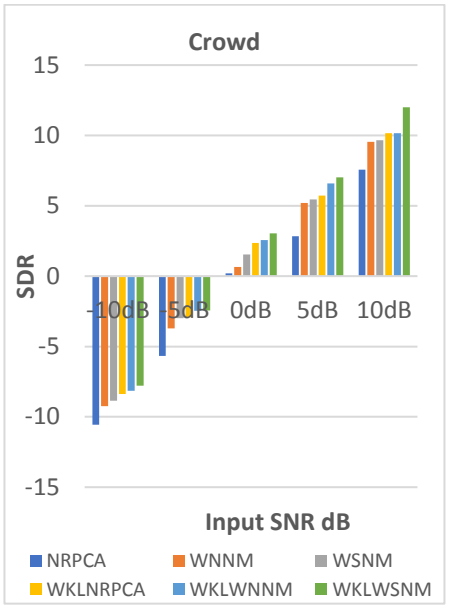


(c)

Fig.5.4(a-c) Comparison of the proposed SE Techniques with classical methods in terms of Average SDR, PESQ, and STOI values.

Further, the histograms in figures (5.5-5.7) show the comparisons of three existing baseline SE techniques: NRPCA, WNNM, and WSNM with three possible alternate cascade combinations: Wavelet-KL-NRPCA(WKLNRPCA), Wavelet-KL-WNNM) WKLWNNM, Wavelet-KL-WSNM(WKLWSNM) SE methods for five types of noise under various SNR situations. The proposed WKLWSNM SE technique demonstrates that it outperforms the other five methods. When compared to NRPCA, WNNM, and WSNM, the proposed cascade formations: WAVELET-NRPCA, KL-RPCA, and WAVELET-KL-NRPCA methods improve SDR and STOI significantly.

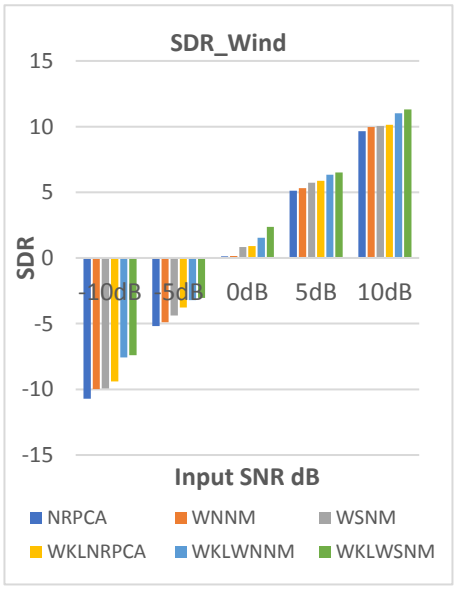
The suggested algorithm and the existing approaches were compared in terms of SDR in Figures 5.5(a-e), which was used to look at the reduction of distortion in enhanced speech. The results show that the suggested algorithm consistently produced the greatest SDR under all noisy environments and at all SNR levels. The high SDR in Figure 5.5(a-f) shows that the proposed technique is capable of reducing distortion in situations with high levels of noise (-10 dB, -5 dB). The suggested approach performed significantly well in Traffic & car-noisy environments and scored higher than competing methods in terms of SDR, although only at 5 dB (SDR = 6.23dB) and 10 dB (SDR = 11.02dB) for machine and 10 dB (SDR = 10.71dB) for water noise, respectively. When compared to competing approaches, WKLWSNM and WKLWNNM performed better. At all SNR levels and in every noisy condition, significant improvements have been seen, especially at low SNR levels noise (-10 dB, -5 dB). The Traffic & car noise level of -10 dB (ΔSDR = 13.18dB) and water noise level of 10 dB (ΔSDR = 0.17dB) have the highest and lowest SDR improvements, respectively.


Crowd
15
10
5
0
-5
-10
-15
SDR
-10dB
-5dB
0dB
5dB
10dB
Input SNR dB
NRPCA
WNNM
WSNM
WKLNRPCA
WKLWNNM
WKLWSNM

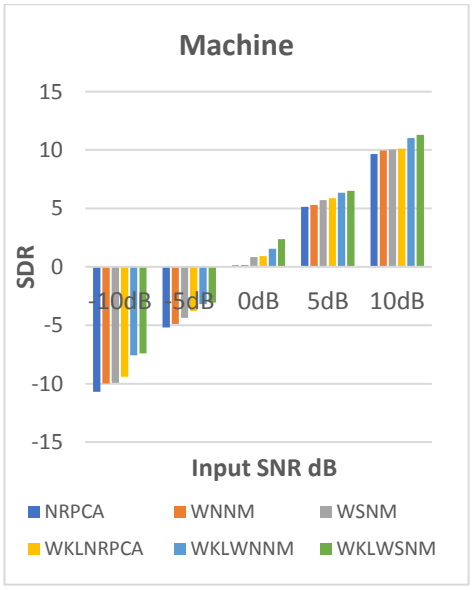
(a)


SDR_Water
15
10
5
0
-5
-10
-15
SDR
-10dB
-5dB
0dB
5dB
10dB
Input SNR dB
NRPCA
WNNM
WSNM
WKLNRPCA
WKLWNNM
WKLWSNM

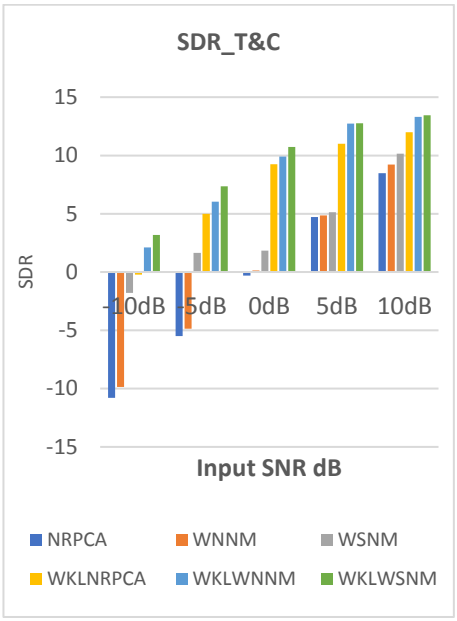
(b)


SDR_Wind
15
10
5
0
-5
-10
-15
SDR
-10dB
-5dB
0dB
5dB
10dB
Input SNR dB
NRPCA
WNNM
WSNM
WKLNRPCA
WKLWNNM
WKLWSNM

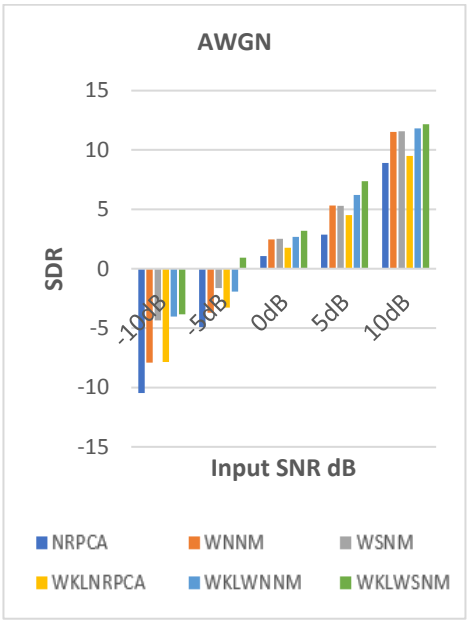
(c)


Machine
15
10
5
0
-5
-10
-15
SDR
-10dB
-5dB
0dB
5dB
10dB
Input SNR dB
NRPCA
WNNM
WSNM
WKLNRPCA
WKLWNNM
WKLWSNM

(d)



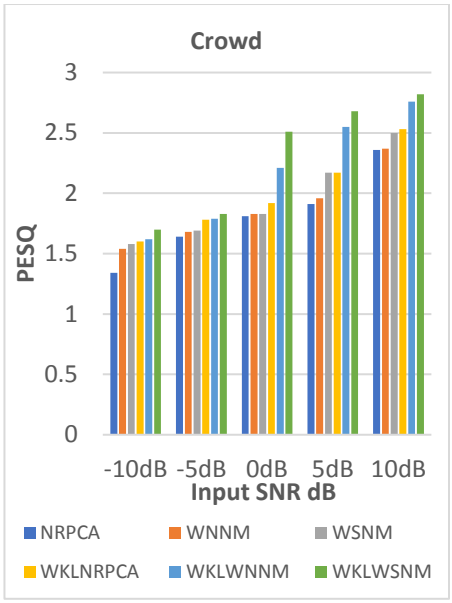
(e)



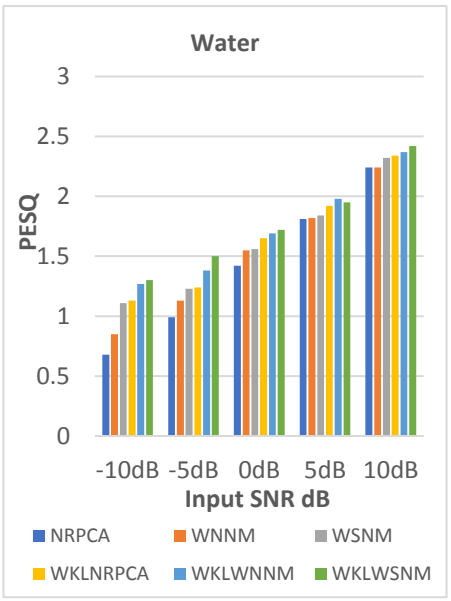
(f)

Fig. 5.5(a-f) Comparison of the suggested SE algorithms against baseline methods in terms of SDR for a) Crowd b) Water c) Wind d) Machine e) Traffic & Car f) AWGN noise

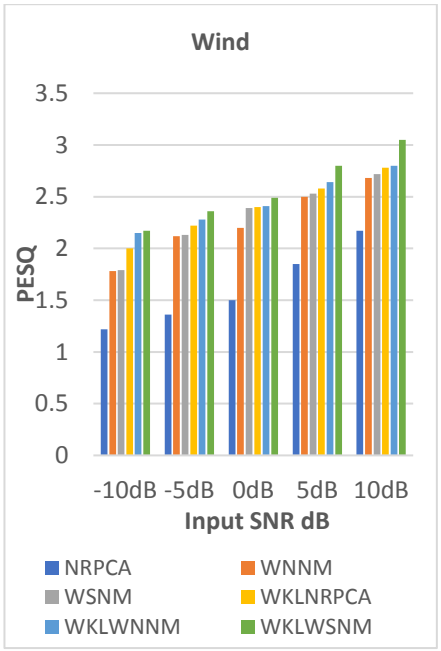
A Comparision of the performance of the suggested algorithm and competing approaches using PESQ, which measures the overall quality of enhanced speech, is shown in Figure 5.6 (a-f). In all noisy environments, the proposed algorithm consistently produced the highest PESQ scores, primarily at low SNR levels (-10 dB and -5 dB). PESQ is improved at two different levels: -10 dB (ΔPESQ = 0.346) for Traffic & car noise and 10 dB (ΔPESQ = 0.083) for machine noise. However, less significant PESQ scores of 1.53 and 1.88 were noticed for machine noise at low SNR levels (-10 dB and -5 dB) respectively. The suggested method reported a less significant loss in speech quality than WKLNRPCA and WKLWNNM methods. At all SNR levels, WKLWSNM outperformed other competing techniques and in every noisy condition, an improvement was seen, notably at low SNR levels (-10dB and-5dB).


Crowd
PESQ
3
2.5
2
1.5
1
0.5
0
-10dB -5dB 0dB 5dB 10dB
Input SNR dB
NRPCA
WNNM
WSNM
WKLNRPCA
WKLWNNM
WKLWSNM

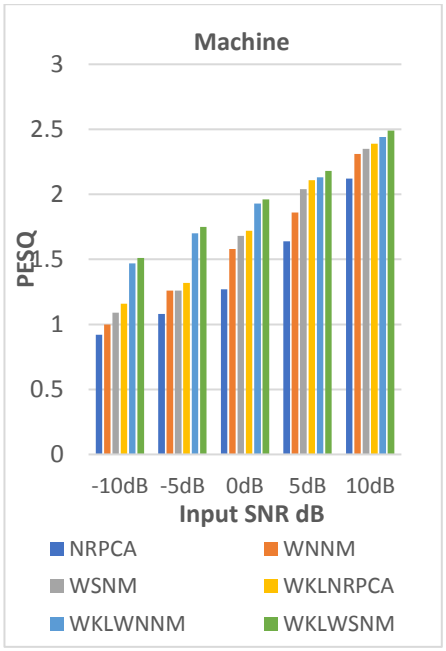
(a)


Water
PESQ
3
2.5
2
1.5
1
0.5
0
-10dB -5dB 0dB 5dB 10dB
Input SNR dB
NRPCA
WNNM
WSNM
WKLNRPCA
WKLWNNM
WKLWSNM

(b)


Wind
PESQ
3.5
3
2.5
2
1.5
1
0.5
0
-10dB -5dB 0dB 5dB 10dB
Input SNR dB
NRPCA
WNNM
WSNM
WKLNRPCA
WKLWNNM
WKLWSNM

(c)

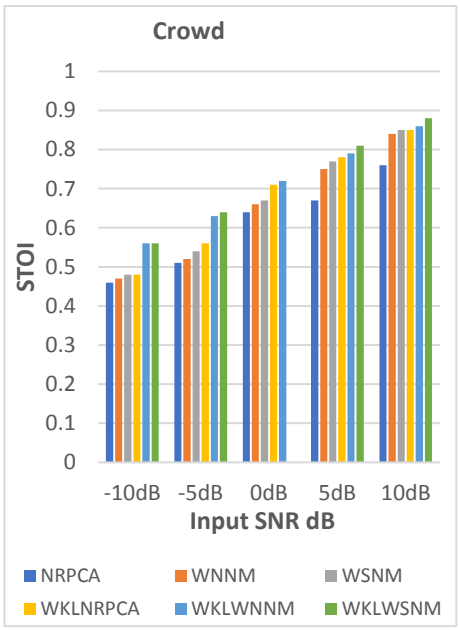

Machine
PESQ
3
2.5
2
1.5
1
0.5
0
-10dB -5dB 0dB 5dB 10dB
Input SNR dB
NRPCA
WNNM
WSNM
WKLNRPCA
WKLWNNM
WKLWSNM

(d)

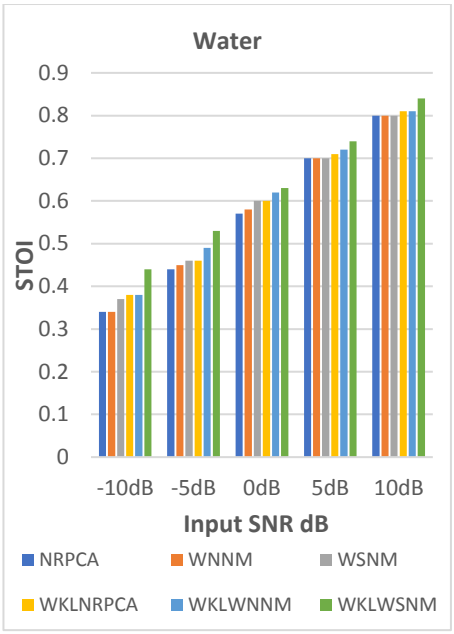
(e) (f)

Fig. 5.6(a-f) Comparison of the suggested SE algorithms against baseline methods in terms of PESQ for a) Crowd b) Water c) Wind d) Machine e) Traffic & Car f) AWGN noise

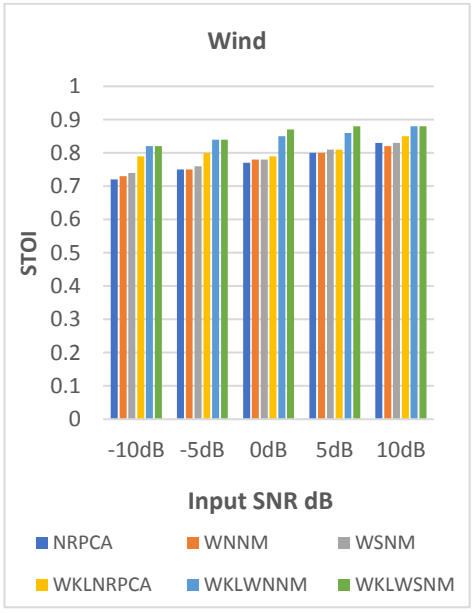
Figs. 5.7(a-f) displays the intelligibility prediction rates obtained using the STOI measure for processed speech using the suggested and alternate approaches. For the suggested method, the traffic & car and machine noise achieve the highest and lowest STOI outcomes, respectively. For SNR = 10 dB, all noise sources had good intelligibility scores of STOI > 86%. However, with the low SNR levels, significant discrepancies in prediction rates were observed. It is clear that the suggested algorithm performed better than alternatives at all SNR levels and in all noisy settings. The best overall prediction rate was obtained using the suggested algorithm, which was 88.4 %. Fig. 5.7(a-f) shows that the suggested algorithm produced the best average rates for five nonstationary noise and AWGN noise sources when compared to NRPCA, WNNM, WSNM, WKLNRPCA, WKLWNNM. For instance, at -10 dB, the suggested approach increases the predicted rate of Traffic & car noise on overall average from 54.4 % with NRPCA to 88.2 %. For Traffic & car noise at 10dB, for WKLNRPCA, and WKLWNNM methods, the overall STOI improves by 68.3 %, and 72.8 %, and for wind noise, by 66.5 %, and 70.2 %, respectively.


Crowd
STOI
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-10dB
-5dB
0dB
5dB
10dB
Input SNR dB
NRPCA
WNNM
WSNM
WKLNRPCA
WKLWNNM
WKLWSNM

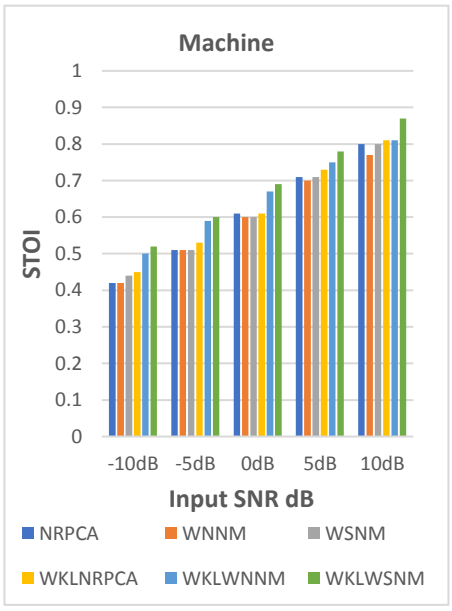
(a)


Water
STOI
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-10dB
-5dB
0dB
5dB
10dB
Input SNR dB
NRPCA
WNNM
WSNM
WKLNRPCA
WKLWNNM
WKLWSNM

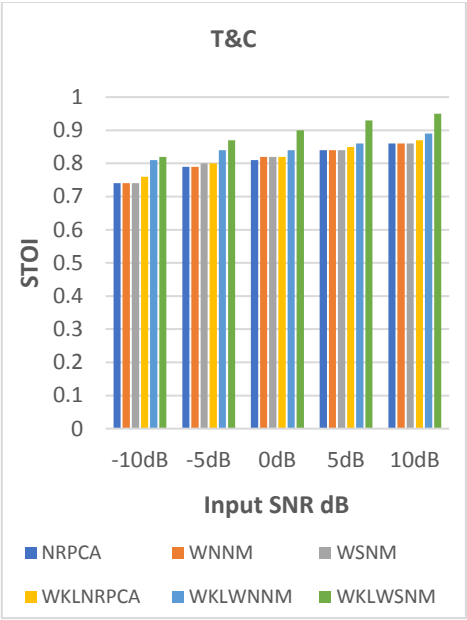
(b)


Wind
STOI
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-10dB
-5dB
0dB
5dB
10dB
Input SNR dB
NRPCA
WNNM
WSNM
WKLNRPCA
WKLWNNM
WKLWSNM

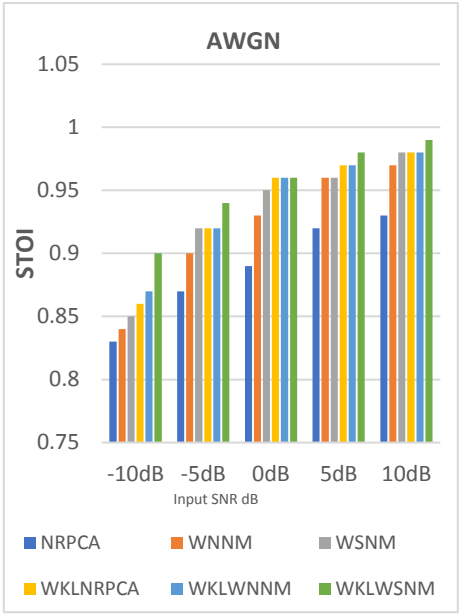
(c)


Machine
STOI
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-10dB
-5dB
0dB
5dB
10dB
Input SNR dB
NRPCA
WNNM
WSNM
WKLNRPCA
WKLWNNM
WKLWSNM

(d)



(e)

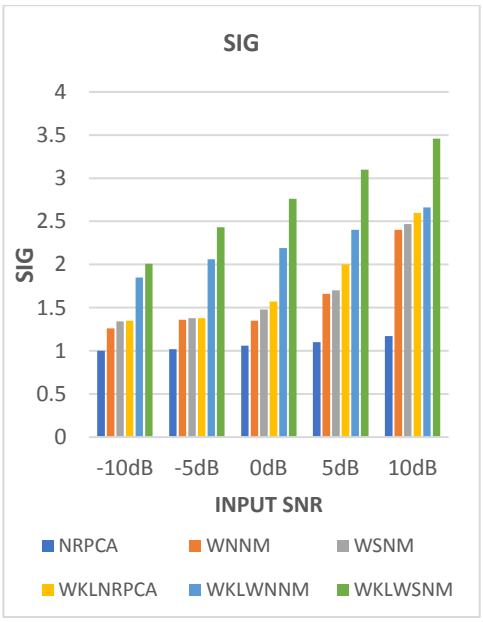


(f)

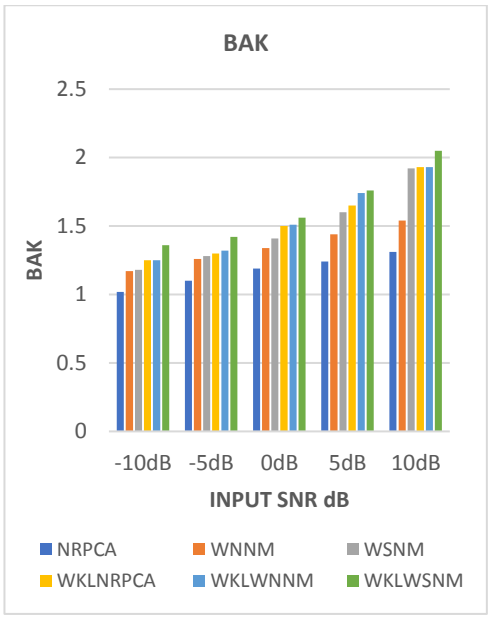
Fig. 5.7(a-f) Comparison of the suggested SE algorithms against baseline methods in terms of STOI score for a) Crowd b) Water c) Wind d) Machine e) Traffic & Car f) AWGN noise

In order to investigate speech distortion and residual noise in enhanced speech, figures 5.8(a-c) compares the proposed algorithm and the competing approaches in terms of average SIG, OVL, and BAK. The SIG calculates the distortion that results from processing noisy speech. Better performance is implied by a high value. Results in fig 5.8(a) show that, in comparison to competing methods, the proposed algorithm significantly reduces speech distortion in all noisy situations and at all SNR levels, except for -5dB (SIG = 2.76) crowd noise and -10dB (SIG = 1.98) machine noise, respectively. Compared to the suggested technique, the NRPCA and WNNM methods achieved a more observable loss of speech contents. Although rigorous noise estimates in such systems can efficiently reduce background noise it also removes the crucial speech content. Consequently, this adds significant speech distortion. The BAK results are shown in figure 5.8(b). High BAK scores suggest that enhanced speech has low residual noise. In comparison to competing methods, the suggested algorithm effectively decreased background noise, resulting in low residual noise at all SNR levels and in all noise situations, except for -10dB (BAK = 1.35) crowd noise and -5dB (BAK = 1.43) machine noise. Low BAK values suggest significant residual noise in NRPCA-processed speech. Such methods provide disturbing residual noise as a result of too low noise estimation.

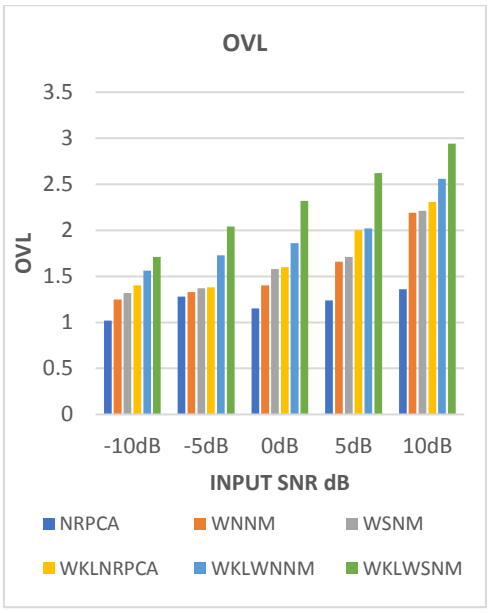
The OVL results of enhanced speech are seen in fig.5.8(c). At all SNR levels and in all noisy conditions, significant improvements have been made, especially at low SNR levels (-10dB and -5dB).



(a)



(b)



(c)

Figure 5.8(a-c). Average SIG-BAK-OVRL scores for proposed speech enhancement compared to the existing methods at different SNR input levels

5.7 Summary

The effectiveness of standard RPCA-based methods has been shown to be favourable for SE in low SNR environments. However, the performance of these SE approaches is constrained by the use of overlap-add in STFT process, noisy phase, the biased estimate of low rank in nuclear norm minimization, and Euclidian distance in the cost function, which may result in information loss in the reconstructed speech signal. To address these limitations, we propose to integrate the DWPT, gamma-tone filter bank, KLD, and WLRSD models in this work to develop an unsupervised SE framework. DWPT/IDWPT permits the decomposition and reconstruction of noisy input speech, while WLRSD model can improve sub-band signals. To improve noisy speech, cochleagrams were employed rather than magnitude spectrograms. The DWPT-WLRSD model produces less distortion compared to STFT-based WLRSD (WNNM, WSNM) models which improve speech quality and intelligibility. The low-rank noise component and the sparse speech component were separated from the noisy speech cochleagram. The speech component is extracted using a binary mask. In low SNR conditions, low residual noise and speech distortion were detected. The estimate of the clean speech is possible by applying regularization constraints. For each noisy input, all regularization settings were updated and changed. The findings indicate that the proposed algorithm has great potential for enhancing speech quality and intelligibility in strong noisy environments without the need for a noise estimator.

Chapter 6

Validation of the Proposed Speech Enhancement system

This chapter presents the validation results of proposed Speech enhancement techniques that are carried out by training Kaldi ASR to achieve low WER using different noises with SNRs ranging from -10dB to 10dB.

6.1 Motivation

In spite of a wide range of noise interferences that exist in the real world, human speech perception is robust. The effectiveness of ASR systems in recognizing words is close to 100% in conditions with no background noise. However, in the presence of strong background noise, the accuracy of single-channel ASR systems decreases significantly. One of the best ways to improve the robustness of a speech recognition system is to include a noise reduction (i.e SE) stage. Most single-channel speech enhancement (SE) approaches (denoising) have only provided marginal performance improvements over state-of-the-art ASR backends trained on multi-condition training data SE is a prominent technique for making the ASR more robust. Numerous attempts have been made to improve single-channel SE algorithms in terms of signal-based metrics, such as increased signal-to-noise ratio or decreased speech distortion.

6.2 Introduction

ASR is the process of translating a sequence of words spoken by a human into readable text using machines or software. ASR has progressed into a technology that is becoming increasingly prevalent in daily life and is emerging as an essential requirement

for many Human-Machine applications, including command and control applications, navigation, entertainment, etc. Modern ASR systems are getting close to performing at levels comparable to human recognition. However, due to acoustic interferences such as noise, it is still difficult to recognise speech via a distant microphone. Increased attention has been given to the issue of distant microphones. By using a multi-channel SE pre-processing with an ASR backend trained on multi-conditioned data, ASR performance may be significantly enhanced when a microphone array is available. However, there are many of circumstances where there is just one microphone accessible. In such situations, a single microphone's performance lags substantially behind that achieved by a microphone array. Though novel SE techniques were proposed in this work, the effectiveness of these methods is to be validated through ASR application. Further study is therefore needed on the developed SE approaches as frontends for ASR.

6.3 Overview of Speech Recognition system

The automatic speech recognizer shown in figure 6.1 have a speech input, feature extraction, decoder, and word output.

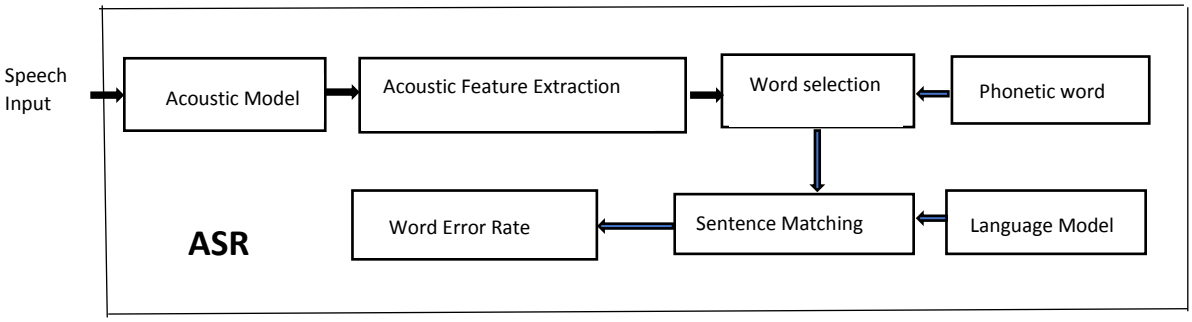


Fig.6.1 Basic components of Automatic Speech recognition system

Important components of speech recognition algorithms are acoustic modelling and language modelling. Acoustic models, dictionaries, and language models are used for decoding.

Acoustic Model: An acoustic model contains statistical representations of all the different sounds that make up a word. Each of these statistical representations is given the name

"phoneme." For speech recognition, the English language has about 40 different sounds that can be separated into 40 different phonemes.

Language Model: Word sequences are matched to sounds in order to distinguish between words that sound similar. Even if the audio sample is not perfectly grammatically correct or has skipped words, it assumes it is semantically and grammatically sound. As a result, adding a language model to the decoding process can improve ASR precision.

The acoustic models are trained using acoustic features from labelled data, such as the Wall Street Journal Corpus(WSJ), TIMIT, or any other transcribed speech corpus. The ASR employed in this work corresponds to kaldi recipe available as an open source for use by experts and researchers in the field of speech processing. The ultimate performance measure of SE algorithm for ASR will be always WER achieved over a particular dataset. To convert speech into text and increase transcription accuracy, a variety of algorithms and computational methods are used.

The top-level Kaldi's directory structure are : egs, src, tools, misc, and windows are top-level directories. We use egs and src directories. egs contains training recipes for major speech corpora. There are training recipes for wsj, timit, rm, and others. Each directory has several versions (s3, s4, s5, etc.) The latest version, s5, is used for new development or training. src contains most of the training recipes' source code.

Standard sub-directory structure for each training recipe directory. This is shown in egs/rm/s5. Top directory contains run.sh and two other scripts (cmd.sh and path.sh). conf, data, exp, local, steps, and utils are sub-directories (utilities). We'll mostly use data and exp. The data directory will eventually contain transcripts, dictionaries, etc. The exp directory will include the training, alignment, and acoustic model outputs.

In the testing stage, the language model weight is set to the value that provides the lowest WER in the development set.

$$WER = \frac{(\text{Substitutions+ Insertions+ Deletions})}{\text{Number of words spoken}}$$

Newly developed SE algorithms can be objectively compared by training Kaldi ASR to obtain low WER.

The procedure adopted for training acoustic model is as follows:

1. Get a speech transcript: For a more precise alignment, utterance (sentence) start and end times are useful, but not essential.

2. Transcription of Kaldi ASR : Kaldi requires different forms for acoustic model training. Start and finish times of each utterance, speaker ID, and list of all words and phonemes in the transcript are needed.

3. Obtaining acoustic features from audio: Mel Frequency Cepstral Coefficients (MFCC) are the most popular features, however PLP and other features are also available. Acoustic models are based on these qualities.

4.Train monophone models : A monophone model lacks context for the previous or next phone. It is a basic block for triphone models, which incorporate contextual information. While monophone models describe a single phoneme's auditory properties, phonemes vary greatly depending on context. The triphone models show a phoneme variant alongside two others (left and right).

5. Align audio with acoustic models: Acoustic training steps estimate the model's parameters, but the process can be enhanced by cycling through training and alignment phases. Viterbi training (related, but more computationally expensive procedures include the Forward-Backward algorithm and Expectation Maximization). By aligning the audio to the reference transcript with the latest acoustic model, additional training algorithms can refine the model's parameters. After each training step, audio and text will be realigned.

The figure 6.2(a) demonstrate the performance of kaldi ASR for trained ground truth audio data from Noizeus corpus and the figure 6.2 (b) the tested output as a transcript.

Ground truth :

sp1_f.wav: He wrote down a long list of items

sp1_m.wav: The birch canoe slid on the smooth planks

sp2_f.wav: The drip of the rain made a pleasant sound

sp2_m.wav: He knew the skill of the great young actress

sp3_f.wav: Smoke poured out of every crack

sp3_m.wav: Her purse was full of useless trash

sp4_f.wav: Hats are worn to tea and not to dinner

sp4_m.wav: Read verse out loud for pleasure

sp5_f.wav: The clothes dried on a thin wooden rack

sp5_m.wav: Wipe the grease off his dirty face

(a)

Transcript:

sp1_f.wav: he wrote down a long list of items

sp1_m.wav: the birds canoes slid on a smooth planks

sp2_f.wav: the trap of the rain made a pleasant sound

sp2_m.wav: he knew the skill of the great young actress

sp3_f.wav: smoke poured out of every crack

sp3_m.wav: her purse is full of useless trash

sp4_f.wav: have a warrant to t and not to dinner

sp4_m.wav: read reverse out loud for pleasure

sp5_f.wav: the clothes dried on a thin would and rack

sp5_m.wav: why degrees off his dirty face

(b)

Figure 6.2 screen shots of a) Trained text Ground truth audio data b) Transcript output

6.4 Simulation results on ASR

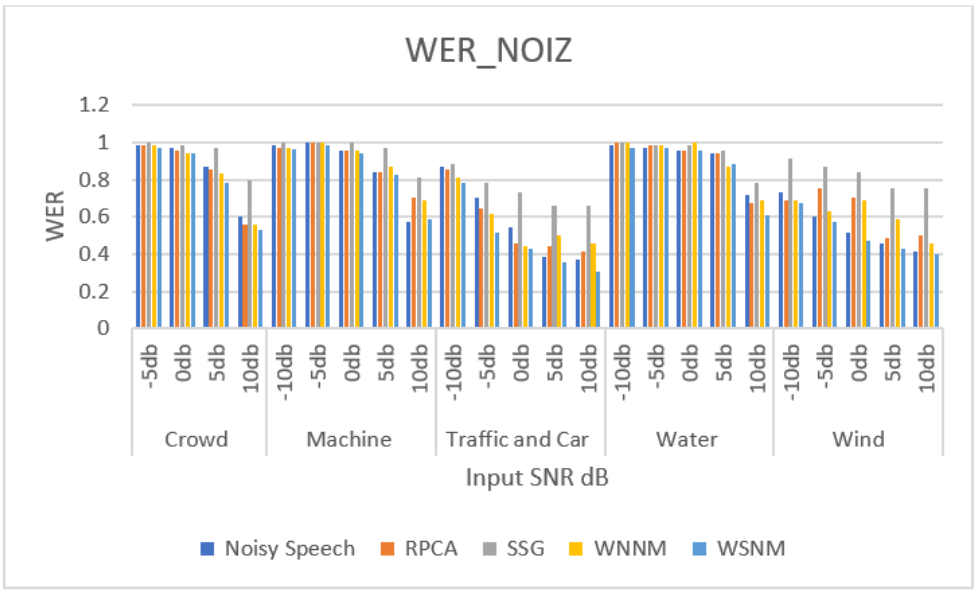
The experimental findings of our suggested approach are presented in this subsection, along with some discussions. The word error rate (WER) indicates that at varying input SNRs, the human ability to recognize speech contents remains resilient. However, in the presence of background noise, the accuracy of single-channel ASR systems decreases significantly. Most single-channel speech enhancement (SE) approaches (denoising) have only provided marginal performance improvements over state-of-the-art ASR backends trained on multi-condition training data. One of the best ways to improve the robustness of a speech recognition system is to include a noise reduction (SE) stage.

Testing the performance of each SE algorithm throughout the complete spectrum of acoustic circumstances takes a long time. As a result, it is preferable to estimate WER scores using more easily computed metrics during the development of the SE algorithm, where the clean speech reference is available. Predicting the performance of the SE algorithm is beneficial to correlate the improvements in WER with improvements in bss_eval metrics.

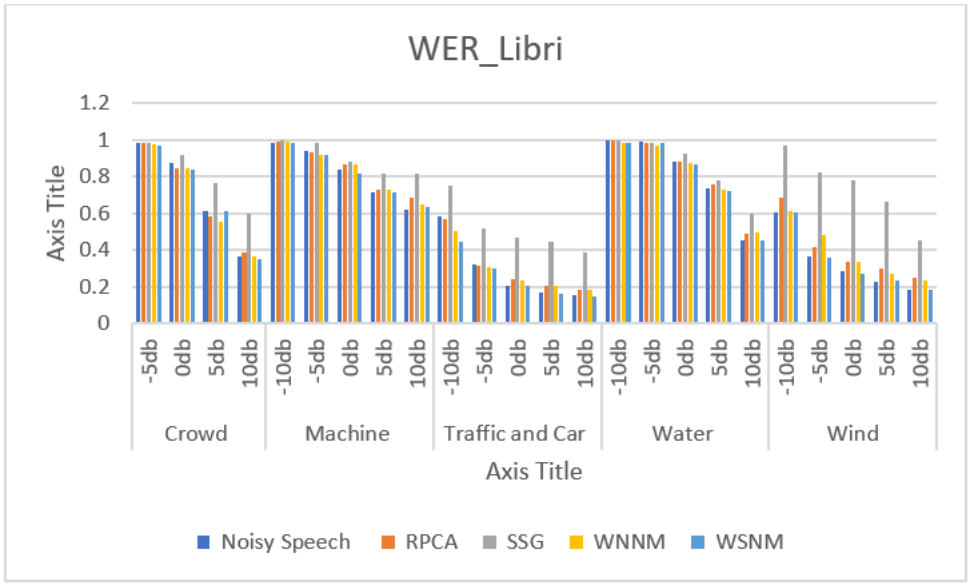
The second experiment is carried out to examine and contrast the SE algorithms for speech recognition with the RPCA models. This uses the Kaldi ASR repository to train and test WER on enhanced speech signals from the first experiment. We examined the Noizeus[145], Libri[217], TIMIT[218] databases and ASR backends to test the generalisation capability of proposed SE approaches.

From the clean speech data available, Khaldi ASR was trained, and WER was estimated as follows: NOIZEUS database WER= 0.2058, Libri speech database WER = 0.1, and TIMIT speech database WER= 0.3046**.** Figures 6.1(a-c) shows the performance comparison of baseline algorithms such as KSVD, NMF, and RPCA with initially proposed methods: WNNM and WSNM on Kaldi ASR using the above three databases. in terms of WER.

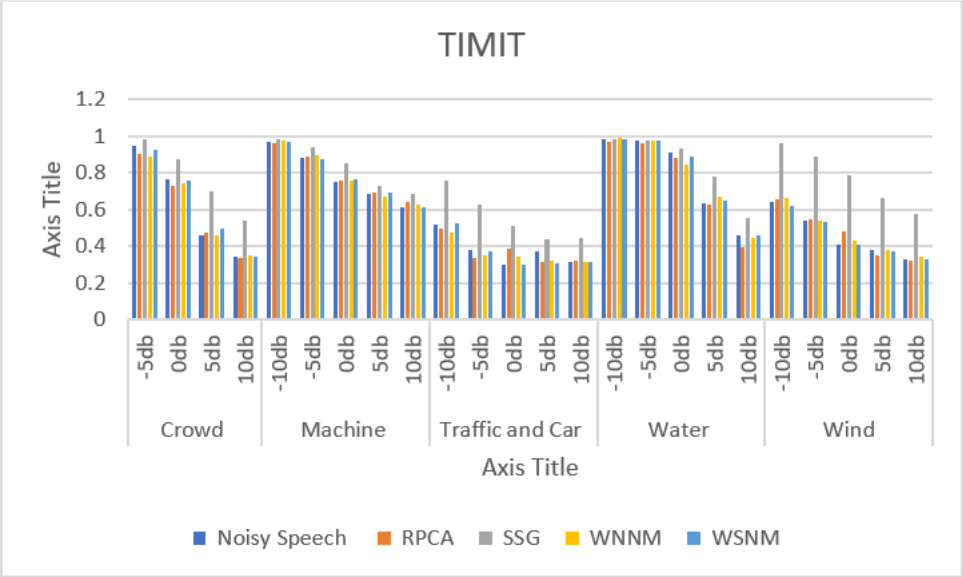
Figure 6.3(a-c) demonstrate the performance of the initially proposed speech enhancement algorithm in terms of Word Error Rate (WER) over noisy and baseline algorithms using a) Noizeus b)Libri c) TIMIT Database



(a)



(b)

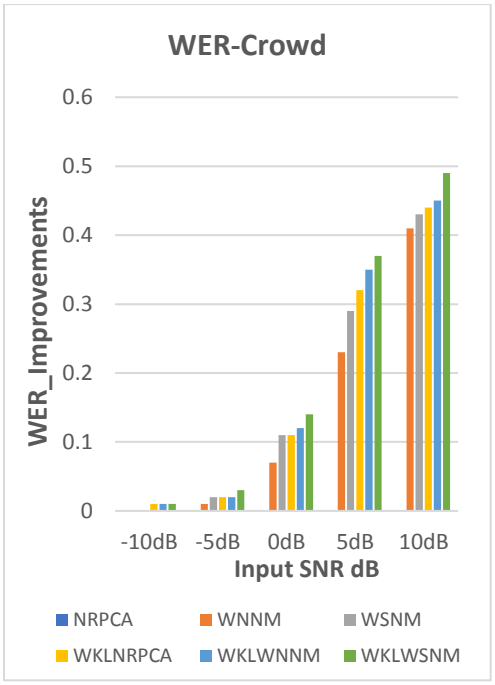


(c)

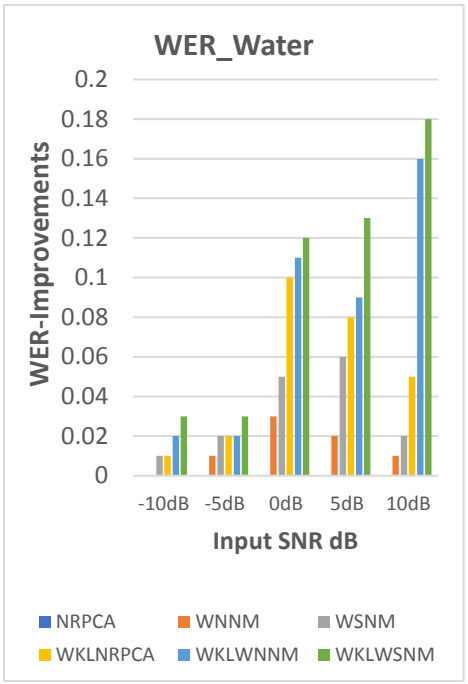
Figure 6.3(a-c).Performance comparison of the initially proposed speech enhancement algorithm in terms of WER over noisy and baseline algorithms using a) Noizeus b)Libri c) TIMIT Database

Among the speech processing schemes experimented, particularly proposed speech enhancement algorithms performed well in terms of the WER over noisy speech, depicted in Figures 6.3(a-c). The ASR results show that the performance of our proposed approach with the Libri database produced the lowest WER values. It is noticed that in low SNR conditions among the various noises Traffic & car, and wind noise cases shown a better performance in terms of lowest WER.

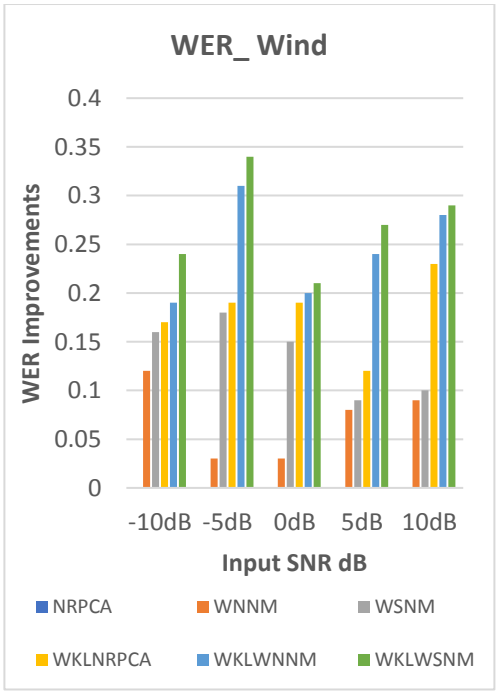
Studies on proposed wavelet-based speech enhancement algorithms outperformed the methods proposed earlier. More testing is being done to confirm their efficacy in ASR systems. Figure 6.4(a-f) provides a thorough analysis of each noise as it relates to WER improvements over NRPCA technique. Significant improvements are demonstrated with the proposed approaches. The WER using NRPCA approach is the highest for the crowd and machine noise at -10dB. WKLWSNM techniques shown a significant progress (with lowest WER) at low SNR, traffic & car, and wind noise scenarios. The AWGN was used to capture ASR's best performance across all noise kinds and SNR ranges.


WER-Crowd
WER_Improvements
0.6
0.5
0.4
0.3
0.2
0.1
0
-10dB -5dB 0dB 5dB 10dB
Input SNR dB
NRPCA WNNM WSNM
WKLNRPCA WKLWNNM WKLWSNM

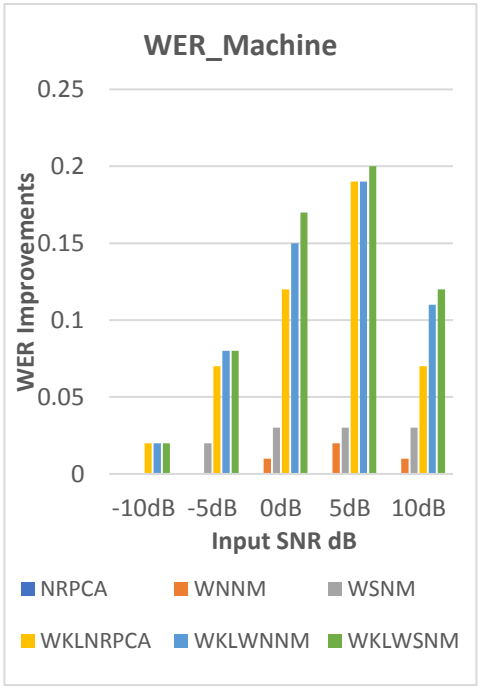
(a)


WER_Water
WER-Improvements
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-10dB -5dB 0dB 5dB 10dB
Input SNR dB
NRPCA WNNM WSNM
WKLNRPCA WKLWNNM WKLWSNM

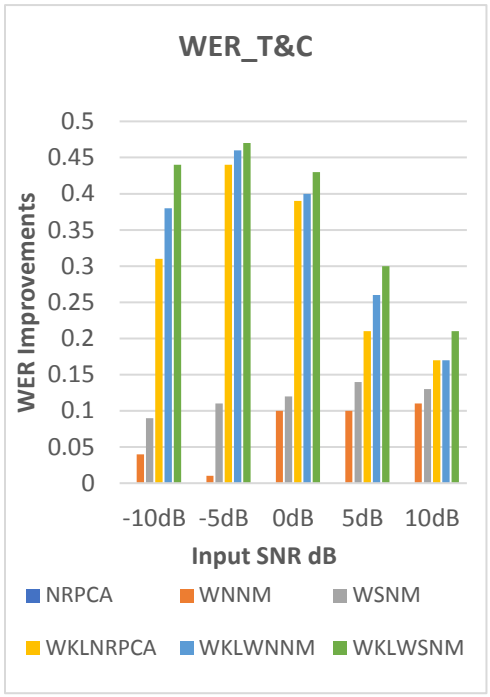
(b)


WER_ Wind
WER Improvements
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-10dB -5dB 0dB 5dB 10dB
Input SNR dB
NRPCA WNNM WSNM
WKLNRPCA WKLWNNM WKLWSNM

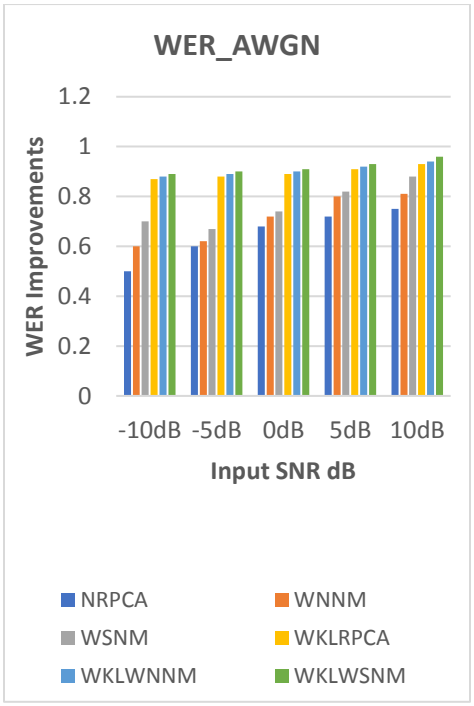
(c)


WER_Machine
WER Improvements
0.25
0.2
0.15
0.1
0.05
0
-10dB -5dB 0dB 5dB 10dB
Input SNR dB
NRPCA WNNM WSNM
WKLNRPCA WKLWNNM WKLWSNM

(d)



(e)



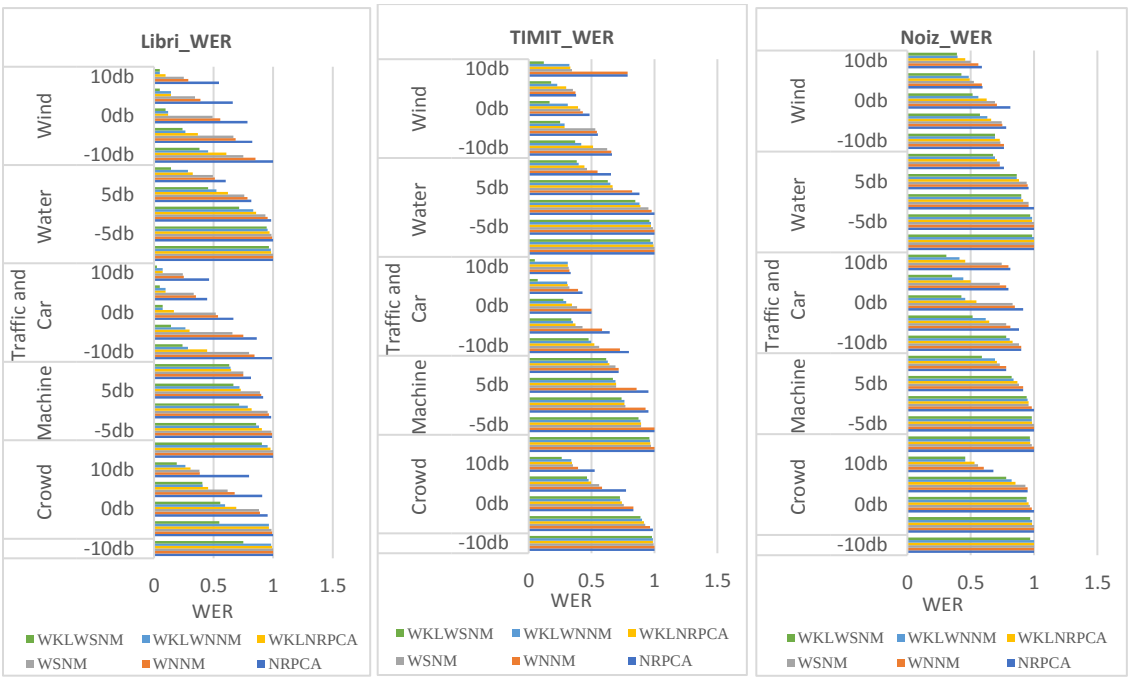
(f)

Figure 6.4(a-c). Performance comparison of the proposed SE algorithms in terms of WER improvements with a) Crowd b) Water c) Wind d) machine e) Traffic & car f) AWGN noises over NRPCA algorithm

6.5 OBSERVATIONS

Based on the preceding chapter and the current studies, the following are the observations:

1. PESQ and WER have partial correlation, which improves PESQ, decreases WER (except for some low SNR cases). Increasing speech quality doesn't decrease recognition accuracy. STOI measures recognition accuracy in noisy settings than PESQ. At low SNR WKLWSNM significantly improves STOI and recognition accuracy for AWGN, Traffic & car and Wind noises. The increase in STOI does not necessarily diminish WER, but the decrease in WER does.
2. The two SE approaches tested here don't show speech recognition gains in noisy environments. In our view, directly correcting speech features for a noisy recognition system reduces WERs far beyond SE approaches.



(a) (b) (c)

Figure 6.5(a-c) Comparison of suggested SE methods over baseline algorithms in terms of Word Error Rate (WER) using a) Libri b) TIMIT c) Noizeus data bases

The generalization capability of the established SE methods was tested as shown in figures 6.5(a-c), utilizing Noizeus, Libri, and TIMIT speech corpora on the kaldi ASR backend (The performance of our proposed methodology with Libri database resulted in the lowest WER values, according to ASR results.Comparing the WSNM, WNNM, and NRPCA SE methods on both the training and test sets demonstrates inferior recognition performance (greater WERs) than proposed WKLWSNM, WKLWNNM, WKLNRPCA methods of SE at all low SNR(< 0dB). With the Libri speech corpus, WKLWSNM and WKLWNNM displayed 24.4% and 28.7% WERs at -10 dB SNR with traffic and car noise, respectively, which was better than the baseline results of 69.8% and 71.2% for WSNM and WNNM.

In all noise situations, the trained ASR with libri speech corpus performed effectively in low SNR levels (< 0 dB) and significantly decreased WER when compared to baseline techniques. Overall, the results demonstrate that performance of our proposed approach was better than those of existing state of art methods.

Chapter 7

Conclusions and Future Scope

This chapter gives an insight into the thesis obtained from the contributions made towards the development of speech enhancement system under low SNR conditions using convex optimization techniques, overcoming the issues addressed in earlier chapters. The scope for future research is discussed with reference to some of the potential areas of advancements in the research field of speech enhancement.

7.1 Conclusions

This thesis offers five contributions, in which a Dictionary-Based speech enhancement methods was evaluated to test the use of sparsity property in clean and noisy speech signals for single channel speech enhancement (SE) using fixed dictionaries (like : DFT, DCT, CEPSTRAL), adaptive dictionaries(like: KSVD,NMF) and low-rank sparse decomposition approaches (like RPCA, SSGODEC). Compressive sensing (CS) recovers a sparse signal from random samples. CS recovery removes background noise and restores structured speech. The study compares fixed and adaptive dictionaries for CS based SE. It established that fixed dictionaries can't represent non-stationary signals sparsely while being simple and fast. Adaptive dictionaries enhance noisy speech better than fixed ones. Prior models are necessary to estimate speech and noise. The experimental results showed that Low-rank sparse decomposition techniques prevent musical noise by estimating the matrix rank using NNM.

The second contribution presented in chapter 3, examined how STFT parameters affect matrix decomposition-based SE. The influence of transform windows is explored on a low-rank or sparse spectrogram. The length N of the STFT influences T-F frequency resolution. The longer the N, the more are the frequency bins. The best STFT length is observed for M = 1536, while the lowest was for N = 512. The RPCA-based SE's SDR performance was impacted by STFT parameters. SDR results rely on noise types, which

is unsolved. Wind, traffic & car noises are easily hidden than the other types. The values of m_l and m_s on speech signal analysis provide the best low-rank, sparse model for speech and noise. The SE algorithm's performance could be evaluated for an infinite number of parameter combinations, and the results could be entirely independently evaluated. The present research on modifying STFT parameters are not exhaustive, and each parameter's value could be changed in a variety of ways.

Low Rank sparse Decomposition splits the input noisy spectrum into low-rank noise and sparse speech components to estimate noise and speech when neither is provided beforehand. Inaccurate rank estimate creates tone-like components in LRSD. Weighted Low-rank and sparse decomposition models with a different objective function than RPCA can minimise musical noise and distortion. At -10dB, WSNM method enhanced Traffic & Car and Wind SDR by 8.14 and 6.17 dB. PESQ improved highest with -10 dB for traffic & car noise (ΔPESQ=0.49) and least with 10 dB of wind and other noise. Figure 4.2 compares suggested and baseline SE SDR performance. In AWGN as a stationary noise situation was also covered. Adopting a binary T-F mask improved speech intelligibility in noisy circumstances. The recommended approach preserves speech content while achieving high SIG values at all SNR and noise levels. The strategy worked effectively at low SNR levels (-10 dB) and minimised residual noise compared to baseline techniques.

RPCA-based techniques improved SE, but have drawbacks. The majority of algorithms maximise their Euclidean distance(ED) -based cost function. ED can produce substantial reconstruction problems since it over emphasises large values. Most of SE approaches improve STFT-based spectral magnitude while preserving input noise-corrupted phase component unaltered. DWPT-based SE techniques directly process data in the temporal domain to correct distortion owing to noisy phase. Cochleagrams can better distinguish monaural mixed audio than spectrograms. Due to the cochleagram's non-uniform time-frequency (T-F) transform, low-frequency T-F units offer better resolution. Cochleogram responses differ for speech and noise. A sparse and low-rank decomposition model may help estimate cochleagram mask. The Frobenius norm

measure is used in cost functions for sparse and low-rank models without regularisation requirements. Decomposed speech must be positive. Based on these findings, we used NRPCA with DWPT to improve unsupervised RPCA-SE.

A novel approach using a weighted low-rank sparse decomposition method, the discrete wavelet packet transform (DWPT), and the KL Divergence were developed, this being contribution 4 and presented in chapter 5. The said approach increases speech enhancement effectiveness in different noise environments by separating speech and noise cochleagram.

NNM-based NRPCA uses auditory perceptual information incorrectly due to an improper matrix rank computation. The findings from chapter 4 show that accurate rank estimation is possible with WSNM method. Therefore, combining DWPT-KL with nonnegative RPCA (WKLNRPCA) improves speech quality and intelligibility over STFT-based SE. The strategy recommended worked well in Traffic & car-noisy settings and scored higher than competing methods in SDR. WKLWSNM provided the greatest PESQ scores in all noisy situations, especially at low SNR levels (10dB and -5dB). Figure 5.6(a-f) compares the proposed method with other techniques that assess increased speech quality.

The method recommended by our study predicted with 88.4% accuracy. All noise sources showed STOI > 86% at 10dB SNR. At -10dB the proposed strategy improves traffic and car noise from 54.4% to 88.2%. The proposed approach decreases speech distortion in all noisy scenarios and at all SNR levels, except for -5dB (SIG = 2.76) crowd noise and -10dB (BAK = 1.43) machine noise. High BAK values indicate reduced residual noise in improved speech. The WKLWSNM technique reduced background noise, resulting in low residual noise.

As a part of fifth contribution (chapter 6), the validation results of the proposed Speech enhancement techniques that were carried out by training Kaldi ASR to achieve low WER using different noises with SNRs ranging from -10dB to 10dB presented. Compared to existing speech enhancement techniques, the recommended algorithms

performed better in terms of WER over noisy speech. The trained ASR utilizing the libri speech corpus worked well in low SNR (< 0 dB), improving WER by more than 85% compared to baseline techniques.

7.2 Future Scope

The proposed SE methods, however, are unable to completely remove background noise since the convex optimization techniques are inaccurate in estimating exact low rank. The problem of developing robust speech enhancement algorithms that can fully remove background noise yet maintaining good quality and intelligibility in highly nonstationary and adversely noisy situations has yet to be solved. For Superior performance, models to estimate exact low-rank and noise type are to be explored.

It is quite remarkable how well the WSNM-RPCA-based unsupervised and untrained technique performs, especially when applied to challenging unstable noises like the sound of a bubbling stream of water or a crowd of people cheering. Even under extremely noisy settings and with low input SNR values, improvement is still possible. The use of larger test signal corpora and more complex objective speech quality metrics is therefore possible. Further research should be done to determine which noise types the SE approach performs best with and which noise kinds are too difficult. The development of real-time realizations of this algorithm may be motivated by the promising findings of WKLWSNM-based RPCA method for particular noise types. Real-time realizations of this algorithm may be useful for hands-free mobile communication in automobiles or hearing aids, for example. High BAK and SIG values, respectively, demonstrate that the suggested approach led to little speech distortion at low SNR levels and that little residual noise was detected in speech processed by the proposed algorithm.

References

[1] S. Boll, "Suppression of Acoustic noise in speech using spectral subtraction." *IEEE Trans. Acoust.,Speech,SignalProcess*,27,pp.113120.1979,http://dx.doi.org/10.1109/TASSP.1979.1163209.

[2] Eberhard Zwicker, Hugo Fastl, "Psychoacoustics facts and models," *Springer Series in Information Sciences*, Berlin, Germany, Vol.22, 1990.

[3] H. Najafzadeh, H. Lahdidli, M. Lavoie, and L. Thibault, "Use of auditory temporal masking in the MPEG psychoacoustics model 2," in *Proc. of the 114th Convention. Audio Engineering Society*, 2003. Available: http://www.aes.org/e-lib/browse.cfm?elib=12484.

[4] Philipos C Loizou , " Speech enhancement theory and practice, CRC press, 2007 , doi.org/10.1201/9781420015836.

[5] J. Gonzalez, J.L Sanchez-Bote and J.O Garcia, " Speech dereverberation and noise reduction with a combined microphone array approach in IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), vol.2, 2000. **doi:** 10.1109/ICASSP.2000.859140.

[6] Kostas Kokkinakis and Philipos C. Loizou ," Multi-microphone adaptive noise reduction strategies for coordinated stimulation in bilateral cochlear implant devices ," *J Acoust Soc Am.* vol. 127(5): 3136–3144. pp. 3136-3144, May, 2010, doi: 10.1121/1.3372727.

[7] V. F. Pisarenko, " The Retrieval of Harmonics from a Covariance Function," *Geophysical Journal International*, Volume 33, Issue 3pp.347–366, 1973, https://doi.org/10.1111/j.1365-246X.1973.tb03424.x

[8] R. Schmidt, " Multiple Emitter Location and Signal Parameter Estimation," *IEEE Transactions on antennas and propagation*, VOL. AP-34, NO. 3, pp 276-280, MARCH 1986.

[9] M.Berouti, R.Schwaartz, and J. Makhoul, " Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech , and Signal Processing, IEEE International Conference on ICASSP 1979., Vol.4*, pp. 208-211.

[10] Y. Hu and P.C. Loizou, " Subjective comparison and evaluation of speech enhancement algorithms,"*SpeechCommun.;vol:49(7)*:588–601.Jul-2007.doi: 10.1016/j.specom.2006.12.006.

[11] Weiss, M., Aschkenasy, E., Parsons, T.,. Study and the development of the INTEL technique for improving speech intelligibility. *Technical Report NSC-FR/4023, Nicolet Scientific*

Corporation.

[12] D.W. Tufts, R kumaresan, I Kirsteins, " Data adaptive signal estimation by singular value decomposition of the data matrix, " *Proceedings of IEEE*, vol.70, no.6,pp. 684-685,1982.

[13] B.D. Moore, " The Singular value decomposition and long and short spaces of noisy matrices," *IEEE Trans. On Signal Processing*, vol.41,no. 9 , pp 2826-2838, 1993.

[14] M.Dendrinos, S Bakamidis, G.Carayannis, " Speech enhancement from noise: A regenerative approach," *Speech Comm*. Vol. 10,, no.1,pp 45-57, 1991.

[15] Y. Ephraim, H.L.V. Trees, "A signal subspace approach for speech enhancement", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[16] Gerhard Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands,"*Proc. Eurospeech*, pp. 1513-1516, 1995.

[17] Rainer Martin," Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics ," IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 9, NO. 5, JULY 2001.

[18] Israel Cohen, " Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement ," IEEE SIGNAL PROCESSING LETTERS, VOL. 9, NO. 1, JANUARY 2002.

[19] S. Rangachari, P.C.Loizou, Yi Hu , "A noise estimation algorithm with rapid adaptation for highly nonstationary environments," IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1 , ppI-305-8, May 2004, doi: 10.1109/ICASSP.2004.1325983.

[20] H. Sameti, H. Sheikhzadeh, L. Deng and R. L. Brennan, "Hmm-based strategies for enhancement of speech signals embedded in nonstationary noise", *IEEE Trans. Audio Speech Lang. Process.*, vol. 6, no. 5, pp. 445-455, Sep. 1998.

[21] S. Srinivasan, J. Samuelsson and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement", *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 1, pp. 163-176, Jan. 2006.

[22] D. Y. Zhao, W. B. Kleijn, A. Ypma and B. de Vries, "Online noise estimation using stochastic-gain hmm for speech enhancement", *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 4, pp. 835-846, May 2008.

[23] J. Bai and M. Brookes, "Adaptive hidden markov models for noise modelling", *Proc. 19th Eur. Signal Process. Conf. (EUSIPCO'11)*, pp. 494-499, 2011-Aug.

[24] N. Mohammadiha, R. Martin and A. Leijon, "Spectral domain speech enhancement using hmm state-dependent super-gaussian priors", *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 253-256, Mar. 2013.

[25] P. Smaragdis, "Non-negative matrix factor deconvolution: Extraction of multiple sound sources from monophonic inputs", *Proc. 5th Int. Conf. Independent Component Analysis*, pp. 494-499, 2004.

[26] M. N. Schmidt, J. Larsen and K. Lyngby, "Wind noise reduction using non-negative sparse coding", *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, pp. 431-436, 2007.

[27] Z. Chen and D. P. Ellis, "Speech enhancement by sparse low-rank and dictionary spectrogram decomposition", *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, pp. 1-4, 2013.

[28] P. K. Ghosh, A. Tsiartas and S. S. Narayanan, "Robust voice activity detection using long-term signal variability", *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 3, pp. 600-613, Mar. 2011.

[29] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition", *Proc. ICASSP*, pp. 7092-7096, 2013.

[30] K. K. Wu, L. Wang, F. K. Soong and Y. Yam, "A sparse and low-rank approach to efficient face alignment for photo-real talking head synthesis", *Proc. ICASSP*, pp. 1397-1400, 2011.

[31] J.O' Donnell, "60- looking through the right window improves spectral analysis," in *Electronic Circuits, Systems, and Standards*, I Hickman, Ed Newnes, pp. 192-198.1991.

[32] J. Benesty, M.M. Sondhi, Y.A. Huang," Springer Handbook of Speech Processing," Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[33] E.Murnolo, L.Pivetta, and C.Chiaruttini, "Speech Enhancement in Wireless Digital Communication via Heuristic Rules and Image Relaxation Techniques," in 9th European Signal Processing Conference(EUSIPCO 1998,pp 1-4,), Sept. 1998.

[34] M. Rahmani, N. Yousefian, A. Akbari, " Energy based speech enhancement technique for hands free communication ," Volume 45, Issue 1, January 2009, p. 85-86. 2009, DOI: 10.1049/el:20092177.

[35] Wee-Tong Lim, " A speech enhancement Technique for SAF Army Communication Systems," in Defense Science Research Conference and Expo, Aug 2011, pp 1-4.

[36] F.Beritelli and C.Rametta, " HSDPA Dual Streaming Approach for Improving VoIP Speech Quality in Forensic Applications," IEEE CSNDSP Manchester,2014,pp 739-743. DOI:10.1109/CSNDSP.2014.6923924.

[37] Sriram Srinivasan, Ashish Pandharipande," Optimal rate allocation for speech enhancement using remote power-constrained wireless microphones, " IET Signal Processing, Vol. 8, no. 7, pp. 792-799, Sept. 2014, Doi: 10.1049/iet-spr.2013.0417

[38] W Nabi, Noureddine A, Adnane C, " An improved speech enhancement algorithm based on wavelets for mobile communication," *2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP),* 2016, pp 622-626, DOI: 10.1109/ATSIP.2016.7523171.

[39] Jinyu Li, Li Deng, Y Gong, and Reinhold H," An Overview of Noise-Robust Automatic Speech Recognition," IEEE TRANS. AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL.22, N0.4, pp. 745-777, April 2014.

[40] C. Guan; Y. Chen; B. Wu, "Direct modulation on LPC coefficients with application to speech enhancement and improving the performance of speech recognition in noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1993.

[41] Colin Breithaupt and Rainer Martin ,"Statistical analysis and performance of DFT domain noise reduction filters for robust speech recognition," INTERSPEECH 2006-ICSLP.

[42] Douglas O' Shaughnessy ," Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, Vol.41, no.10, pp 2965-2979.2008. https://doi.org/10.1016/j.patcog.2008.05.008.

[43] John H. L. Hansen, and Mark A. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Transactions on signal processing*, Vol. 39, NO. 4. Pp.795-805, APRIL 1991.

[44] T Athanaselis, Stavroula-Evita F, Stelios B, I Dologlou and Georgios G, " Signal Enhancement for Continuous Speech Recognition*," Springer -Berlin Heidelberg,* 2003, pp. 1117-1124, [online], DOI: 10.1007/3-540-44989-2_133

[45] C.N Quoc, D.T.Tien, K.N.Dang and B.N.Huu. " Robust speech recognition based on binaural speech enhancement system as a preprocessing step," in proceedings of the third symposium on information and communication Technology, SoICT

[46] Delcroix M, K.Kinoshita, Nakatani. T, Araki S and others " Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," Computer Speech & Language, Volume 27, Issue 3, May 2013, Pages 851-873. doi:10.1016/j.csl.2012.07.006.

[47] W.Li, Wang.L, Y.Zhou, Robust log Energy estimation and its dynamic change enhancement for in-car speech recognition ," *IEEE Transactions on Audio, speech, and language processing*, Vol.21, issue no.8, pp. 1689-1698, Aug 2013.

[48] H.Y.Lee, .W Cho, M Kim, HM Park, " DNN-Based Feature Enhancement Using DOA-Constrained ICA for Robust Speech Recognition," IEEE Signal Processing Letters, Volume: 23, Issue: 8, pp. 1091-1095, August 2016.

[49] J W Cho, J H Park, J H.Chang, " Bayesian feature enhancement using independent vector analysis and reverberation parameter re-estimation for noisy reverberant speech recognition," Computer Speech & Language, Volume 46, November 2017, Pages 496-516. Available on: https://doi.org/10.1016/j.csl.2017.01.010.

[50] C.H. YOU, Bin MA, "Spectral-domain speech enhancement for speech recognition," Speech Communication, Vol. 94, , pp. 30-41, Nov 2017. Available on : https://doi.org/10.1016/j.specom.2017.08.007.

[51] J. W. Cho, Hyung min park," Independent vector analysis followed by HMM-based feature enhancement for robust speech recognition," *Signal Processing*, Vol.120, March 2016, pp. 200-208. https://doi.org/10.1016/j.sigpro.2015.09.002.

[52] "Deafness and hearing loss" http:// www.who.int/mediacentre/factsheets/fs300/en/, updated ,Apr 2021.

[53] Nathaniel Whitmal, Janet Rutledge and Jonathan Cohen, " Denoising Speech Signals for Digital Hearing Aids: A Wavelet Based Approach ," Applied and Numerical Harmonic Analysis book series (ANHA), pp 299-331, Feb 2011, Available on : DOI: https://doi.org/10.1007/ 978-0-8176-8095-4_14.

[54] L Álvarez, E Alexandre, Cosme L, Roberto GP and Lucas C ," Speech Enhancement in Noisy Environments in Hearing Aids Driven by a Tailored Gain Function Based on a Gaussian Mixture Model," International Conference on Artificial Intelligence and Soft Computing - ICAISC 2013: Artificial Intelligence and Soft Computing, pp 503–514, DOI: 10.1007/978-3-642-38658-9_45

[55] J. Thiemann, M Müller, Daniel Mt, Simon D and Steven van de Par ,"Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene," *EURASIP Journal on Advances in Signal Processing* volume , Article No.12 ,2016.Available online: DOI: 10.1186/s13634-016-0314-6.

[56] Chandan K A Reddy, Nikhil Shankar, Gautam S Bhat, Ram Charan, Issa Panahi, "An individualized super Gaussian single microphone Speech Enhancement for hearing aid users with smartphone as an assistive device," IEEE SIGNAL PROCESSING LETTERS, Vol 24 No. 11, pp. 1601-1605, Nov 2017.

[57] Nayan.M, Y.Karuna, Sourabh.T,“ Design of multichannel wiener filter for speech enhancement in hearing aids and noise reduction technique,” 2016 Online International Conference on Green Engineering and Technologies (IC-GET), pp. 1-4, Nov 2016. DOI: 10.1109/GET.2016.7916626.

[58] Simon J. Godsill, Peter J. W. R,“ Digital Audio Restoration,”, Boston, MA Springer US,2002,pp. 133-194 Available on line: http://doi.org/10.1007/0-306-47042-X_4

[59] Paulo A.A. Esquef , “ Audio Restoration,” Springer NY,2008 pp 773-784 , Available on : DOI: 10.100+7/978-0-387-30441-0_40

[60] S.V. Vaseghi, P.J.W. Rayner, “ A new application of adaptive filters for restoration of archived gramophone recordings, *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pp: 2548-2551, Apr 1988, DOI Bookmark: 10.1109/ICASSP.1988.197163.

[61] D. O'Shaughnessy, Peter Kabal, D. Bernardi, “Applying speech enhancement to audio surveillance,” IEEE International Carnahan Conference on Crime Counter measures, pp.69-72, Nov. 1988. DOI: 10.1109/CCST.1988.75991.

[62] O.Cappe and J Laroche, “ Evaluation of short time spectral attenuation techniques for the restoration of musical recordings ,” IEEE Transactions on speech and Audio processing, Vol.3, No. 1, pp 84-93, Jan 1995.

[63] Andrzej Czyżewski, , Marek Dziubiński, , Łukasz Litwic ,Przemysław Maziewski ,” Intelligent Algorithms for Movie Sound Tracks Restoration,” Springer , 2006, Transactions on Rough Sets , pp. 123-145. Available on DOI: 10.1007/11847465_6.

[64] Xiao; R. M. Nickel , “Speech Enhancement With Inventory Style Speech Resynthesis,” IEEE Transactions on Audio, Speech, and Language Processing , Vol: 18, Issue: 6, pp 1243-1257, Aug. 2010.

[65] Jacob Benesty, Shoji Makino, Jingdong Chen, “Speech Enhancement,” Oxford , Academic Press,2014, Available on line: htttp://www.sciencedirect.com/ science/ book/9780128001394.

[66] J.S. Lim, A.V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” Proceedings of the IEEE ,Vol. 67, Issue: 12, Dec 1979, pp. 1586 – 1604, DOI: 10.1109/PROC.1979.11540.

[67] Y. Ephraim, “Statistical-model-based speech enhancement systems,” Proceedings of the IEEE , Vol. 80, Issue: 10, , pp.1526 – 1555, October 1992, Available on online : DOI: 10.1109/5.168664

[68] Sunil Kamath, Philipos Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002 May 2002, Available on online : DOI:10.1109/ICASSP.2002.5745591.

[69] S LeeDavid, K.Han, and H.Ko , "Single-channel speech enhancement method using reconstructive NMF with spectro-temporal speech presence probabilities," Applied Acoustics,Vol.117, Feb 2017, pp. 257-262, Available on https://doi.org/10.1016 / j.apacoust.2016.04.024

[70] Pejman Mowlaee, Johannes Stahl, and Josef Kulmer, , "Iterative joint MAP single-channel speech enhancement given non-uniform phase prior," Speech Communication, Vol. 86 Issue C, pp 85–96, Feb. 2017. Available online: https://doi.org/10.1016/j.specom.2016.11.008

[71] K Lee, " Application of non-negative spectrogram decomposition with sparsity constraints to single-channel speech enhancement," Speech Communication, Vol.58, Mar, 2014, pp 69–80, Available online: https://doi.org/10.1016/j.specom.2013.11.008

[72] StephenSo, Kuldip K.Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," Speech Communication, Volume 53, Issue 6, July 2011, Pages 818-829, Available on: https://doi.org/10.1016/j.specom.2011.02.001.

[73] Allan Kardec Barros, Noboru Ohnishi, "Single channel speech enhancement by efficient coding," Signal Processing, Vol. 85, Issue 9, Sept 2005, pp.1805-1812, Available on : https://doi.org/10.1016/j.sigpro.2005.03.011

[74] Kuldip Paliwal, Kamil Wo´jcicki , Belinda Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," Speech Communication vol.52, no. 5 pp. 450–475, 2010, Available on: doi:10.1016/j.specom.2010.02.004

[75] Pejman Mowlaee Begzade Mahale, M. Blass, W. Kleijn, "New Results in Modulation-Domain Single-Channel Speech Enhancement,"IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 25 , no. 11, pp 2125-2137,Nov 2017.

[76] Martin Krawczyk and Timo Gerkmann, "STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement," IEEE/ACM Transactions on Audio, Speech, and Language processing, Vol. 22, No. 12, Dec 2014.

[77] C.S.J Doire, M. Brookes, P.A. Naylor, C M Hicks, D. Betts, M A Dmour and S H Jensen, " Single-Channel Online Enhancement of Speech Corrupted by Reverberation and Noise,"

IEEE/ACM Transactions on Audio, Speech, and Language Processing , Vol. 25, no. 3 pp 572-587,Dec 2016, DOI:10.1109/ TASLP.2016.2641904

[78] M Pirolt, J.Stahl, P. Mowlaee, " Phase estimation in single-channel speech enhancement using phase invariance constraints," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5585-5589, Mar 2017, DOI:10.1109/ICASSP.2017.7953225

[79] Pejman Mowlaee B, Josef Kulmer, " Harmonic Phase Estimation in Single-Channel Speech Enhancement Using Phase Decomposition and SNR Information," IEEE/ACM Transactions on Audio, Speech, and Language Processing , Vol. 23, no. 9 pp 1521-1532, Sept. 2015.

[80] Manohar, K., and Rao, P, " Speech enhancement in nonstationary noise environments using noise properties," Speech communication, , 48(1), pp.96–109, 2006

[81] J. Meyer, K.U. Simmer, " Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," IEEE International Conference on Acoustics, Speech, and Signal Processing , pp. 1167-1170, April 1997 , Available on DOI:10.1109/ICASSP.1997.596150.

[82] J.Tu, Y. Xia, " Effective Kalman filtering algorithm for distributed multichannel speech enhancement," Neurocomputing, Vol 275, Jan 2018, pp.144-154, Available on : https://doi.org/10.1016/j.neucom.2017.05.048.

[83] S M Kim; H K Kim, S J Lee; Y K Lee, " Adaptation mode control with residual noise estimation for beamformer-based multi-channel speech enhancement," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 301-304, Mar 2012, DOI**:** 10.1109/ICASSP.2012.6287876.

[84] Niko Moritz, Kamil Adiloğlu, JörnAnemüller, StefanGoetze, BirgerKollmeier, "Multi-Channel Speech Enhancement and Amplitude Modulation Analysis for Noise Robust Automatic Speech Recognition," Computer Speech & Language, Vol. 46, 558-573 Nov 2017, Available on: https://doi.org/10.1016/j.csl.2016.11.004.

[85] Jörn Anemüller; Hendrik Kayser, Multi-channel signal enhancement with speech and noise covariance estimates computed by a probabilistic localization model, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 156-160. Mar 2017, Available on **DOI:** 10.1109/ICASSP.2017.7952137.

[86] Djamila Mahmoudi; Andrzej Drygajlo, " Wavelet transform based coherence function for multi-channel speech enhancement," 9th European Signal Processing Conference (EUSIPCO 1998), pp. 1-4, Sep. 1998.

[87] WahbiNabi, NoureddineAloui and AdnaneCherif, "An improved speech enhancement algorithm for dual-channel mobile phones using wavelet and genetic algorithm," Computers & Electrical Engineering , Vol 62, pp. 692-705, Aug 2017, Available on.: https://doi.org/10.1016/j.compeleceng.2017.01.020.

[88] Jahn Heymanna, Lukas Drudea , Reinhold Haeb-Umbacha , "A Generic Neural Acoustic Beamforming Architecture for Robust Multi-Channel Speech Processing," Computer Speech and Language, Vol. 46, Issue C , pp 374–385, Nov. 2017. https://doi.org/10.1016/j.csl.2016.11.007.

[89] Renjie Tong; Guangzhao Bao; Zhongfu Ye, " A Higher Order Subspace Algorithm for Multichannel Speech Enhancement," IEEE Signal Processing, Letters , Vol:22, Issue: 11, pp. 2004-2008, Nov 2015, DOI**:** 10.1109/LSP.2015.2453205

[90] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," IEEE Transactions on Signal Processing ,Vol: 40, Issue: 4, pp. 725 – 735, Apr 1992, DOI**:** 10.1109/78.127947.

[91] H. Sameti; H. Sheikhzadeh; Li Deng; R.L. Brennan, " HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," IEEE Transactions on Speech and Audio Processing , Vol: 6, Issue: 5, pp. 445 – 455, Sept. 1998, DOI**:** 10.1109/89.709670.

[92] David Y. Zhao; W. Bastiaan Kleijn, " HMM-Based Gain Modeling for Enhancement of Speech in Noise," IEEE Transactions on Audio, Speech, and Language Processing, Vol: 15, Issue: 3, pp. 882-892,Mar. 2007, DOI:10.1109/TASL.2006.885256.

[93] Nasser Mohammadiha, , Rainer Martin, and Arne Leijon , " Spectral Domain Speech Enhancement using HMM State-Dependent Super-Gaussian Priors," IEEE SIGNAL PROCESSING LETTERS, Vol.20, No. 3, pp. 253-256, Mar 2013.

[94] HadiVeisi, HosseinSameti, " Speech enhancement using hidden Markov models in Mel-frequency domain," Speech Communication, Vol 55, Issue 2, pp.205-220, Feb 2013, , https://doi.org/10.1016/j.specom.2012.08.005.

[95] SarangChehrehsa, Tom James Moir, "Speech enhancement using Maximum A-Posteriori and Gaussian Mixture Models for speech and noise Periodogram estimation ,"Computer Speech & Language, Vol 36, Mar2016, pp. 58-71, https://doi.org/ 10.1016/j.csl.2015.09.001.

[96] Sarang Chehrehsa, M.H. Savoji, " MMSE speech enhancement using GMM," The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012),May 2012, pp. 266-271, DOI**:** 10.1109/AISP.2012.6313756.

[97] T.V.Sreenivas and Pradeep K, " Codebook constrained Wiener filtering for speech enhancement, IEEE Transactions on Speech and Audio Processing, Vol.4 , No. 5, Sept. 1996.

[98] S. Srinivasan, J. Samuelsson, W.B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement, IEEE Transactions on Audio, Speech, and Language Processing ,Vol: 14, Issue: 1, pp. 163-176, Jan 2006.

[99] Yong Xu; Jun Du; Li-Rong Dai; Chin-Hui Lee , "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," IEEE Signal Processing , Letters ,Volume: 21, Issue: 1, pp.65-68, Jan 2014, DOI**:** 10.1109/LSP.2013.2291240

[100] Yong Xu, Jun Du, Li-Rong Dai, Chin-Hui Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing ,Vol: 23, Issue: 1, pp 7-19,.Jan 2015, DOI**:** 10.1109/TASLP.2014.2364452.

[101] Kevin W. Wilson, Bhiksha Raj, Paris Smaragdis, Ajay Divakaran, " Speech denoising using nonnegative matrix factorization with priors ," 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4029-4032. Mar 2008, DOI: 10.1109/ICASSP.2008.4518538.

[102] Nasser Mohammadiha, Jalil Taghia, Arne Leijon, "Single channel speech enhancement using bayesian nmf with recursive temporal updates of prior distributions ," ICASSP2012, 2012 IEEE International Conference on Acoustics, Speech and Signal Processing, Mar 2012, pp. 4561-4564. DOI: 10.1109/ICASSP.2012.6288933.

[103] Hao-Teng Fan, Jeih-weih Hung; Xugang Lu; Syu-Siang Wang; Yu Tsao, Speech enhancement using segmental nonnegative matrix factorization," ICASSP2014, IEEE International Conference on Acoustics, Speech and Signal Processing, May 2014, pp. 4483-4487.

[104] Lin Qiao, Xiongwei Zhang , Xushan Chen , JibinYang, " Speech Enhancement Using Non-negative Matrix Factorization Solved By Improved Alternating Direction Method Of Multipliers," 2016 International Conference on Progress in Informatics and Computing (PIC) , pp 34-378, Dec. 2016, DOI: 10.1109/PIC.2016.7949529 .

[105] PhilipHarding, BenMilner, "Reconstruction-based speech enhancement from robust acoustic features," Speech Communication, Vol 75, pp. 62-75,Dec. 2015, https://doi.org/10.1016/j.specom.2015.09.011.

[106] Tian Gao, Jun Du , Li-Rong Dai,Chin-Hui Lee, "A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR

environments, Speech Communication, Vol 95, pp. 28-39, Dec 2017, Available on : https://doi.org/10.1016/j.specom.2017.10.003.

[107] A. Petrovsky, M. Parfieniuk, A. Borowicz, Warped DFT Based Perceptual Noise Reduction System, Journal of Audio Engineering Society Convention 116, May 2004, Available online: http://www.aes.org/e-lib/browse.cfm?elib=12674.

[108] R.M. Udrea, Nicolae D. Vizireanu, Silviu Ciochina, "An improved spectral subtraction method for speech enhancement using a perceptual weighting filter," Digital Signal Processing, vol. 18 No. 42008, pp. 581–587, 2008 , Available online: http://www.sciencedirect.com/science/article/pii/S10511200407001145

[109] Yang Lu, Philipos C. Loizou , "A geometric approach to spectral subtraction," Speech Communication vol.50, no.6, 2008, pp. 453–466, Available online: http://www.sciencedirect.com/science/article/pii/S0167639308000125

[110] P.Scalart and J Filho, "Speech enhancement based on a priori signal to noise estimation, " International Conference Acoustics, Speech, and Signal Processing, 1996. ICASSP-96, vol. 2, DOI:10.1109/ICASSP.1996.543199.

[111] Ch.V.Rama Rao, M.B.Rama Murthy, K.Srinivasa Rao, Speech enhancement using sub-band cross-correlation compensated Wiener filter combined with harmonic regeneration, AEU - International Journal of Electronics and Communications, Vol 66, Issue 6, June 2012, pp.459-464 , https://doi.org/10.1016/j.aeue.2011.10.007.

[112] Saeed V. Vaseghi , "Wiener Filters," Advanced Digital Signal Processing and Noise Reduction, Second Edition, John Wiley & sons, Chapter 6, September 2001, https://doi.org/10.1002/0470841621.ch6

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

[114] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, May 2002,pp. I-253-I-256.

[115] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," in IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 845-856, Sept. 2005, doi: 10.1109/TSA.2005.851927

[116] B. Chen and P.C Loizou, “A laplacian-based mmse estimator for speech enhancement,” Speech Communication, vol. 49, no. 2, pp. 134 – 143, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639306001956

[117] M. B. Trawicki and M. T. Johnson, “Speech enhancement using Bayesian estimators of the perceptually- motivated short time spectral amplitude (stsa) with chi speech priors,” Speech Communication, vol. 57, no. Supplement C, pp. 101 – 113, 2014. [Online]. Available on : http://www.sciencedirect.com/ science/rticle/ pii/S0167639313001301

[118] H. R. Abutalebi and M. Rashidinejad, “Speech enhancement based on Is-order mmse estimation of short time spectral amplitude and Laplacian speech modelling,” Speech Communication, vol. 67, no. Supplement C, pp. 92 – 101, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639314000879

[119] B. M. Mahmmod, A. R. Ramli, S. H. Abdulhussian, S. A. R. Al-Haddad, and W. A. Jassim, “Low-distortion mmse speech enhancement estimator based on lapacian prior,” IEEE Access, vol.5, pp. 9866-9881, 2017.

[120] B. BabaAli, H. Sameti, and T. H. Falk, “A model distance maximising framework for speech recognizer-based speech enhancement,” {AEU} – International Journal of Electronics and Communication, vol. 65, no. 2,pp. 99 – 106,2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1434841110000336

[121] X. Lu, M. Unoki, S. Matsuda, C.Hori, and H. Kashioka, “Controlling tradeoff between approximation accuracy and complexity of a smooth function in a reproducing kernel Hilbert space for noise reduction, “ IEEE Transactions on Signal Processing, vol. 61, no. 3, pp. 601 – 610, Feb 2013.

[122] C. H. You, S. Rahardja and S. N. Koh, "Audible Noise Reduction in Eigen domain for Speech Enhancement," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 6, pp. 1753-1765, Aug. 2007, doi: 10.1109/TASL.2007.899288.

[123] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," in IEEE Transactions on Speech and Audio Processing, vol. 9, no. 2, pp. 87-95, Feb. 2001, doi: 10.1109/89.902276.

[124] P. S. K. Hansen, P. C. Hansen, S. D. Hansen and J. A. Sorensen, "Experimental comparison of signal subspace based noise reduction methods," 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258), 1999, pp. 101-104 vol.1, doi: 10.1109/ICASSP.1999.758072.

[125] Y. Hu and P. C. Loizou, "A Comparative Intelligibility Study of Speech Enhancement Algorithms," 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, 2007, pp. IV-561-IV-564, doi: 10.1109/ICASSP.2007.366974.

[126] F. Jabloun; B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," IEEE Transactions on Speech and Audio Processing , Vol: 11, Issue: 6, pp.700 – 708,Nov 2003.

[127] Gwo-Hwa Ju; Lin-Shan Lee, "A Perceptually Constrained GSVD-Based Approach for Enhancing Speech Corrupted by Colored Noise," IEEE Transactions on Audio, Speech, and Language Processing, Vol: 15, Issue: 1, pp**:** 119 – 134. Jan. 2007.

[128] Sigg, Cd., Dikk, T., et.al, *speech enhancement using generative dictionary learning*, IEEE Trans. on AUD, sp., and lang proc. vol: 20 , issue: 6 , aug. 2012 ,pp 1698-1712,

[129] Amin H A, Sid Ahmed S., *Speech enhancement using PCA and variance of the reconstruction error model identification,* interspeech,2007.

[130] Balcan, D.C., Rosca, J., *Independent component analysis for speech enhancement with missing tf content*, proc 6th int. conf. ind. comp. anal blind signal separation, charleston, USA, Mar 2006, pp552-560.

[131] Guo, X., Hairong, J., *Speech enhancement using the improved k-svd algorithm by subspace*, journal of xidian university (natural science edition) 2016, 43(6). DOI: 10.3969/j.issn.1001-2400.2016.06.019

[132] Kevin Wilson, W., Bhiksha, R., *Regularized Non-negative matrix factorization with temporal dependencies for speech denoising,* Interspeech 2008, Brisbane, Australia.

[133] *Y.V.Varshney, Zia Ahmad, A M Raza A, O Farooq, Frequency Selection Based Separation of Speech Signals with Reduced Computational Time Using Sparse NMF*. Archives of Acoustics, 42, 2, 2017, pp. 287–295, , 10.1515/aoa-2017-0031.

[134] Hadi Veisi, Hossein Sameti, Ali Aroudi, *Hidden Markov model-based speech enhancement using multivariate Laplace and Gaussian distributions*, 2015, Signal Processing, IET 9(2):177-185, DOI: 10.1049/iet-spr.2014.0032.

[135] N Saleem, E Mustafa, A Nawaz, A Khan, *Ideal binary masking for reducing convolutive noise,* International Journal of Speech Technology 18 (4), 2015,pp 547-554,

[136] Y Wang, Arun N, DeLiang Wang, *On Training Targets for Supervised Speech Separation* IEEE/ACM Trans Audio Speech Lang Proc, 2014;22(12):1849-1858. DOI: 10.1109/TASLP.2014.2352935.

[137] N Saleem, M Irfan Khattak, MY Ali, M Shafi, *Deep neural network for supervised single-channel speech enhancement,* Archives of Acoustics 44, 2019.

[138] Cheng Yu, Ryandhimas,Zezario, SS Wang, Jonathan S, Yi-Yen, Xugang Lu, HsWang, Tsao, *Speech Enhancement based on Denoising Autoencoder with Multi-branched Encoders,* arXiv:2001.01538v3 [eess.AS] Dec 2020.

[139] Candès, E. J., Li, X., Ma, Y., Wright, Robust principal component analysis. journal of the ACM (jacm), 2011, 58(3), 11.

[140] Wright, J., Ganesh, A., Rao, S., Peng, Y., & Ma, Y. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. Advances in neural information processing systems, 2009, pp. 2080–2088.

[141] C. Sun, Q. Zhang, J. Wang, J. Xie Noise reduction based on robust principal component analysis, Journal of Computational Information Systems, 2014,10(10):4403-4410, DOI:10.12733/jcis10408.

[142] Dennis L. Sun, Fovette, Alternating Direction method of multipliers for non-negative matrix factorization with the beta-divergence, IEEE- ICASSP 2014. DOI**:** 10.1109/ICASSP.2014.6854796

[143] ITU Recommendation ITU-R bs . 1387-1 method for objective measurements of perceived audio quality, tech. rep., 2001.

[144] Vincent, E., Jafari, M. G., Preliminary guidelines for subjective evaluation of audio source separation algorithms, in ica research network international workshop, Liverpool, UK, 2006, pp. 93–96.

[145] Hu, Y., & Loizou, P. C, Evaluation of objective quality measures for speech enhancement. IEEE transactions on audio, speech, and language processing, 2008, 16(1), 229–238. http://www.utdallas.edu/~loizou/speech/noizeus/

[146] Nicolas Gillis And Francois Glineur. low-rank matrix approximation with weights or missing data is np-hard. siam journal on matrix analysis and applications, 32(4):1149–1165, 2011.

[147] Sun, Pengfei and Qin, Jun.: Low Rank and Sparsity Analysis Applied to Speech Enhancement via Online Estimated Dictionary. IEEE Signal Processing Letters23, No.12,(Fall2016):1862-1866. DOI:10.1109/LSP.2016.2627029.

[148] Cai. J F., Cand`Es, E J. A singular value thresholding algorithm for matrix completion, siam j. optimz. 2010, vol. 20, no. 4, pp. 1956–82.

[149] Boyd, S., Parikh, N., "Distributed optimization and statistical learning via the alternating direction method of multipliers," foundations and trends in machine learning, 2010,3(1), pp. 1–122.

[150] Huang, P. S., Chen, S. D., et.al singing voice separation from monaural recordings using robust principal component analysis. IEEE Intl. conference on acoustics, speech and signal processing,2012, pp. 57–60.

[151] Sorensen, K. V., & Andersen, S. V., speech enhancement with natural-sounding residual noise based on connected time-frequency speech presence regions. eurasip journal on advances in signal processing, 2005(18), 305909.

[152] Rangachari, S., Loizou, " A noise-estimation algorithm for highly non-stationary environments," journal . Speech Comm,2006,48(2), pp 220–231.

[153] Zhou, T., & Tao, D, godec randomized low-rank & sparse matrix decomposition in noisy case. proc 28th international conference on machine learning,2011, pp 33-40.

[154] S. Gu, L. Zhang, W.et .al "weighted nuclear norm minimization with application to image denoising," 2014 IEEE conference on computer vision and pattern recognition, Columbus, 2014, pp. 2862-2869, DOI: 10.1109/cvpr.2014.366.

[155] Yuan, Xie., Shuhang, Gu., Yan, Liu., " weighted Schatten p-norm minimization for image denoising and background subtraction, IEEE trans on image processing 2016, vol: 25, issue: 10, pp 4842-4857.

[156] S Islam, Tarek Hasan, Wasim U Khan, and Zhongfu Ye, Supervised Single-Channel Speech Enhancement Based on Stationary Wavelet Transforms and Non-negative Matrix Factorization with Concatenated Framing Process and Subband Smooth Ratio Mask, Journal of Signal Processing Systems 92(2), 2020 DOI: 10.1007/s11265-019-01480-7.

[157] *H Liu*, *W Wang*, *Lin X*, *J Yang*, *Z Wang*, *Chunli*, Speech Enhancement Based on Discrete Wavelet Packet Transform and Itakura-Saito Nonnegative Matrix Factorisation, Archives of Acoustics, 45**,** 4, 2020, pp. 565–572,. DOI: 10.24425/aoa.2020.134072**.**

[158] Bin. G, W.L.Woo, Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic sparsity adaptation, The Journal of the Acoustical Society of America 135, 1171, 2014; https://doi.org/10.1121/1.4864294.

[159] N Saleem, G Ijaz, Low-rank sparse decomposition model based speech enhancement using gammatone filterbank and Kullback–Leibler divergence, IJST 21 (2), 2018,pp 217-231.

[160] Gang M, X wang, X Zou, Mask estimate through Itakura-Saito nonnegative RPCA for speech enhancement, IEEE IWAENC, 2016, DOI: 10.1109/IWAENC38542.2016.

[161] Venkata Sridhar K, Kishore Kumar T, Speech Enhancement For Robust Speech Recognition Using Weighted Low Rank and Sparse Decomposition Models Under Low SNR Conditions, Traitement du Signal, Vol. 39, No.2, April 2022, pp 633-644, DOI: 10.18280/ts.390226.

[162] D.L.Donoho, "Compressive sensing". IEEE Trans. on Inf. Th., Vol. 52, NO. 4, April 2006.

[163] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," IEEE Trans. Inform. Theory, vol. 54, no. 5, pp. 2210–2219, May 2008.

[164] Emmanuel J. Candes and Michael B. Wakin, An Introduction to Compressive sampling. IEEE signal Processing Magazine, Vol.25, , Issue.2, pp.21- 30,2008

[165] E. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," Inverse Problems, vol. 23, no. 3, pp. 969–985, 2007.

[166] TV. Sreenivas and W. Bklejin, "Compressive sensing for sparsely excited speech signals." IEEE Intl Conference on Acoustics, Speech and Signal Processing, Taipei , pp.4125 – 4128, 2009.

[167] Daniele Go, M G Christensen, Manohar N. Murthi, S Holdt Jenson, "Retrieving Sparse patterns using a compressed sensing framework: Applications to speech coding Based on sparse linear prediction" IEEE Signal processing letters, vol.17, Issue.1, pp.103-106, 2010.

[168] Fereshteh Fakhar Firouzeh , Seyed Ghorshi, Sina Salsabili ," Compressed sensing based speech enhancement, 8th International Conference on Signal Processing and Communication Systems (ICSPCS) , Dec 2014,DOI: 10.1109/ICSPCS.2014.7021068.

[169] Houria HANECHE, Bachir Boudraa, A. Ouahabi, Speech Enhancement Using Compressed Sensing-based method, 2018 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM), DOI: 10.1109/CISTEM.2018.8613609.

[170] Viet Hang D, Manh "Dictionary Learning Based speech Enhancement, Chapter, InTech open., DOI: 10.5772/Intechopen.85308.

[171] Elad M, " Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer", 2010.

[172] I. Tosic and P. Frossard, "Dictionary learning: What is the right representation for my signal?," IEEE Signal Process. Mag., vol. 28, no. 2, pp. 27–38, 2011.

[173] Michal Aharon, "Overcomplete Dictionaries for Sparse Representation of Signals", Phd Thesis, the Technion - Israel Institute of Technology,Haifa,2006.

[174] M.S. Manikandan, " Sparse Representation and Compressive sensing, slides share.net, Amrita univ.2011.

[175] CD.Sigg, T Dikk, J Buhmann, " Speech enhancement using Generative Dictionary

Learning" , IEEE Tran. on Aud, Sp. and Lang Proc. Vol: 20 , Issue: 6 , pp 1698-1712, Aug. 2012 .

[176] Siddhi desai, N Nakrani, " Compressive sensing in speech processing: A survey Based on Sparsity andSensing Matrix," IJETAE, Vol. 3 , Iss 12, 2013.

[177] J.C.Wang, Yuan lee, chang, shu, " Compressive Sensing-Based Speech Enhancement" IEEE/ACMTrans. On Aud., spch and lang. proc. Vol: 24 , Issue:11 , Nov. 2016.

[178] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", IEEETransactions on Pattern Analysis and Machine Intelligence, Vol. II, No. 7, pp. 674-693, 1989.

[179] Candes, DL Donoho et al. Curvelets: A surprisingly effective nonadaptive representation for objects with edges. Department of Statistics, Stanford University.1999.

[180] MN Do and M Vetterli."The contourlet transform: an efficient directional multiresolution image representation," IEEE Transactions on Image Processing.; 14(12): 2091–2106, 2005.

[181] S Chen, DL Donoho, M A. Saunders, "AtomicDecomposition by Basis Pursuit,"SIAM Journal on Scientific Computing, Volume 20, Number 1, , pages 33–61. 1998

[182] Joel , Tropp, Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," IEEE Trans on Inf. Th. Vol: 53 , Issue: 12 ,2007.

[183] S Joshi ; K V Siddamal ; V. S. Saroja," Performance analysis of compressive sensing Reconstruction", 2nd IEEE International Conference on Electronics and Communication Systems , Coimbatore, 2015.

[184] Dalei Wu, Wei-Ping Zhu and M.N.S. Swamy, "A Compressive Sensing Method for Noise Reduction of Speech and Audio Signals" , IEEE 54th International Midwest Symposium on Circuits and Systems, Seoul, South Korea, 2011.

[185] Heung-No lee. Introduction to compressive sensing. Lecture notes INFONET. pp. 26-29, 2011.

[186] TV. Sreenivas and W. Bklejin, "Compressive sensing for sparsely excited speech signals." IEEE Intl Conference on Acoustics, Speech and Signal Processing, Taipei , pp.4125 – 4128, 2009.

[187] Yue Wang, Zhixing Xu, Gang Li, LChang and C Hong, Compressive Sensing Framework for Speech Signal Synthesis Using a Hybrid Dictionary. 4th International Congress on Image and Signal Processing , Vol.5, Shanghai, pp. 2400 – 2403, 2011.

[188] J.C.Wang, Yuan lee, chang, shu, " Compressive Sensing-Based Speech Enhancement" IEEE/ACMTrans. On Aud., spch and lang. proc. Vol: 24 , Issue:11 , Nov. 2016.

[189] Vinayak Abrol, Pulkit Sharma and Sumit Budhiraja,"Evaluating performance of compressed sensing for speech signal", IEEE 3rd International advance computing conference (IACC), Ghaziabad, pp.1159-1164, 2013.

[190] D.Needell, J.A.Tropp, ,CoSaMP: Iterative signal recovery from incomplete and inaccurate samples, Applied and Computational Harmonic Analysis, Vol 26, Issue 3, May 2009, pp.301-321.

[191] D.L.Donoho, "Compressive sensing". IEEE Trans. on Inf. Th., Vol. 52, NO. 4, April 2006.

[192] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," IEEE Trans. Inform. Theory, vol. 54, no. 5, pp. 2210–2219, May 2008.

[193] E. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," Inverse Problems, vol. 23, no. 3, pp. 969–985, 2007.

[194] TV. Sreenivas and W. Bklejin, "Compressive sensing for sparsely excited speech signals." IEEE Intl Conference on Acoustics, Speech and Signal Processing, Taipei , pp.4125 – 4128, 2009.

[195] Daniele Go, M G Christensen, Manohar N. Murthi, S Holdt Jenson, "Retrieving Sparse patterns using a compressed sensing framework: Applications to speech coding Based on sparse linear prediction" IEEE Signal processing letters, vol.17, Issue.1, pp.103-106, 2010.

[196] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," Proc. IEEE, vol. 98, no. 6, pp. 1045–1057, 2010.

[197] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," IEEE Trans. Patt. Anal. Mach. Intell., vol. 34, no. 4, pp. 791–804, 2012.

[198] Emmanuel J. Candés, Xiaodong Li, Yi Ma, and John Wright, " *Robust Principle Component Analysis,"* Journal of the ACM, 58(3):11:1–37, May 2011.

[199] Tianyi Zhou and Dacheng Tao. "GoDec: *Randomized Low-rank & Sparse Matrix Decomposition in Noisy Case*," Pro- ceedings of the 28th International Conference on Machine Learning, ICML '11, pages 33–40, New York, USA, June 2011. ACM.

[200] Jianjun Huang, Xiongwei Zhang, Yafei Zhang, Xia Zou, and Li Zeng. *Speech Enhancement via Low-Rank and Sparse Matrix Decomposition*. ETRI Journal, 36(1):167–170, February 2014.

[201] Po-Sen Huang, Scott D. Chen, Paris Smaragdis, and Mark Hasegawa-Johnson. *Singing-*

Voice Separation From Monaural Recordings Using Robust Principal Component Analysis.In IEEE International Conference on Acoustics, Speechand Signal Processing, 2012.

[201] Rahil Mahdian Toroghi, Friedrich Faubel, and Dietrich Klakow. Multi-*Channel Speech Separation with Soft Time- Frequency Masking*. In SAPA - SCALE Conf., pages 86–91, Portland, Oregon, USA, September 2012. SAPA Workshops.

[202] Tianyi Zhou. "Go Decomposition News. https://sites.google.com/site/ godecomposition/.

[203] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Blind sparse source separation with spatially smoothed timefrequency masking," Proc. IWAENC, Sept. 2006.

[204] Y. Li and D.L.Wang, "Musical Sound Separation Based on Binary Time-Frequency Masking," EURASIP J. Audio, Speech, Music Process., vol. 2009, July 2009, pp. 1-10.

[205] Zhuo Chen and Daniel P. W. Ellis. *Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition.* In *IEEE* Workshop on Applications of Signal Processing to Audio and Acoustics, pages 1–4. IEEE, 2013.

[206] Ö zgür Yilmaz and Scott Rickard. Blind *Separation of Speech Mixtures via Time-Frequency Masking.* IEEE Transactions on Signal Processing, 52(7):1830–1847, July 2004.

[207] Robert G. Leonard. A database for *speaker-independent digit recognition*. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 9:328–331, March 1984.

[208] Guoning Hu. 100 *Nonspeech Sounds*. http://web.cse.ohio-state.edu/ dwang/ pnl/ corpus/HuNonspeech/HuCorpus.html.

[209] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. *Performance Measure- ment in Blind Audio Source Separation*. IEEE Transactions on Audio, Speech, and Language Processing, 14(4):1462–1469, July 2006.

[210] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems," Proc. IEEE, vol. 98, no. 6, pp. 948–958, 2010.

[211] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," IEEE Trans. Signal Process., vol. 54, no. 11, pp. 4311–4322, 2006.

[212] Shuhang, G., Xie, Qi., et.al, weighted nuclear norm minimization, and its applications to low-level vision, intl journal of computer vision, 2017, 121(2). DOI: 10.1007/s11263-016-0930-5.

[213] LiWang, DiXiao,Wen S.Hou, Xiao Y.Wu, " Weighted Schatten ***p***-norm minimization for impulse noise removal with TV regularization and its application to

medical images," Biomedical Signal Processing and Control , Vol. 66, April 2021,

[214] A.W.Rix, J.B.Beerends, Holler, " perceptual evaluation of speech quality -a new method for speech quality assessment of telephone networks and codecs," IEEE Int. Conf. On Acc, Speech Signal processing. Proceed 2001, DOI: 10.1109/ICASSP.2001.941023.

[215] C H. Taal, Richard C. Hendriks, Richard H, and Jesper J " An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," IEEE Transactions on audio, speech, and language processing, vol. 19, no. 7, Sept 2011

[216] Daniel, P., Arnab, G., Gilles, B., the Kaldi speech recognition toolkit, speech recognition project, 2011. http://www.kaldi-asr.org, and https://github.com/kaldi- asr/kaldi.git Kaldi

[217] Vassil Panayotov; Guoguo Chen; Daniel Povey; Sanjeev Khudanpur **, "**Libri speech: An ASR corpus based on public domain audiobooks, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), DOI: 10.1109/ICASSP.2015.7178964.

[218] JOHN, S., GAROFOLO, F., LAMEL., "TIMIT Acoustic-phonetic continuous speech corpus, " DOI: 10.35111/17gk-bn40.

List of Publications

International Journals:

[1*] Venkata Sridhar K, Kishore Kumar T, "Speech Enhancement For Robust Speech Recognition Using Weighted Low Rank and Sparse Decomposition Models Under Low SNR Conditions**" Traitement du Signal**, Vol. 39, No.2, April 2022, pages: 633-644, DOI: 10.18280/ts.390226 **(SCI).**

[2*] Venkata Sridhar K, Kishore Kumar T "Investigation on the Influence of Parameterization in Speech Enhancement using Low Rank Sparse Decomposition Models under Low SNR Conditions", **NeuroQuantology**, June 2022, Volume 20, Issue 6, Pages: 9461-9476, DOI: 10.14704/nq.2022.20.6. (**SCOPUS**).

[3*] Venkata Sridhar K, Kishore Kumar T ," Wavelet-Based Weighted Low-rank sparse Decomposition Model for Speech Enhancement Using Gammatone filter Bank Under Low SNR Conditions" **Fluctuation and Noise Letters journal [SCI]**- (Answered reviewer comments- awaiting for acceptance soon).

International Conferences:

[1] K.V. Sridhar, Prof. T. Kishore Kumar, "performance evaluation of cs based speech enhancement using adaptive and sparse dictionaries " 4th International Conference on Recent Advances & Innovations in Engineering (ICRAIE-2019) ,27–29 Nov. 2019 at Penang, Malaysia. (IEEE).